MASTERARBEIT

Discriminant Analysis of Three Rhythmic Descriptors in Musical Genre Classification

Ausgeführt am Institut für
Softwaretechnik und Interaktive Systeme
der Technischen Universität Wien

unter der Anleitung von
ao. Univ. Prof. Dr. Andreas Rauber
und
DI Thomas Lidy

durch

Bernhard Pflugfelder
Malborghetgasse 27-29/4/6
A-1100 Vienna, Austria

Wien, am 25. 03 2008

To my family

"Music is the one incorporeal entrance into the higher world of knowledge which comprehends mankind but which mankind cannot comprehend."

Ludwig van Beethoven (1770-1827), German Composer

Acknowledgments

I want to thank Andreas Rauber and Thomas Lidy for suggesting me this interesting topic, for giving me useful advices and also for being patient in some difficult times.

My special thanks goes to my parents who gave me the opportunity to study computer science and greatly supported me all the time. Without their lovely help I would not have written these lines.

Many thanks goes to my grandmother who also greatly supported me during my studies and therefore played a major role in giving me the possibility to write this thesis as well.

Finally, I also dedicate many thanks to my brother Roman who helped me with many very useful advices and fruitful discussions during all the years of my studies and beyond.

Sincere thanks to all of you!

Bernhard Pflugfelder

Danksagung

Ich möchte auf das Herzlichste Andreas Rauber und Thomas Lidy dafür danken, dass sie mir dieses interessante Thema näher gebracht haben und mir mit wichtigen Ratschlägen zur Seite gestanden sind. Weiters möchte ich mich ganz besonders dafür bedanken, dass sie auch in etwas schwierigeren Zeiten Geduld aufgebracht haben.

Meinen ganz besonderen Dank widme ich meinen Eltern, welche mir überhaupt erst die Möglichkeit gegeben haben, Informatik zu studieren. Weiters möchte ich mich auch für die großartige Unterstützung bedanken, welche sie mir zu Teil haben lassen. Ohne diese großartige Unterstützung hätte ich wohl diese Zeilen nicht schreiben können.

Auch meiner Großmutter möchte ich herzlich für ihre großartige Unterstützung danken. Diese Unterstützung war ebenfalls eine große Hilfe dafür, dass ich mein Studium beenden konnte.

Schließlich möchte ich auch einen ganz besonderen Dank meinem Bruder Roman widmen. Seine zahlreichen und hilfreichen Ratschläge sowie viele ergiebige Diskussionen gaben mir Motivation und Inspiration während all der Jahre meines Studiums und darüber hinaus.

Vielen Dank!

Bernhard Pflugfelder

Abstract

The introduction of digital music representation considerably altered the ways of creating, accessing and using music. Until today an immense number of music archives has been made available so that the actual attitude of "music consumption" has changed fundamentally. Both the commercial domain as for instance represented by music producers or music distributors and the private domain play a major role in the increasing importance of digital music archives. Yet, the size of music archives which can often be enormous demands new requirements according to the internal organization of included musical pieces as well as the individual access and search of musical pieces. Consequently, this means that scalable methods must be provided to automatically establish organizations of music archives according to specific musical semantics. The research field of *Music Information Retrieval* (MIR) aims to develop such methods which make possible a grouping, i. e. clustering or classification, of music pieces according to specifically defined musical semantics. Basically, such a musical semantics refers to the measuring the similarity of the underlying musical content. The definition of this content-based similarity is based on individual musical characteristics like for instance rhythm, melody, instrumentation or others.

Musical genres represent a very popular and frequently used musical category to organize music collections. In comparison to other possible musical categories, musical genres provide an intuitive understanding for categorization and are frequently used by humans to organize music. For instance, music retailers or music libraries widely use genre categorization to provide an effective organization of offered music collections. Within the MIR community the assumption generally holds true that the understanding of genres is potentially based on the descriptive power of certain content-based characteristics of the included musical pieces. Consequently, specific genres may be actually related to a certain rhythmic, melodic or other musical characteristics. Unfortunately, this assumption of music genre representation based on content-based semantics appears to be insufficient as not content-based characteristics like for instance the cultural origin of artists and the cultural context of lyrics also play a role in the definition of musical genres.

Based on that potentially descriptive power of genres this master thesis examines the discrimination of musical genres in terms of rhythmic characteristics. Since various rhythmic descriptors exist in MIR, the three descriptor *Rhythm Pattern* (RP), *Statistical Spectrum Descriptor* (SSD) and *Rhythm Histogram* (RH) have been used throughout this thesis only. Each of these three descriptors contains a large number of features to constitute the specific rhythmic component of

an individual piece of music. In particular two key questions were analyzed based on extensive empirical evaluations. The first question was dedicated to the possible discrimination of genres based on specific feature patterns within the descriptor which were suited for genre discrimination. Such feature patterns were determined by applying five different heuristic discrimination models in order to estimate the contribution of every feature to distinguish a specific genre. For this purpose the *DiscriminationAnalyzer* tool was designed to compute and to visualize the discriminative power of features according to class discrimination. Moreover, the processing of arbitrary feature sets, the selection of different calculation models and an appropriate visualization of the results are key properties of this application. The second question focused on the evaluation of the usefulness concerning dimensionality reduction, i. e. feature selection, based on the discriminative power of the features to correctly distinguish the underlying classes. In usual classification applications, feature selection is especially important not only because of the potential run-time optimization but also because of the deteriorating influence of the *curse of dimensionality*. The presented application also includes an embedded evaluation of the most discriminative features with arbitrary learning algorithms.

Zusammenfassung

Die Einführung der digitalen Repräsentation von Musik hat die Erstellung, den Zugriff sowie die Verwendung von Musikarchiven entscheidend verändert. Bis zum heutigen Zeitpunkt existiert bereits eine immense Anzahl von Musikarchiven, so dass sich die Einstellung gegenüber des "Konsumierens von Musik" fundamental verändert hat. Sowohl der kommerzielle Bereich, wie u. a. Musikproduzenten oder Musikdistributoren, als auch der private Bereich tragen dazu bei, dass die Bedeutung von digitalen Musikarchiven weiterhin zunimmt. Aufgrund der Größe von Musikarchiven müssen jedoch neue Anforderungen an die Organisation von Musikstücken bzw. an den individuellen Zugriff auf und die Suche von Musikstücken berücksichtigt werden. Um einen effektiven Zugriff bzw. eine effektive Suche zu garantieren, sind skalierbare Methoden zur automatisierten Erstellung einer Organisation von Musikarchiven basierend auf einer spezifischen musikalisch-orientierten Semantik notwendig. Das Forschungsgebiet *Music Information Retrieval* (MIR) widmet sich der Entwicklung von solchen Methoden, welche einerseits die Bildung von semantischen Gruppen (*clusters*) von Musikstücken und andererseits dem Klassifizieren von Musikstücken ermöglichen soll. Grundsätzlich bezieht sich eine solche musikalische Semantik auf ein spezifisches Maß, welches die Ähnlichkeit des musikalischen Inhaltes abbildet. Eine solche inhaltliche Ähnlichkeit wird über den Vergleich von musikalischen Aspekten wie Rhythmus, Melodie, Instrumentierung, usw. der einzelnen Musikstücke ermittelt, welche zuvor automatisiert extrahiert werden müssen. Die semantische Beziehung von Musikstücke derselben Gruppe wird somit über eine spezifische Ähnlichkeit des musikalischen Inhaltes definiert.

Eine der interessantesten Kategorien für die Organisation von beliebigen Musikarchiven stellen *Genres* dar. Im Vergleich zu anderen möglichen Kategorien bieten Genres ein intuitives Verständnis zur Organisation von Musikarchiven. Beispielsweise findet man etwa in Musikgeschäften oder in Musikbibliotheken sehr häufig nach Genres strukturierte Musiksammlungen. Innerhalb der Forschungsgemeinschaft von MIR wird angenommen, dass dieses intuitive Verständnis der Kategorisierung von Musikstücken auf die inhärente Aussagekraft von Genres bezüglich bestimmter musikalischer Aspekte der einzelnen Musikstücke desselben Genres basiert. Daraus könnte man folgern, dass ein bestimmtes Genre durch spezifische rhythmische, melodische oder andere musikalische Aspekte eindeutig beschrieben werden kann. Jedoch ist diese Annahme bezüglich einer rein inhaltlich basierenden Repräsentation von musikalischen Genres ungenügend, da auch nicht inhaltsbezogene Aspekte wie beispielsweise die kulturelle Herkunft von Musiker bzw. der kulturelle Kontext von Songtexten die Definition von Genres beeinflussen.

Ausgehend von dieser potentiellen Aussagekraft von Genres wurde im Rahmen dieser Masterarbeit die Unterscheidbarkeit von Genres hinsichtlich des musikalischen Aspektes Rhythmus untersucht. Da viele unterschiedliche Ansätze zur Rhythmusrepräsentation innerhalb von MIR existieren, wurden ausschließlich die drei Deskriptoren *Rhythm Patterns* (RP), *Statistical Spectrum Descriptor* (SSD) sowie *Rhythm Histogram* (RH) verwendet. Diese drei Deskriptoren definieren jeweils eine große Anzahl an einzelnen Merkmalen (*features*) zur Repräsentation des Rhythmus einzelner Musikstücke im weitesten Sinn. Zwei grundsätzliche Fragenstellungen wurden in der Masterarbeit durch eine eingehende empirische Analyse evaluiert. Die erste Fragestellung widmete sich der möglichen Beschreibung von Genres durch eindeutige, nur dem jeweiligem Genre zugeordnete, Rhythmusmuster. Diese Zuordnung wurde jeweils mit Hilfe von fünf verschiedenen heuristischen Berechnungsmodellen ermittelt, welche die Unterscheidbarkeit eines spezifischen Genres durch das jeweilig untersuchte Rhythmus-Merkmal ermittelt. Zu diesem Zwecke wurde eigens die Applikation *DiscriminationAnalyzer* entwickelt, mit dessen Hilfe beliebige Merkmale aus einer gegebenen Merkmalsmenge dahingehend untersucht werden können, ob und wie stark eine Unterscheidbarkeit von Genres bezüglich der untersuchten Merkmale gegeben ist. Vor allem die Verarbeitung beliebiger Deskriptoren, die Auswahl von verschiedenen Berechnungsmodellen und eine angemessene Visualisierung der Ergebnisse zeichnen diese Applikation aus. Die zweite Fragestellung beschäftigte sich mit der Evaluierung einer Merkmalsreduktion basierend auf diesem Unterscheidbarkeitspotential der einzelnen Merkmale im Zusammenhang mit der automatischen Klassifizierung von Musikstücken nach Genres. Die Reduktion von Merkmalen hat im Kontext der Klassifizierung eine besondere Bedeutung, da neben der Laufzeitoptimierung insbesondere auch der negative Einfluss des so genannten *Fluch der Dimensionalität* (*curse of dimensionality*) durch eine entsprechende Reduktion der verwendeten Merkmale möglichst minimiert wird. Als Kriterium für diese Merkmalsreduktion wurde die jeweilige Unterscheidbarkeit jedes einzelnen Merkmals bezüglich der Genres verwendet. Hierzu bietet *DiscriminationAnalyzer* eine integrierte Evaluation mittels beliebiger Lernalgorithmen an.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The digital representation of music has become a very important part of not only information science and technology but also to a huge number of people around the world who recognize the advantages of using digital music in comparison to traditional musical representations. Music collections of almost innumerable musical works have been collected, distributed and published for commercial or private use in recent years. The global community of users who want to capitalize on the public access to such music collections has immensely grown by now and, certainly, will steadily grow in the future.

Music Information Retrieval (MIR) has been founded as an interdisciplinary research field to encounter various problems which are directly related to the needs of automatic organization and browsing of such music collections. As already today numerous musical collections are publicly available over the World-Wide-Web, for example, automatic techniques to organize music collections are crucial to provide effective handling of digital music. Section 1.1 gives an introduction into motivation and main contributions of MIR and defines a selection of the most relevant research areas which are concerned by MIR. The successive section 1.2 lists the key questions and goals which are focused in this thesis. Eventually, section 1.3 presents the structural outline of the thesis.

1.1 Motivation

Digital information affects deployment and common use of computers and computer networks as it was unimaginable few decades ago. Due to the rapid development of hardware resources and network capacities the creation, processing and transfer of digital information like text, images, music or video have become very important and the size of legally, and unfortunately also illegally, accessible digital information has grown extensively. In recent years, large collections of digital information have been created by commercial vendors and private users. Considering the increasing popularity of Web 2.0 applications around the world, the proportion of information which is created by private users becomes significant and main requirement is the public availability over the world–wide web. Consequently, this can lead to the assumption that the World-Wide-Web

itself can be defined as a huge single and globally accessible collection of very different information contents. However, the immense and still growing scale of both the availability of collections containing arbitrary digital content and the World-Wide-Web itself raises eligible questions how to access particular content effectively and to provide an intuitive organization of such collections to everyone.

Since the early 1990's the importance of digital music has continuously increased. Intuitively, this development can be explained by two main arguments. First, music has a particular meaning to humans regardless of ethical, cultural or political origin. Almost every human being perceives music either directly by listening to or playing some specific musical work, or indirectly by watching movies, TV/Radio shows or even stores and shopping malls. Thus, music constitutes an important good for humans in a global sense. But why is music so popular? The answer obviously lies in the human perception itself. Humans do not only perceive the basic musical information alone, but also connects musical information with certain impressions influenced by emotions, moods or memories. This means that music is also a kind of medium for additional meta data. On the one hand, this meta data is highly subjective but, on the other hand, it is also a key factor for humans to decide which musical pieces will be listened in a particular situation.

Second, the digital representation of musical content has become an ideal way for editing, saving or transferring music. As efficient coding methods, like MP3 in particular, had been introduced, various applications have been developed to handle digital music content in whatever way possible. Until today large number of publicly available music databases or collections have been created[1] by either companies like music labels as well as retailers or private users. This increasing demand of digital music urges for effective ways to browse, organize and dynamically update music collections by the aggregation of sufficient meta data or annotations describing the content of musical pieces. Consequently, the key goal of MIR research is to develop new techniques to automatically extract content descriptions of musical pieces which can be directly used to organize music collections. Organization of musical collections means the corresponding categorization of every musical work and, respectively, or defining certain musical similarity measures to define relationships between musical works. But this obviously implies that some meta data about certain categories of similar musical works must be either annotated manually or automatically by some extraction process. In order to guarantee scalability of music collections, a manual annotation process can not provide a satisfying organization of music collection. Due to the limitations of manual annotation in terms of scalability, the question of effective access can only be solved by introducing new techniques which automatically perform musical categorization and compute similarity structures in connection to a scalable and extensible music collection. As a consequence, the performance of these methods strongly determines the efficiency and usability of music retrieval.

Music Information Retrieval (MIR) has been founded as the research area for giving solutions and methods to above questions. MIR constitutes an interdisciplinary research area and

[1] Unfortunately, the progress of handling digital music has also led to the still ongoing problem of illegal distribution of music mostly over the World-Wide-Web.

combines knowledge of diverse scientific fields like information retrieval and machine learning, musicology and music theory, audio engineering and digital signal processing, cognitive science, library science, publishing, and even law. The aims of MIR are to develop methods which, more or less, automatically organize large music collections and provide semantic structures to guarantee access, comparison or even research requirements of arbitrary musical works. A recommended introduction into MIR research is provided in [16] by Stephen Downie. Virtually all problems of MIR research include information retrieval and machine learning techniques which must be adequately adapted due to the very specific musical content representation. Again, Downie describes several important aspects of the musical content representation in [16] but also Byrd and Crawford [10] give a comprehensive insight into the problem of music representation. In order to demonstrate the complexity of MIR problems concerning musical representation, the following list introduces some important and widely recognized musical representation aspects affecting MIR research and its embedded challenges.

- *The existence of symbolic-based or audio-based or a mixture of both to represent music in digital form.* Symbolic representation assumes an abstract view of the underlying musical work which explicitly includes structural elements of the underlying music. Examples are either notation (scores, charts), event-based recordings (MIDI), or hybrid representations. Since the larger part of music recordings and in particular audio CDs and DVDs use composite audio signals to represent the musical content, audio-based representations of musical pieces are far more available. Basically, audio-based representation involves the discretization of composite audio signals into sets of measurement samples, whereas the composite audio signal describes the corresponding musical work physically. As a consequence, audio-based representation originally includes no meta data concerning musical structure, instrumentation or notation.

- *The complex interaction between music's pitch, temporal, harmonic, timbral facets.* One of the main MIR research areas is the extraction of these musical facets to get better representation of the musical content itself. The complexity of this problem depends on the musical representation. In particular audio-based representation does not directly contain any information about the pitch of single notes or temporal information. Thus, the annotation of those musical facets must be entirely extracted from raw audio signals involving the unpleasant ambiguous nature of musical facets.

- *Possible distortion of the original musical content during recording and, especially, encoding.* The limits of storage devices, network bandwidth or computational resources often require the encoding, i. e. compression, of the originally recorded musical representation. Unfortunately, most encoding methods introduce the possible generation of artifacts which are certainly not desirable but, eventually, unavoidable since a certain degree of compression has to be reached.

- *Music strongly depends on its cultural origin.* In order to characterize the problem of ex-

panding MIR research from usual Western music (popular or classical music) to other music styles, the specific and origin-depended treatment of different instrumentation, notation or scores must be considered as well. Thus, MIR must also recognize multicultural aspects of music and music representation.

Considering these relevant aspects of music and music representation, it becomes apparent that MIR and its main tasks have to challenge a variety of hard and complex problems. A comprehensive overview of MIR problems concerning applications in real world can be reviewed in a very interesting work of Byrd and Crawford [10]. As basic differences to other applications of information retrieval and machine learning are obvious, MIR tasks can not be solved by knowledge of information or computer science alone.

Although the core motivation of MIR is to develop effective retrieval techniques for digital music collections, continuing advance in various interrelated research areas must also be considered to describe MIR research entirely. In recent years, numerous MIR systems have been introduced to provide tools and methods to organize musical works on different levels and in different musical representations. However, MIR systems strongly depend on the chosen musical representation and further explicitly defined requirements. An overview of some interesting and exemplary MIR systems is described in [51, 52]. In spite of MIR research being always directly related to retrieval and access techniques, a wide range of subproblems have been formulated. The main MIR subproblems are:

- **Music search** describes approaches to find musical pieces based on a given query defined by users. Although the aim of music search always concerns the retrieval of specific music pieces, possible types of queries refer to different musical representations and meta data. Therefore, music search is divided into several search applications which depend on very different preconditions and properties. Possible types of search queries are introduced in table 1.1.

Query-by-example or *Query-by-similarity*	Searching by a given exemplary music sample
Query-by-humming/singing	Searching by a hummed or sung example
Query-by-category	Searching by a certain content category (e. g. style, genre, mood)

Table 1.1: Possible types of search queries for music search applications

- **Music identification** aims to assign specific annotations like for instance title or artist to given musical pieces. In other words, the application of music identification is to retrieve meta data of basically unknown musical pieces. The basic approach to realize this task is to build a unique content representation of musical pieces contained in a sample collection. To evaluate a particular identification query, the given query example will be compared with all sample representations. Consequently, the crucial task of music identification is

to effectively extract a unique representation. A common approach is the deployment of *acoustic fingerprints* which constitute uniquely generated code sequences from the original audio waveform of musical pieces. An important requirement of acoustic fingerprints is the robustness against transformations such as e. g. encoding, change of bit rate, etc.

- **Music classification** creates content-based annotations with respect to entire music pieces or segments of those pieces. Since humans categorize and organize music by a large number of different categories and attributes, an automatic approach to generate similar content-based annotation is crucial. Possible categories can be mood, genre, style classes but many more are imaginable. Unfortunately, those categories are related to each other in a very restricted sense. Moreover, classes or single categories are often ambiguous and therefore they can hardly be divided due to fuzzy borders. Considering the organization of music collection and the problem of effective and suitable access, correct annotation of human-defined categories are obviously important and, thus, have to be an intrinsic part of a MIR system. Also segments or elements of notations (e. g. scores) of musical pieces can also be annotated. Numerous practical examples can be found which need to annotate structural components or notations. For example, imagine the problem of identifying typical song-based structures like verses, choruses or bridges within a given music piece.

- **Music similarity** establishes relational structures within music collections. The similarity of music pieces is indicated by some predefined measure which is built upon a mathematical model based on a selection of musical descriptors. The choice of musical descriptors to be used is crucial in order to guarantee sufficient musical similarity structures. But which descriptors are actually effective to correctly represent musical similarity? This question is still to be answered yet. Music similarity is strongly related to *music recommendation* which aims to create user-specific play lists based on a selection of musical samples.

- **Music synchronization and music matching** describe the task of finding similar musical structures embedded in arbitrary music pieces. Typically, music collections contain for a single musical work various musical variations depending on different interpreters or artists, different recordings or digital representations. Effective browsing or retrieval in such music collections must consider these possible variations of musical works. Furthermore, music synchronization also provides techniques to identify similar parts of such interpretations despite of temporal or timbral differences. Exact time position synchronization according to separate digital representations of similar musical works has also become an interesting task of MIR research.

This thesis focuses on the assignment of musical genre categories to given musical pieces of underlying collections and will be further called *musical genre classification*.

Musical genres introduce a very popular and frequently used musical category to organize arbitrary music collections. In comparison to other possible musical categories, musical genres provide an intuitive understanding for categorization and are frequently used by humans to

organize music. Just think of record retailers or music libraries where offered music pieces are usually organized by corresponding musical genres. Due to the long use and its popularity, the automatic extraction of musical genres as a meta data of the underlying musical piece is of great interest to MIR research. This conclusion is particularly emphasized in the work of Aucouturier and Pachet [3]. Unfortunately, musical genres inherit an undesirable ambiguity due to imprecise genre definitions and boundaries as they rise through a complex interaction between public, marketing, historical, and cultural factors. By now a full description of musical genres on the base of music content alone can only be insufficiently solved.

Nevertheless, musical genre classification has matured to a top way for categorizing musical pieces, where two basic assumptions are essentially made within MIR research. First, for most current musical genres it can be assumed that the members of a particular genre share certain characteristics typically related to the instrumentation, rhythmic structure, and/or timbral content of the music. Second, genres can be related to each other in a hierarchical manner. This aspect can be specifically used by MIR techniques, for instance by hierarchical classification. Although a suitable hierarchical structure of genres (i. e. taxonomy) can not be automatically generated yet, the use of a user-defined taxonomy can increase the classification performance according to the genre definitions of the user itself.

This assumed descriptive power of musical genres is the initial point of two key questions raised in this thesis. The motivation of the fist question directly lies in the possible relation of some rhythmic content to a particular musical genre. “Do some specific rhythmic feature subsets exist which uniquely determine the genre of the corresponding musical piece?” To answer this question, a discriminant analysis will be applied based on music representations of three specific rhythmic descriptors. The discriminant analysis determines those features which significantly distinguish the correct genre of the corresponding musical piece. The second key question of this thesis examines the effects of using those discriminative features only for musical genre classification. Based on the results of the discriminant analysis, a feature subset selection approach will be evaluated which reduces the dimensionality by selecting discriminative features only. The following two aspects are evaluated in particular. First, which musical genres are related to a significant feature set reduction. Perhaps observable reductions can be even obtained for all examined genres? The second aspect deals with the question whether such a feature set reduction does affect the classification performance and in which way. The conclusions these two key questions aim to increase the understanding of genre-based rhythmic representation and usefulness of feature selection based on the discriminative power of class determination.

To summarize, MIR research develops techniques to provide effective ways to organize, to browse or to search large and scalable collections of digital music which are already frequently available today and become even more important in the near future. To meet the challenges of automatically creating meta data for annotation, appropriate descriptors of musical content must be extracted. In order to guarantee consistency and arbitrary scale of music collections, an automatic approach of describing musical pieces by certain meta data is crucial. Obviously, manual annotation can not sufficiently encounter these challenges. MIR research consists of several

interrelated subproblems which aim to offer possibilities to browse or autonomously generating play lists. For instance, music classification or music recommendation provide very promising prospects to interact within future music collections and applications. In recent years, MIR research has rapidly grown and due to the amazingly dynamic and creative MIR research community, it appears to be very likely that this progress will continue in the future.

1.2 Thesis contributions

The following list precisely formulates the contributions of this thesis:

- An examination of rhythmic descriptors which are introduced in [38, 42, 46, 47] empirically evaluates the question whether feature ranking based on the discriminative power of features to distinguish the correct musical genre is a promising approach for effective dimensionality reduction. Basically, ranking methods based upon the discriminative power of features for class determination have been widely used in information retrieval applications for many years. However, an employment of this approach on musical data in context of genre classification has not been adequately discussed yet. Another question of this examination tends to evaluate whether some rhythmic descriptors or patterns of rhythmic descriptors are significantly discriminative to particular musical genres.

- Based upon the results of the discriminative feature ranking, a straight forward filter-based feature selection will be examined in context of musical genre classification. An extensive evaluation presents conclusions concerning the usefulness and potential improvement in genre classification performance of this feature selection approach.

- In order to effectively perform a discriminant analysis on arbitrary music collections and various musical descriptors, a comprehensive application has been developed in MATLAB. The core of this application is an examination tool providing various heuristic discrimination models to calculate and to evaluate the discrimination of genres according to specific features. Successively, basic evaluations of given feature selections can be performed. As input both the popular dataset format ARFF [57] of the open-source data mining workbench WEKA and the more specific format SOMLib [49] are accepted and can also be used to save selected feature sets into the corresponding data format.

- As hierarchical genre classification constitutes another promising application of the discriminant analysis, another focus of this thesis is the implementation of a particular hierarchical genre classification within the open source machine learning WEKA workbench. The key goal of this implementation is to provide a generic learning algorithm based on arbitrarily defined class taxonomies. Together with an additional ensemble learning algorithm and other required components, the final implementation constitutes an extension of the WEKA workbench. An analysis of the classification performance achieved by this implementation is not included in this thesis, only the framework and aspects according to the

implementation are discussed.

1.3 Thesis outline

In order to investigate the research questions which are raised in this thesis, expertise and techniques concerning multiple research areas of computer science must be considered. Chapter 2 reviews particular methods and techniques originally developed in the research areas *machine learning*, *Information Retrieval* and its sub-domains *Text Information Retrieval* and *Music Information Retrieval*. An overview of fundamental approaches and definitions is given as well as references to relevant literature. Moreover, an introduction into three specific audio-based rhythmic descriptors known as Rhythm Patterns, Statistical Spectrum Descriptor and Rhythm Histogram is also given in this chapter. It should be noted that only these three audio descriptors are used for evaluations within this thesis. Eventually, the three standard benchmark music collections used in this thesis are outlined briefly.

Chapter 3 focuses on the main research contribution of this thesis which is the analysis of the discriminative power of features defined by rhythmic descriptors according to musical genre classification. On the one hand this analysis is based on the three rhythmic audio descriptors introduced in the preceding chapter. On the other hand five heuristic discrimination models are employed to estimate the discriminative power of particular features. The following heuristic discrimination models are used in this thesis: *Chi-square statistics*, *Information Gain*, *Gain Ratio*, *Balanced Information Gain* and *ReliefF*. Detailed definitions and summarizations of the properties of these heuristic discrimination models are given as well. The final conclusion gives answers to two basic questions concerning the existence of discriminative features or feature patterns with respect to single musical genres. The first question is dedicated to which features or feature patterns are actually discriminative according to the five heuristic discrimination models. The second question is given by how consistent are the results of those five calculation models for each of the three rhythmic descriptors and different music collections.

Chapter 4 evaluates a filter-based feature selection approach for musical genre classification based the results of the discriminative feature ranking analysis of the previous chapter. In order to compare the effectiveness of this feature selection approach, music pieces from three music collections are deployed to three separate learning algorithms – to the *Support vector machine* learner, to the *Decision tree* learner (J48) and also to the probabilistic learner *Naive Bayes*. These three learning algorithms have been chosen because they represent quite different approaches to learning and classification. Moreover, the crucial definition of an adequate problem-based similarity measure to compare instances within the feature space can be omitted because the SVM learner usually works well with the Euclidean distance measure and the other two learning algorithms do not use such a similarity measure at all. Contrarily, the classification performance of learning algorithms like for instance *Nearest neighbor* learners or *Gaussian mixture models* is strongly influenced by whether intrinsic similarity measure has been employed. The key conclusion at the end of this chapter is whether a notable feature set reduction with constant

or even increased classification accuracy can be obtained by this feature selection approach in comparison to use the complete feature set.

Chapter 5 presents an overview of two particular applications which are created to perform the evaluation tasks of this thesis. Most importantly, usage and functionality of the discriminative feature analysis tool *DiscriminationAnalyzer* are explained. The DiscriminationAnalyzer tool is developed under MATLAB and, furthermore, is built upon the powerful open-source, Java-based data mining workbench WEKA [57]. Another part of this chapter focuses on the implementation and usage of a hierarchical learning algorithm as well as a specific ensemble learning algorithm working on various feature sets simultaneously. Both learning algorithms are completely embedded into the WEKA framework.

Finally, a summarization of this theses as well as proposals for future work are given in chapter 6.

Chapter 2

Related Work

This chapter gives an overview of five major scientific research application areas and some particular considerations of using audio benchmark collections which were crucial to achieve all defined contributions of this master thesis. Each of the following sections includes a short topical introduction as well as an overview of *state-of-the-art* techniques and emphasizes references to relevant literature sources.

At the beginning, the sections 2.1 and 2.2 provide an overview of concepts and literature concerning important learning algorithms and feature selection methods. Since ensembles of learning algorithms play a major role in the applications of this thesis, this section reviews basic approaches and, in particular, the combination of multiple learning algorithms. Feature selection is also a fundamental part of this thesis and is introduced in this section. Feature selection involves methods to reduce the original, usually high dimensional feature space. Considering the assumption of adequate and meaningful feature selection, this feature space reduction can actually increase the classification accuracy. A crucial factor for the significance of discriminant analysis is the choice of an appropriate heuristic method. The aim of this heuristic is to estimate the actual discrimination value for a particular feature and the underlying classes. Section 2.3 refers to literature which introduce such methods. Section 2.4 focuses on musical genre classification and reviews relevant methods and approaches which solve this specific problem of music classification. Since rhythm is one of the most frequently used musical content descriptors, section 2.5 introduces some main aspects of extracting rhythmic descriptors. Both audio and symbolic representation can be the source of extracting quantitative information which represents rhythm. In this thesis, three certain rhythmic descriptors are being used as feature input and are discussed as well. Finally, section 2.6 refers to the importance of music collections and introduces the three music collection which are used to evaluate the key questions of this thesis.

2.1 Learning algorithms

In order to provide effective techniques for organizing and browsing large music collection, music information retrieval is fundamentally related to the general research areas of machine learning. In particular, machine learning provides effective learning algorithms which can be applied to a wide range of real-world applications regarding for instance information retrieval, pattern recognition or regression problems. These applications have at least one common aspect, namely, the effort to describe some complex problem which is usually influenced by various and mostly unknown factors. In computer science, the usual approach of solving a given problem starts with the explicit and precise mathematical formulation of the given problem model. Unfortunately, many real-world problems involve a high degree complexity since the functional relation between included factors and/or the factors themselves are unknown. Thus, an explicit mathematical formulation of such problems can not be defined at all. Actually, today's most interesting problems do have such a high complexity. Yet, approximated solutions of such problems can be estimated by directly simulating the transition from some particular input to the corresponding output by a particular approximation. This approach is called *learning* and is a crucial part of any machine learning method. The goal of learning is to find an approximation $\hat{y}_i = h(\mathbf{x}_i)$, $1 \leq i \leq m$ of some unknown function $y_i = f(\mathbf{x}_i)$[1] for each data point $(\mathbf{x}_i, y_i)$ of a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$. This dataset $\mathcal{D}$ is known as *training set* (or *learning set*) and the number m of included data points $(\mathbf{x}_i, y_i)$ is crucial to retrieve a good approximation of $f(\cdot)$. In some literature, $\mathbf{x}_i$ of a the data point $(\mathbf{x}_i, y_i)$ is frequently denoted as the feature vector and y_i the corresponding output. The function h represents the learning algorithm or *classifier*. If y is known for each available data point, the determination of h is called *supervised learning*. Otherwise it is called as *unsupervised learning*.

To get an overview of basic machine learning methods and techniques, the textbooks of Bishop [6] and Witten [57] are of particular interest. Both books include a detailed introduction into most popular learning algorithms like *Decision Trees*, *Support Vector Machines* (SVM), *Hidden Markov Models* (HMM) *Gaussian Mixture Models* (GMM) or probabilistic learners suchg as the *Naive Bayes*. More learning methods exist but these mentioned learning algorithms are employed most frequently in applications. Additionally, [57] also includes a detailed documentation of the machine learning workbench WEKA, which is an open-source, Java-based library and is frequently used within the MIR community.

Besides these two introductory machine learning references, a selection of literature regarding different aspects and general concepts of learning should also be mentioned. In most machine learning applications the assumption holds true that the classification accuracy obtained by a single classifier is lower than the corresponding accuracy archived by the simultaneous use of multiple classifiers. Such a use of multiple classifiers is known as *ensemble learning*. One possible justification of this assumption is that a single learning algorithm is more affected by the learning problems of *overfitting* or *underfitting*. The first problem means that the approximated function

[1] The unknown function $f(\cdot)$ is the mathematical synonym for some complex and unknown problem.

h depends too strongly on the training data. This leads to unsatisfying behavior to arbitrary real-world data, although the approximation error corresponding to the training data can be extremely low. The latter problem describes the inverse situation. The applied learning algorithm produces significant approximation errors in relation to the training data. One approach to avoid overfitted learning is to apply an ensemble of similar or different learning algorithms, where the output of each model will be accordingly aggregated to a final approximation. Several approaches for realizing ensembles are known and the work of Diettrich [14] gives a detailed insight into existing ensemble techniques. Kittler et al. [30] and Tax et al. [50] introduce a specific concept of ensembles in which the probabilistic outputs of learning algorithms are combined by applying a certain probability aggregation function, e. g. averaging, minimum or maximum probability. This ensemble technique has been implemented for this thesis as a specific WEKA extension and is presented in section 5.2. Unfortunately, this approach can only include such learning algorithms which output a probability density of the possible output realizations. Especially, SVM do not support this requirement originally but, fortunately, there exists a transition of the marginal output to probabilistic output according to the work of Platt [43]. Ensembles can also be used to split an originally multiple class problem into separate binary class problems. Allwein et al. [2] describe this important approach of reducing a multiple class problem to an ensemble of binary class problems.

2.2 Feature selection

Another important aspect of machine learning is the choice of features which are used as the quantitative input for the underlying problem. Learning strongly depends on the quality of features which means whether a certain feature type possesses sufficient power to describe the underlying data to be classified. Apart from the basic questions of using which feature type or feature extraction algorithm, the number of features can also significantly affect the approximation accuracy of learning algorithms. More precisely, using a large set of features does not necessarily imply better accuracy. Quite the contrary, the inverse effect of decreasing performance can often be observed. This effect is known as *curse of dimensionality* and basically means that the number of training data must grow exponentially with the number of used features. Thus, a large set of features in relation to few training data will actually decrease the approximation accuracy. The usual way of avoiding the curse of dimensionality is to select a certain number of most descriptive features by considering some criteria, all other features will be omitted during learning. This approach is commonly called *feature selection* (or *variable selection*). Guyon et al. [24] published a fundamental introduction to feature selection. In [31], Kohavi and John describe two alternative basic feature selection approaches which they denote as *filter-based* and *Wrapper* feature selection. The Wrapper feature selection refers to a subset of feature selection techniques wherein each candidate feature subset visited in the algorithm's search is evaluated by training and testing a learning model using only that feature subset. Consequently, an individual training set which must be independent from the training set of the actual classification realization

is required to evaluate a statistically reliable feature selection. The classification accuracy for the visited candidate subsets is used to guide the search to new subsets where forward selection and genetic algorithms are frequently used algorithms for finding new (better) feature subsets to evaluate. Contrarily, the filter-based feature selection does not involve any learning model but the search of a proper feature subset is guided by a particular measure. Such a measure can be for instance the correlation among the features, the dependencies between features and genres.

Basically, many feature selection studies concluded that a general recommendation of either the Wrapper of the filter-based feature selection can not be given but the actual choice of the feature selection algorithm depends on various decisive factors whereupon the dimensionality of the feature set and the applied learning model are most important. Nevertheless, the feature selection studies [25,31] proposed that the Wrapper feature selection achieves better classification accuracy as the filter-based feature selection based on popular learning models like Support Vector Machines if speed is not an issue. Thus, the key advantage of the filter-based feature selection is the fastness of the feature selection compared with the Wrapper feature selection. Wu and Zhang [58] particularly emphasized this key property of filter-based feature selection for high-dimensional numeric data sets. It is important to note that music classification often involves high-dimensional numeric data sets. Another important advantage is that a separate training set for feature selection is not necessarily required. In this thesis, only the filter-based feature selection is used. As proposed in section 2.3, the use of heuristic discrimination models to rank features represents a particular implementation of the filter-based feature selection by ranking every feature based on its discriminative power to distinguish correct class membership.

Considering music classification and, musical genre classification in particular, frequently rather large sets of features are extracted to describe the musical content. From the perspective of machine learning this content representation can actually lead to deteriorated classification accuracy, especially, if few training samples are available. Thus, feature selection must be considered. Chapter 4 of this thesis proposes a specific feature selection approach for musical genre classification based on discriminative ranking of features. Alternatively, Fiebrink et al. [20] suggest a feature selection approach based on feature weighting by estimating the contribution of every feature to the classification task. These estimations of according feature weights were done by a genetic algorithm implementation. Grimaldi et al. [23] compared the performances of the three filter-based feature selection approaches Information Gain (☞ see 3.2.2), Gain Ratio (☞ see 3.2.3) and Principal Component Analysis based on two different ensemble strategies and a *k-Nearest Neighbor* learning algorithm in terms of musical genre classification. Furthermore, Fiebrink and Fujinaga [19] summarize some very interesting conclusions of effective feature selection and emphasize existing pitfalls in music classification. In particular they pointed out in terms of the Wrapper feature selection that a feature subset must be verified with a separate test set. This means that the Wrapper feature selection actually requires a separate training and test set for feature selection.

2.3 Heuristics discrimination models

The basic idea of evaluate the contribution of a feature to distinguish the correct genre evolved from the use of impurity functions in the context of rule-based learning algorithms. Such learning algorithms use explicitly created rules based on feature values to determine a class label. The popular rule-based learning algorithm *Decision tree* was the first application in which various heuristic discrimination models had been applied as an approximation of the underlying impurity function. According to decision trees the impurity function formally defines the interdependencies among included features. In other words, the impurity function is a quality measure of how strong a particular feature depends on others. Considering the problem of finding an effective rule structure, the order of rules related to different features is crucial. If an insufficient order of features is chosen, the size of the rule structure will grow exponentially in the worst case – this effect is also called *combinatorial explosion*. Thus, the definition of an impurity function implies a systematical choice of feature order because it favors features with low interdependencies. In that sense the impurity function is directly related to the estimation of the genre discrimination because a feature which has low interdependencies also possesses more discriminative power in order to distinguish correct class labels.

Various heuristic discrimination models exist in to estimate the discriminative power[2] of features as well as to approximate the impurity function. The most frequently used models are probabilistic models based on the concept of mutual information which is contributed to a particular feature. In literature, such models are also frequently called *impurity* models or *impurity* functions. In the case of classification which implies that the target concept is a discrete variable, one of the best known impurity models is the *Information Gain* [26] which is often used to efficiently construct rule-based learning models like Decision trees. Unfortunately, Information Gain tends to overestimate multi-valued features because the estimate of the Information Gain also grows with the entropy of the features. In order to avoid this tendency, various normalization heuristics have been introduced like the *Gain Ratio* [45], the *Symmetrical Uncertainty* [58], the *Balanced Information Gain* [58]. Also the *Gini-index Gain* [8] as well as the *Chi-square* $\mathcal{X}^2$ and G statistics are used to estimate the quality of features.

Kira and Rendell [29] introduce another probabilistic but different model known as *Relief* which utilizes the nearest-neighbor algorithm to estimate the quality of features. Relief does not measure the uncertainty of a certain feature as entropy models do, but alternatively, it estimates how well the feature values distinguish between data points that are close to each other. In other words, if a small value change of a certain feature will cause a different class assignment, that feature probably possesses a significantly higher discriminative power for genre determination. Based on this work, Kononenko introduced the enhanced *ReliefF* measure [33] which has actually replaced Relief. A decisive theoretical and empirical study of Relief and ReliefF models can be reviewed in [48]. In [32], Kononenko also publishes an interesting evaluation of the heuristic discrimination models mentioned so far. In particular Robnik-Šikonja and Kononenko [48] pointed

[2] In this thesis, the quality and the discriminative power of a feature are synonymous terms.

out an interesting aspect of Relief and ReliefF. To estimate the quality of a specific feature, both algorithms also take into account the context of other features, i. e. the conditional dependencies between the features given the observed value. This inclusion of the feature dependencies is due to the intrinsic nearest-neighbor algorithm which uses the neighborhood of the input space to estimate the quality of a feature. Heuristic discrimination models based on impurity functions only use the correlation between the feature and the class disregarding the dependencies to other features. The authors of [48] concluded that the power of Relief and ReliefF is the ability to exploit information locally, taking feature dependencies into account, but still regard the correlation between the feature and the class.

Another heuristic discrimination model has been originally developed in context of text retrieval and is called *Attribute Discrimination* (or *term-discrimination* within text information retrieval). Contrary to the previous models, it is not based on probabilistic foundations, but rather, it determines the discriminative power of a particular feature according to its contribution in order to reflect changes on the average class similarity. The average class similarity can be seen as a measure to compare the affinity concerning location and expansion of included class structures within the multi-dimensional feature space. The works [11,18] give a detailed insight into the definition of the Attribute Discrimination value measure.

In this thesis, the heuristic discrimination models Chi-square, Information Gain, Gain Ratio, Balanced Information Gain and ReliefF were used. The precise definition and main properties of these calculation models can be reviewed in section 3.2.

2.4 Musical genre classification

Musical genres are widely used and very popular descriptors for categorizing and organizing music collections. Contrary to other existing musical descriptors, genres are directly related to the way humans do browse and select musical pieces within large music collections. Consequently, genres are very suitable to describe the musical content and musical genre classification, which aims to automatically assign genres, represents an important application of MIR research.

Unfortunately, musical genre definitions are often very fuzzy and, therefore, clear borders can not be recognized to separate genres from each other accurately and uniformly. This genre ambiguity strongly depends on the power of those descriptors which are used to represent musical content. Considering the preferred way of representing musical content by using audio-based descriptors containing pitch, tempo or timbre facets only, the raise of doubts concerning effective classification is justified. The fundamental question is whether audio-based descriptors are really sufficient for describing musical genres. In general, there are limitations to automatically define and classify musical genres by using audio-based descriptors only. Many genres do not really differ on a strict musical level but rather on cultural origin due to interpreter or artist. Even language and specific forms of lyrics can determine genre membership. A direct conclusion of this possible genre fuzziness is that non-audio content descriptions should also be taken into account to correctly define and classify musical genres. In spite of the limitations of automatic

genre classification, it still possesses promising potential for organizing music collections because it provides an intuitive and very common approach for humans. Several scientific works indicate this unavoidable ambiguity of musical genre classification and give some interesting insight. In [3], Aucouturier and Pachet introduce an interesting view of genre ambiguity by suggesting that every genre always consists of an intentional and an extensional concept which both do not coincide in real world. Despite of that they actually emphasize the importance of genre classification. Additionally, McKay and Fujinaga [40] formulate an interesting argument in which they also emphasize the importance of musical genre classification, although little progress has been made in recent years due to the fuzziness of genre descriptions.

Another interesting aspect of musical genre classification is the application of genre taxonomies to perform hierarchical classification. Basically, hierarchical classification has two very useful effects in context of genre classification. First, it provides an enhanced way to browse musical collections by starting from some very general genre and continuously refine the music search by choosing some more specific sub-genre. Second, it introduces a promising way to divide a flat but complex genre classification problem into smaller subproblems. These subproblems potentially yield better classification accuracy individually. But in general, a significant improvement of classification accuracy can not always be expected in comparison to flat genre classification approaches. However, the hierarchical approach of genre classification can improve the overall needed calculation time because each classifier has usually to deal with a more easily separable subproblem. Moreover, every classifier can use an independently optimized feature set where feature can be respectively reduced.

Obviously, the definition of a genre taxonomy is mainly responsible for successful hierarchical genre classification. In order to underline the importance of genre taxonomies, Pachet and Cazaly [41] describe a conceptional guideline to build effective genre taxonomies. In [34], Tao Li and Mitsunori Ogihara introduce a basic outline of using a taxonomy-based musical genre classification. Alternatively, Burred and Lerch [9] show another hierarchical classification approach based on a more complex taxonomy which includes speech content along with music content. Contrary to those works, Brecheisen et al. [7] use an individual feature subset selection at every node of the taxonomy instead of using the same the feature subsets at every node. Consequently, the actual feature set selection depends on those genres only which are incorporated by the respective node. All these works suggest that the primary choice of a more general genre already determines the possible set of sub genres which can be finally assigned. In other words, if a certain branch of the taxonomy is chosen, all genres located in different branches will be omitted. Yet, this restricted approach can significantly deteriorate classification accuracy. Thus, a less rigorous approach in which transitions exist between genres of different branches of the same refinement level probably promises better accuracy. DeCoro et al. introduce such an approach in [12] in which the inter-branch transitions are established by a Bayesian network.

In recent years, many different approaches to musical genre classification have been introduced. Best known due to the ground-breaking results are the works of Tzanetakis and Cook [54] and Bergstra et al. [5]. Tzanetakis and Cook achieved remarkable genre classification results by

using several different feature sets together with a hierarchical genre classification approach. Those feature sets correspond to different music content descriptors describing timbral texture, rhythmic content and pitch facets. The classification results proved that the combination of features representing different content description can actually improve the class separation of the underlying problem which leads to an improvement of the classification accuracy according to a real-world music collection. Since the different feature sets have been employed this is no contradiction to the previous conclusion that too many features (but from the same feature set) can deteriorate the classification performance due to the curse of dimensionality. Flexer et al. [21] also pointed out the classification improvement by using different feature sets together with a flat musical genre classification system.

Instead of classifying on a song domain only, Bergstra et al. introduce an approach in which every song is divided into several segments and genre classification is applied on those segments individually first. In a second step those partial classification results are aggregated to assign a genre label to the entire song itself. Tao Li publishes comparative studies [35,36] based on Tzanetakis and Cook in which influences of audio feature and learning algorithms according to musical genre classification are examined. In [34], Li also shows significant improvement by using Support Vector Machines in musical genre classification. Another work which definitely approves the positive affects of SVM in musical genre classification is done by Xu et al. in [59]. Although they use rhythmic descriptors and standard *Mel-Frequency Cepstral Coefficients* (MFCCs) together with a rather simple classification approach, remarkable improvements are obtained compared with other learning algorithm. Besides this excellent performance of SVM in musical genre classification, another advantage is that SVM can be applied without any further adoptions or modification. Contrarily, Hidden Markov Models require definitions of a specific transition structure and probabilities or Gaussian Mixture Models inherit the assumption of a certain similarity metric. Unfortunately, these assumptions strongly depend on the musical descriptors which are actually used.

As audio-based representations are far more available, most works concerning musical genre classification use combinations of audio-based content descriptors. Symbolic content descriptors extracted from MIDI representations are rather seldom used, in spite the fact that the inclusion of symbolic descriptors introduces better determination of musical facets like timbre of tempo. As an example of the use symbolic descriptors, McKay and Fujinaga [39] develop a remarkable approach by using several musical features from symbolic MIDI. The classification is done hierarchically whereas different sets of features are used on different taxonomy levels. It should be noted that a comparison of classification systems where some systems work with audio-based descriptors or some other systems use symbolic descriptors are invalid because of the intrinsic difference of these two representation types.

2.5 Rhythmic descriptors in MIR

According to Western music and to Western musicology, the rhythm component of a musical work contains numerous different temporal indicators. As mentioned in [16], temporal indicators are for instance tempo, meter, the duration of pitches[3], the duration of harmonic and accents but many more temporal indicators affect the rhythmic perception of humans. Rhythm constitutes an important musical content descriptor because humans perceive even minor rhythm differences very well. As a consequence various approaches to extract the rhythmic content have been proposed in MIR. As outlined in [22], two separate concepts can be basically recognized for extracting rhythmic information. One group of extraction algorithms focuses on the measurement of metrical elements to describe the underlying rhythmic structure. Popular metrical elements are for instance tempo, fastest pulse, quantized durations or tempo variations. Another group of rhythm descriptors is tightly linked to physical properties of the audio signal itself. This means that raw signal descriptors, e. g. frequency, modulation or amplitude properties, are used to describe the rhythmic structure of a corresponding musical work. Also the identification of continuous periodical frequency components can be used as a descriptor. Unfortunately, rhythmic descriptors based on physical properties tend to represent the rhythmic content less explicit.

Apart from the question which extraction approach yields the better description of the rhythmic component, many researchers emphasize the importance and effectiveness of using rhythmic descriptors in various music classification applications including musical genre classification. Although musical genres inhere an ambiguous definition due to the diversity of influences which actually determine a certain genre assignment, rhythmic content propose a very valuable contribution to identify genres sufficiently.

Considering rhythmic descriptors in a narrow sense, Dixon et al. [15] conclude that only a small selection of metrical elements including tempo, beat and measure are sufficient to describe rhythm and to predict the genre of the musical work. The metrical representation is based on the analysis of relationships of detected periodicity patterns. Those extracted periodicity patterns establish a certain metrical hierarchy containing the selected metrical elements for rhythmic description. The effectiveness of this approach is verified by using a set of standard and Latin dance music. Based on these results, the work of Gouyon et al. [22] propose an interesting evaluation concerning the relevance of different rhythmic descriptors for predicting genre labels. They conclude that a tempo descriptor and a set of 15 Mel-Frequency Cepstrum Coefficients descriptors are actually significant for their genre classification application. This conclusion is verified by classifying a set of standard and Latin dance music, almost the same set of dance music as in [15]. The use of the tempo descriptor yields a classification accuracy over 80 %, whereas the rhythmic representation by the set of MFCC descriptors gives the best result with an accuracy of 90 %.

Another relevant approach of defining a descriptive representation of the rhythmic component is to transform fluctuations according to loudness sensation of selected frequency bands into

[3] The supposition is that metrically prominent pitches are longer in duration

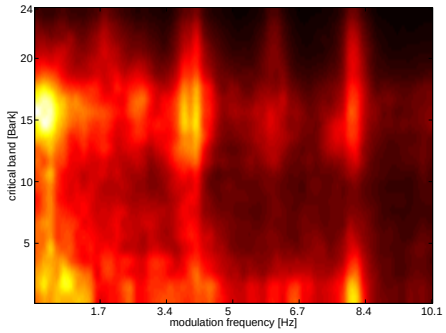
a time invariant representation. Since loudness fluctuations represent the directly perceivable part of the rhythmic component, descriptors based on this approach can also be assumed to describe rhythm in a narrow sense. According to this approach, the *Rhythm Pattern* (RP) feature set has been introduced in [42,46] as a time-invariant representation of loudness variations. In [47], Rauber et al. enhance the basic version of Rhythm Patterns by also incorporating psycho-acoustic phenomena of human audio perception. Two further rhythmic descriptors have been presented by Lidy et al. [37,38] which are basically based on the very same extraction algorithm as Rhythm Patterns. These descriptors are known as *Statistical Spectrum Descriptor* (SSD) and *Rhythm Histogram* (RH). It is worth noting that these three rhythmic descriptors are being used in this thesis only. Further details concerning the processing of these descriptors are discussed in the following subsections. An extensive comparison of the classification performance based on the deployment of these three rhythmic descriptors is given in et al. [38].

2.5.1 Rhythm Pattern

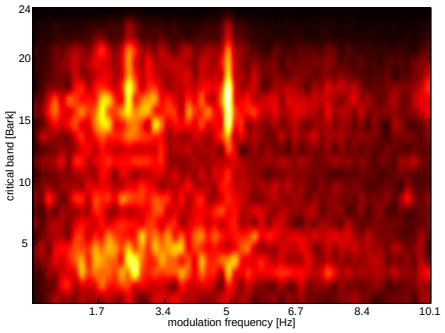
The basic idea of the Rhythm Pattern descriptor is the quantification of loudness sensation fluctuations in a time-invariant description. Since the human perception of rhythm strongly depends on the composed loudness levels and variation of the perceived musical piece, The Rhythm Pattern descriptor can be defined as a rhythm representation in a narrow sense. The Rhythm Pattern descriptor is defined as a matrix representation of the dependency between certain critical frequency bands and corresponding frequency amplitude modulations. These critical frequency bands are directly obtained from a psycho-acoustic frequency transformation of the original frequency spectrum. This frequency transformation is motivated by the specific perception of loudness and energy of certain frequency bands by the human auditory system. In order to obtain a time-invariant description of the rhythmic structure representing an entire musical piece, a modulation frequency is extracted over the time range of the critical bands. Contrary to the raw amplitude energy of frequency bands which is time-depended, the modulation frequency is time-invariant.

The Rhythm Pattern descriptor is first introduced in the works [42,46] in context of a musical jukebox system. The musical organization of this jukebox is implemented with Self-Organizing Maps on the basis of rhythm patterns. In [47], the original processing of Rhythm Patterns has been drastically improved by applying a psycho-acoustic frequency transformation based on the Bark scale. Eventually, Lidy et al. [38] give a detailed description of the entire extraction process of the Rhythm Pattern descriptor with an evaluation of the influence of the individual processing steps.

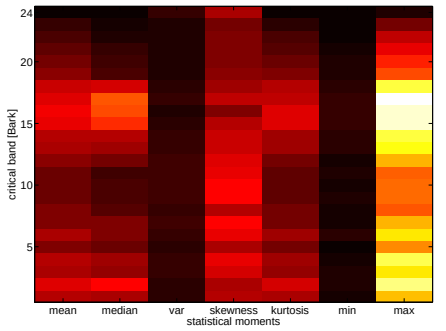
Nevertheless, a short overview of the extraction algorithm of Rhythm Patterns should be given here. The quantization of the original audio source into the final descriptor is organized in two separate parts. The first part concerns the computation of a frequency spectrum by employing a Fast Fourier Transformation (FFT) with overlapping Hanning window. The obtained frequency components are grouped into 24 critical frequency bands according to the Bark scale. Additionally, some further psycho-acoustic enhancements are being applied. First, a transforma-



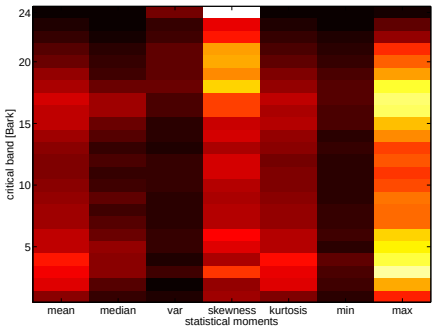
(a) Rhythm Pattern descriptor (RP)



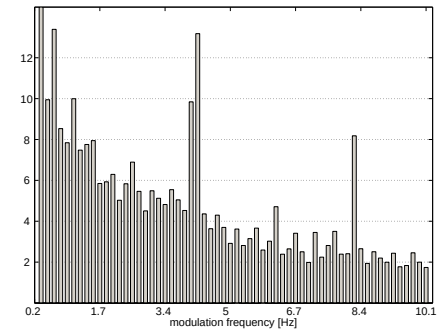
(b) Rhythm Pattern descriptor (RP)



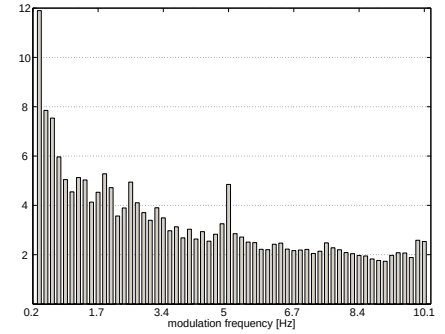
(c) Statistical Spectrum Descriptor (SSD)



(d) Statistical Spectrum Descriptor (SSD)



(e) Rhythm Histogram descriptor (RH)



(f) Rhythm Histogram descriptor (RH)

Figure 2.1: Specific representations of the rhythmic content of two music pieces related to two different musical genres by the descriptors RP, SSD and RH. The descriptors of figures (a), (c) and (e) are extracted from an interpretation of a classical music piece included in the ISMIR 2004 Genre (☞ see 2.6.3) audio collection. Contrarily, the descriptors of figures (b), (d) and (f) are computed from the rock song *Rotten flowers* included in the ISMIR 2004 Genre (☞ see 2.6.2) audio collection.

tion of the spectrogram into decibel scale is employed. In a successive step, the soundness levels are further transformed into Phon scale. The last step concerns the eventual calculation of the specific soundness sensation per critical band (Sone scale). The resulting "Bark scale Sonogram" describes the loudness sensation of human hearing and determines the loudness perception in relation of the underlying frequency.

The aim of the second part is to calculate a time-invariant representation of the enhanced spectrogram which has been obtained from the first part of this algorithm. In order to achieve a time-invariant representation, the time-depended amplitude energy of every critical band is converted into a spectral representation by the application of the FFT. The resulting spectral coefficients describe amplitude energy as amplitude modulations which are actually time-independent. It is worth noting that the amplitude modulation frequency range is defined up to 10 Hz because modulation frequencies above this boundary are mostly related to arbitrary noise (roughness) rather than useful rhythmic content. As the modulation frequency range is divided into 60 equally spaced samples, the entire descriptor contains $24 \times 60 = 1440$ features.

Figure 2.1(a) illustrates the rhythmic representation of a classical music piece according to the Rhythm Pattern descriptor. Due to the high polyphonic composition of orchestral classical music and the variation of tempo according to specific instruments, a wider region of high energy modulation frequencies related to numerous critical bands can be identified. Contrarily, figure 2.1(b) visualizes the rhythmic representation of a rock music piece in which a clearer impression of the underlying rhythm can be observed due to only few critical bands including high modulation frequencies.

In terms of the effectiveness of the Rhythm Patterns, Lidy et al. [38] present a detailed evaluation according to musical genre classification. Two related rhythmic descriptors are also compared which are the Statistical Spectrum Descriptor and Rhythm Histogram. This evaluation shows that classification results based on the Rhythm Pattern descriptor which are extracted form the GTZAN music collection (☞ see 2.6.1) are competitive to results obtained by Tzanetakis [53] or by Li [35]. Even the use of other music collections also implies the effectiveness of the Rhythm Pattern descriptor. However, the best classification results are obtained by the combination of the Rhythm Pattern and Rhythm Histogram descriptors.

2.5.2 Statistical Spectrum Descriptor

The extraction of the Statistical Spectrum Descriptor (SSD) is directly derived from the original Rhythm Patterns algorithm [47] and, therefore, also intends to describe the perceived fluctuations of loudness sensations. Basically, the raw signal data is transformed into a frequency spectrum by a Fast Fourier Transformation with a Hanning window. Additionally, a psycho-acoustic transformation of this spectrum into 24 critical bands is consecutively applied as proposed by Lidy et al. [38]. Similar to the Rhythm Patterns, the psycho-acoustic transformation groups all included frequencies into 24 critical bands according to the Bark scale. This psycho-acoustic transformation contains the same processing steps comparing with the Rhythm Patterns. Thus, the spectrogram is transformed into decibel scale at first. Successively, transformations into Phon scale and Sone scale are performed to get a frequency representation according to the 24 critical Bark bands.

Contrary to the processing of the Rhythm Pattern descriptor, the Statistical Spectrum Descriptor is composed of seven statistical components which are calculated for every Bark scale critical frequency band. This statistical representation provides a very compact representation

of rhythmic influences in relation to the respective critical band energy. The following statistical moments have been chosen to describe the rhythmic band fluctuations: *mean*, *median*, *variance*, *skewness*, *kurtosis* and *min- and max-value*. Thus, the Statistical Spectrum Descriptor contains $24 \times 7 = 168$ features for all 24 critical bands. Successively, a representation of the entire music piece can be aggregated by calculating the mean or median of the descriptors for every extracted part of the track where every part has a duration of 6 seconds. Figures 2.1(c) and 2.1(d) demonstrate exemplary representations of a classical and a rock music piece by using the Statistical Spectrum Descriptor.

2.5.3 Rhythm Histogram

In comparison to Rhythm Patterns and Statistical Spectrum Descriptor, the idea of the Rhythm Histogram description is to aggregate the rhythmic information by building modulation frequency bins. In order to build a descriptive histogram, the sample magnitudes of all 24 critical bands are summarized into 60 modulation frequency bins. Consequently, a Rhythm Histogram intuitively constitutes the relation of *rhythmic energy* per modulation frequency.

The extraction of rhythmic histograms is also similar to the first part of the Rhythm Patterns algorithm and is illustrated in [37]. The Rhythm Histogram descriptor consists of 60 bins which aggregate energy on the entire available frequency ranges for individual modulation frequency between 0.2 and 10.1 Hz. It should be noted that a specific rhythmic representation of a given musical piece is obtained by calculating the median of the histograms of every 6 second segment, Thus, the Rhythm Histogram descriptor contains 60 features. Figures 2.1(e) and 2.1(f) visualize an exemplary representation of a classical and rock music piece by Rhythm Histogram descriptors.

2.6 Audio collections

Openly available music collections are crucial to compare the effectiveness of different music classification systems. Since for most music classification systems a theoretical evaluation is not appropriate or even feasible, an empirical evaluation is the only way to compare such systems with respect to classification accuracy and computation time. Consequently, the creation of carefully selected music collections, which resemble most possible to "real-world", assures reliable benchmark tests. But the following requirements must be considered at least in order to build proper music collections:

- *A sufficient number of included music pieces.* The actual number depends on the specific application but, eventually in order to closer resemble "real-world" applications, the more music pieces are included the better it is.
- *Choose all music pieces systematically.* According to musical genre classification this means that every genre incorporates a sufficient number of related music tracks within the corresponding music collection.

- *The encoding of included music pieces must be considered.* Although music classification systems should be designed to be robust against artifacts derived from the specific encoding, i. e. compression of the input data, a benchmark collection should only contain equally encoded musical pieces. Artifacts constitute specific alterations of original musical content due to the content compression. Basically, the inclusion of artifacts should be avoided in a benchmark collection at all.

- *The duration of music pieces is also important.* Sufficient representations of underlying musical contents significantly depend on the respective duration of included musical pieces.

Since benchmark collections are used in music classification, i. e. a supervised approach, additional annotations for every included music piece must also be provided. Annotations can be assigned to entire music pieces or even to segments or musical facets of a single music piece. This annotation is usually called *ground truth*. It is worth noting that among the MIR community the MIREX project [17] is the most important try for providing benchmark music collections according to various music classification applications and other tasks. MIREX was initiated by the IMIRSEL group [27] led by Stephen Downie and his team at the University of Illinois at Urbana-Champaign, US.

Since this thesis only concentrates on musical genre classification, three distinctive musical genre music collections are reviewed in the following part of this section. These three music collections are the only ones employed in the thesis and a compact description is given in table 2.1.

Name	**Encoding**	**Genres**	**Pieces**	t_{Sample}	$t_{Overall}$
GTZAN	au, 22 kHz, mono	10	1000	30 sec	05:20
ISMIR 2004 Genre	mp3, 44 kHz, stereo	6	1458	full length	18:14
ISMIR 2004 Rhythm	RealAudio	8	698	30 sec	05:39

Table 2.1: Overview of three benchmark music collections which are employed in this thesis. The total duration $t_{Overall}$ of each collection is measured in [hh:min].

2.6.1 GTZAN

The GTZAN musical genre collection is named after its creator George Tzanetakis who introduced this collection in his PhD thesis [53] for evaluating musical genre classification systems. It contains 10 popular musical genres and provides an equal number of music pieces for every genre. Considering the preprocessing of included music samples, the originally uncompressed music data was reduced to mono and was sampled at a sampling frequency of 22 kHz. In order to avoid some unwanted effects like lead-in or lead-out, a 30 seconds segment from the center of each song was extracted. Since many works in MIR employ the GTZAN music collection to evaluate the classification performance, this collection is chosen to be the main benchmark collection of this thesis. The upper section of table 2.2 offers a detailed overview of all included genres.

Class name	Absolute Frequency	Relative Frequency
Blues	100	0.1
Classical	100	0.1
Country	100	0.1
Disco	100	0.1
Hip hop	100	0.1
Jazz	100	0.1
Metal	100	0.1
Pop	100	0.1
Reggae	100	0.1
Rock	100	0.1

(a) GTZAN

Class name	Absolute Frequency	Relative Frequency
Classical	640	0.44
Electronic	229	0.16
Jazz & Blues	52	0.04
Metal & Punk	90	0.06
Rock & Pop	203	0.13
World	244	0.17

(b) ISMIR 2004 Genre

Class name	Absolute Frequency	Relative Frequency
ChaChaCha	111	0.16
Jive	60	0.09
Quickstep	82	0.12
Rumba	98	0.14
Samba	86	0.12
SlowWaltz	110	0.16
Tango	86	0.12
VienneseWaltz	65	0.09

(c) ISMIR 2004 Rhythm

Table 2.2: Classes of the benchmark music collections GTZAN, ISMIR 2004 Genre and ISMIR 2004 Rhythm.

2.6.2 ISMIR 2004 Genre

At the *ISMIR 2004 Audio Description Contest* [28] which was operated by the Music Technology Group (MTG) of the University Pompeu Fabra of Barcelona, Spain, this music collection was introduced to compare algorithms dedicated to the particular music classification problems genre classification, artist identification and artist similarity. The contest organizers made available a training and a development set with 729 musical pieces each before the contest. This thesis used the combined set of 1458 musical pieces for the evaluations. According to genre classification the ISMIR 2004 music collection consists of 1458 music pieces unequally distributed over 6 popular genres. Since musical genres are never represented by equal numbers of included music pieces in real world, this collection simulates this observation by deliberately defining an unequal genre distribution with favor to the classical genre. The middle section of table 2.2 lists all included

genres and the corresponding amount of related music pieces. Concerning the compilation of the ISMIR 2004 genre collection, full length music pieces were used which were encoded in 128 kbps, MP3 stereo format sampled at a frequency of 44 kHz.

2.6.3 ISMIR 2004 Rhythm

Another music collection was introduced for the *ISMIR 2004 Rhythm Classification Contest*. Since this music collection contains pieces of Latin and Ballroom dance music only, the aim of the contest was to compare algorithms for automatic classification of the 8 defined rhythm classes. The organizers divided the total number of 698 musical pieces into a training and test set containing 488 and 210 pieces respectively. This thesis employed all 698 musical pieces for the evaluations, where all pieces are approximately 30 seconds long. The defined rhythmic classes are listed in the bottom section of table 2.2. Again, an unequal genre distribution was deliberately chosen to simulate "real-world" music collections. The collection includes 698 rhythmic music pieces, The original data was originally fetched in Real Music format with a sampling rate of 22 kHz. This music collection is publicly available over the website of BallroomDancers.com [4] and a complete list of the musical pieces used in the Rhythm Classification Contest is available at the ISMIR 2004 contest website [28].

Chapter 3

Discriminant analysis of rhythmic descriptors

This chapter focuses on the discriminant analysis of audio-based rhythmic descriptors in order to distinguish correct genres in terms of musical genre classification. The analysis is based on five different heuristic discrimination models to estimate the discriminative power of rhythmic features according to Rhythm Patterns, Statistical Spectrum Descriptor and Rhythm Histogram. These models also represent different approaches to express the discriminative power of variables. Usually, statistical models are used to define the density discrimination of a random variable by another variable due to statistical variable interdependencies. In this chapter, the heuristic discrimination models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain are applied which all actually represent a specific realization of the impurity function. Additionally, the ReliefF model is used as the fifth calculation model, since it utilizes a different approach to estimate the discriminative power of a feature.

The key goal of this chapter is to examine how the different calculation models performs on different music collections and whether a consistency of the discrimination values according to correlated genres can be concluded among the heuristic discrimination models. Another important goal of this analysis is whether specific features related to each of the three descriptors express consistent genre discrimination or not. All computations of the discrimination analysis will be done based on one-vs.-rest genre situations, i. e. binary class situations, only.

Section 3.1 gives insight into fundamental assumptions of the importance of features for classification and, furthermore, emphasizes meaning and significance of discriminant analysis in terms of musical genre classification. All five heuristic discrimination models which are employed to estimate the discriminative power of features are reviewed in section 3.2. The experiment environment and all relevant experiment results are discussed in section 3.3. Eventually, section 3.4 summarizes key conclusions based on the presented results for music genre classification.

3.1 Overview

In most classification problems, the intrinsic interaction between possible feature realizations and corresponding class values is very complex because class regions are usually inhomogeneous projected within the feature space. Unfortunately, if class regions are not clearly separable then the performance of the employed learning algorithm will deteriorate due to the poor approximation of the real but unknown discrimination function. Various reasons are responsible for obtaining such complex class regions, and the number of used features is definitely not the only factor which has to be considered. Actually, the inclusion of too few features leads to imprecise prediction of class regions which do not correlate with real world. According to the "curse of dimensionality", a proper approximation of class regions can also suffer by taking into account too many features unless the size of the training set will grow accordingly. Interactions between features also influence the complexity and accuracy of class regions in terms of the classification problem.

In order to guarantee a proper number of features and to avoid problems due to the "curse of dimensionality", feature selection represents a frequently used approach which is introduced in section 2.2. The goal of feature selection is to rank the original set of features by the respective contribution of every feature in terms of how good a single feature can distinguish the correct class. Unfortunately, a precise determination of such feature contributions is not feasible in sufficiently complex classification problems and therefore an approximation must be found instead. Basically, two concepts are used to obtain the proper feature ranking according to the discriminative power of every single feature. First, the feature selection can be performed by attempting to identify the best feature subset to use with a particular algorithm. This concept is known as the *Wrapper feature selection* and is extensively reviewed in [31]. Although the ease of implementation and good quality of the yielded feature selection are on the positive side of this approach, calculation time considerations and the need of separate data sets for training and evaluating must also be considered. Especially in case of a restricted number of training data in

first place a Wrapper feature selection is not suitable to yield an effective feature selection.

Thus, the second feature selection approach evaluates the contribution of every feature to distinguish the correct class value from the data alone. In [31] this approach is denoted as the *filter-based feature selection* and also the Wrapper feature selection is compared with this approach. Contrary to the Wrapper feature selection, a feature subset identified by the filter-based feature selection is independent from any learning algorithm. Thus, the key advantage of this approach is that the approximation of the feature contributions can be performed on the very same training set which will be successively used by the learning algorithm. Since no additional learning algorithm is applied in the filter-based feature selection, a statistically independent data set for feature selection is not required. Moreover, the calculation time of most implementations is significantly lower than in the case of wrapper implementations.

However, for both feature selection holds that the actual discriminative power of a feature can only be approximated. In case of the Wrapper feature selection this is obviously due to the intrinsic use of learning algorithms for feature evaluation. In terms of the filter-based feature selection various heuristic discrimination models offer an effective way to evaluate the contribution of features. In this chapter, five heuristic discrimination models are applied to estimate the discriminative power of features extracted from three different music collections in terms of musical genre classification. These five heuristic discrimination models are: *Chi-square statistics*, *Information Gain*, *Gain Ratio*, *Balanced Information Gain* and *ReliefF*. Subsection 3.2 gives precise definitions for each model.

The results of these heuristic discrimination models rate the contribution of a feature to distinguish classes. Besides the key goal of selecting and ranking an effective feature subset for successive learning, those results can also be used to examine possible correlations between certain features and classes. In terms of musical genre classification that implies the interesting question whether some particular features do significantly represent a single musical genre better than others. This question actually promises very interesting conclusions for representing genres or for separating genres from each other. Let me introduce a short example to emphasize the meaning of such feature-genre correlations.

Consider a large music collection which uses rhythmic descriptors to describe the musical content properly. This means that the rhythmic component of every musical piece is defined in some way. Additionally, tag inf ormation like genre, style or mood is also provided by the music collection. A very interesting question is whether a member of one of these musical tags can be directly related to a specific rhythmic pattern or a set of rhythmic pattern at least. A more intuitive formulation of this question can be given as follows: "*Can some specific musical genre be sufficiently described with particular types of rhythm which are represented by the underlying rhythmic descriptors?*". If such rhythmic patterns actually correspond to different musical genres musical genre classification could be adapted accordingly to gain better classification accuracy. But also from the view point of musicology, the possible connection between specific rhythmic content and musical genres opens interesting perspectives for a better description of such genres.

Section 3.3 focuses on an extended discriminant analysis due to the result of those five

heuristic discrimination models in order to conclude such promising feature-genre correlations. Three different audio-based rhythmic descriptors are used to represent the rhythmic content of musical pieces which are contained in the musical collections GTZAN, ISMIR 2004 Genre and ISMIR 2004 Rhythm. Those rhythmic descriptors are Rhythm Patterns, Statistical Spectrum Descriptor and Rhythm Histogram and are reviewed in subsection 2.5. Moreover, each of the included musical collections is described in subsection 2.6. An extensive analysis of discrimination calculations with the five heuristic discrimination models and the three introduced music collections GTZAN, ISMIR 2004 Genre and Rhythm is given in section 3.3. The key goal of this analysis is to conclude whether particular genre-specific feature patterns actually exist. A possible application of genre-specific feature correlations is hierarchical genre classification where decisions between specific genres occur frequently.

Eventually, the results of the discriminant analysis can also be used to select only those features which actually possess a particular minimum class correlation. An empirical evaluation of this feature selection approach is given in chapter 4.

3.2 Heuristic discrimination models

The key idea of heuristic discrimination models is to estimate the contributions of a random variable to predict realizations of another random variable. In terms of an arbitrary classification problem, most discrimination models can be reformulated to the estimation of the discriminative power of some feature in order to distinguish the correct class. The underlying dataset $\mathcal{D}$ of a classification problem can be defined as $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m)$ is the set of data instances and the target vector $\mathbf{Y} = (y_1, y_2, \cdots, y_m)$ constitutes the corresponding class assignments. A single data instance $\mathbf{x}_i \in \mathbb{R}^n$ with $1 \leq i \leq m$ is formally defined as a vector $\mathbf{x}_i = (x_1, x_2, \cdots, x_n)$. The feature set is denoted by the set $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n\}$, while a certain attribute $\mathbf{a_j} \in \mathcal{A}$ with $1 \leq j \leq n$ is related to $\mathcal{D}$ by $\mathbf{a}_j = (\mathbf{x}_1^j, \mathbf{x}_2^j, \cdots, \mathbf{x}_m^j)$. In that sense j denotes the unique index of the attribute $\mathbf{a}$ with respect to the feature set $\mathcal{A}$. To get a convenient notation, the function $\eta : \mathbb{R}^m \mapsto \{1, \cdots, |\mathcal{A}|\}$ will be used to denote the index of $\mathbf{a}$. Eventually, all class labels of the underlying classification problem are aggregated in the set $\mathcal{C} = \{c_1, c_2, \cdots, c_k\}$. The relation of a specific target value $y \in Y$ to a corresponding class $c \in \mathcal{C}$ is uniquely described by the function $\gamma : \mathbb{R} \mapsto \mathcal{C}$. An overview of important mathematical notations used for the definitions of the following five heuristic discrimination heuristic is provided in appendix A.

For the actual implementation of a heuristic discrimination model two fundamental mathematical approaches can basically be used. The primary group of heuristic discrimination model is founded on probabilistic formulations. Because of usual imprecision of random variable realizations due to measurement errors or internal computational representation, the probabilistic formulation of the feature's discriminative power guarantees robust and effective estimates. The key idea of probabilistic discrimination models is to distinguish the distribution of a certain random variable by the distribution of another variable without any particular knowledge of the

underling distributional parameters. Possible variable interdependencies are directly computed from the corresponding random samples. In terms of classification, the application of such models is straight forward since the two random variables can also be interpreted as the feature and the respective class variable. Popular representatives of probabilistic discrimination models are based on the concept of *mutual information*. The mutual information $I(X;Y)$ states the interdependency between two random variables X and Y. Consequently, $I(X;Y)$ also offers good estimates to reduce the uncertainty of X achieved by learning the state of the random variable Y. In information theory, the uncertainty of X with respect to the knowledge of Y is known as the entropy $H(X)$. In terms of classification, the formulation of the class-based entropy $H(c)$, the feature-based entropy $H(\mathbf{a})$ and the joint entropy $H(\mathbf{a}, c)$ are given in the following equations:

$$H(c) = -p(c)\log p(c) \tag{3.1}$$

$$H(\mathbf{a}) = -p(\mathbf{a})\log p(\mathbf{a}) \tag{3.2}$$

$$H(\mathbf{a}, c) = -p(\mathbf{a}, c)\log p(\mathbf{a}, c) \tag{3.3}$$

where $\mathbf{a}$ and c represent a specific feature and class respectively. In statistical-based heuristic models, the elementary feature and class probabilities are denoted with $p(\mathbf{a})$ and $p(c)$ where $\mathbf{a} \in \mathcal{A}$ denotes a selected feature and $c \in \mathcal{C}$ represents a specific class. The respective inverse probabilities are $p(\bar{\mathbf{a}}) = 1 - p(\mathbf{a})$ and $p(\bar{c}) = 1 - p(c)$. Moreover, some models also involve the joint probabilities. In addition to the joint probability $p(\mathbf{a}, c)$ of a specific feature and class the following joint probabilities are defined as

$$p(\bar{\mathbf{a}}, c) = \prod_{\mathbf{a}' \in \mathcal{A} \setminus \mathbf{a}} p(\mathbf{a}', c) \tag{3.4}$$

$$p(\mathbf{a}, \bar{c}) = \prod_{c' \in \mathcal{C} \setminus c} p(\mathbf{a}, c') \tag{3.5}$$

$$p(\bar{\mathbf{a}}, \bar{c}) = \prod_{\mathbf{a}' \in \mathcal{A} \setminus \mathbf{a}} \prod_{c' \in \mathcal{C} \setminus c} p(\mathbf{a}', c') \tag{3.6}$$

where $\bar{\mathbf{a}} := \mathcal{A} \setminus \mathbf{a}$ is the set of all features but feature $\mathbf{a}$ and $\bar{c} := \mathcal{C} \setminus c$ constitutes the set of all classes expect of the class c respectively. A general probabilistic independence is assumed due to obvious simplification considerations.

As various probabilistic models are based on the impurity function and use the entropy as the core mechanism to determine the discriminative power, the *Information Gain* (☞ see 3.2.2), the *Gain Ratio* (☞ see 3.2.3) and the *Balanced Information Gain* (☞ see 3.2.4) are included in this thesis because of their importance and frequent use in machine learning.

Apart from specific realizations of the impurity function, different probabilistic approaches are also known to define interdependency between two random variables and therefore offer possible reformulations of the desired variable discrimination. In this thesis, the probabilistic approaches Chi-square statistics (☞ see 3.2.1) and ReliefF (☞ see 3.2.5) are also deployed.

3.2.1 Chi-square statistics

The Chi-square ($\mathcal{X}^2$) statistics are frequently used in empirical science to measure the difference between observations and their expected results according to an initial hypothesis. More precisely, for given observations X and assumed results Y the $\mathcal{X}^2$ statistics investigates whether the unknown statistical distributions of X and Y actually differ from each other by estimating the dependency of these distributions. Thus, this approach provides a convenient way to empirically verify a given hypothesis by considering measured observations and concluded results only. In particular, from the statistical point of view is not required to describe the mostly unknown distributions of the variables X and Y in some way.

Considering arbitrary classification problems and music genre classification in particular, the $\mathcal{X}^2$ statistics can be used in order to measure how independent a certain feature $\mathbf{a} \in \mathcal{A}$ and a class $c \in \mathcal{C}$ approximately are. To estimate the value $\mathcal{X}^2(\mathbf{a}, c)$, the following definition

$$\mathcal{X}^2(\mathbf{a}, c) = \frac{[p(\mathbf{a}, c)p(\bar{\mathbf{a}}, \bar{c}) - p(\mathbf{a}, \bar{c})p(\bar{\mathbf{a}}, c)]^2}{p(\mathbf{a})p(\bar{\mathbf{a}})p(c)p(\bar{c})} \tag{3.7}$$

can be used where the dependency $\mathbf{a}$ in relation to a class c is described. See (3.6) for further details on the calculation of particular probabilities which are applied in (3.7). In the case of a low value for $\mathcal{X}^2(\mathbf{a}, c)$, the feature $\mathbf{a}$ is relatively independent to class c, and therefore this feature does not possess significant discriminative power to distinguish class c. Contrarily, a high value for $\mathcal{X}^2(\mathbf{a}, c)$ implies a high dependency and good discrimination. In order to obtain an estimation of feature $\mathbf{a}$ over all included classes, the weighted summarization

$$f(\mathbf{a}) = \sum_{c \in \mathcal{C}} \mathcal{X}^2(\mathbf{a}, c)p(c) \tag{3.8}$$

must be calculated and the final value $f(\mathbf{a})$ represents the discriminative power of the corresponding feature. The higher $f(\mathbf{a})$ actually is for the feature $\mathbf{a}$ the more discriminative is this feature.

3.2.2 Information Gain

As an important realization of the impurity function the *Information Gain*, which originated in information theory [26] and machine learning, is a synonym for the *Kullback-Leibler divergence*. It describes the amount of information one random variable X contains about another random variable Y. In other words, $IG(X, Y)$ measures the mutual information $I(X; Y)$, i. e. is the reduction of the uncertainty of X (or the entropy $H(X)$) achieved by learning the state of the random variable Y. Although $IG(X, Y)$ is rather an information-theoretic function, it can also be considered as an estimator of the dependency of the two random variables like the $\mathcal{X}^2$ statistics. The relation between uncertainty reduction and dependency is obvious because if two random variables X and Y are strongly independent, the uncertainty reduction of X due to additional knowledge of Y will be poor and therefore the value for $IG(X, Y)$ is near to zero. Contrarily, a

significant dependency of X and Y leads to a better uncertainty reduction and corresponds to a value for $IG(X,Y)$ closer to $\min\{H(X), H(Y)\}$. Thus, the value range of $IG(X,Y)$ is defined as $0 \leq IG(X,Y) \leq \min\{H(X), H(Y)\}$.

In terms of arbitrary classification problems the Information Gain can be used to measure the amount of information of a certain feature $\mathbf{a} \in \mathcal{A}$ has to determine a particular class $c \in \mathcal{C}$. According to the discriminative power of feature $\mathbf{a}$, if the value for $IG(c, \mathbf{a})$ is higher, then the feature will be more discriminative for class c. Basically, two equivalent formulations of $IG(\mathbf{a}, c)$ exist. The first definition

$$IG(\mathbf{a}, c) = \sum_{i \in \{\mathbf{a}, \bar{\mathbf{a}}\}} \sum_{j \in \{c, \bar{c}\}} p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \tag{3.9}$$

describes the Information Gain value explicitly by using probability definitions of (3.6). As the Information Gain is highly correlated to the entropy which projects the uncertainty of a random variable, $IG(c, \mathbf{a})$ can be alternatively formulated as

$$IG(\mathbf{a}, c) = H(c) + H(\mathbf{a}) - H(\mathbf{a}, c) \tag{3.10}$$

based on the definitions of equation (3.3) where $H(c)$ and $H(\mathbf{a})$ are the corresponding single entropies and $H(\mathbf{a}, c)$ is the joint entropy.

Finally, a weighted summarization is applied to compute the uncertainty reduction of all included classes $f(\mathbf{a})$ by a certain feature $\mathbf{a}$.

$$f(\mathbf{a}) = \sum_{c \in \mathcal{C}} IG(\mathbf{a}, c) p(c) \tag{3.11}$$

This yielded value $f(\mathbf{a})$ actually states the discriminative power of the corresponding feature. The higher the value for $f(\mathbf{a})$ is, the more discriminative is feature $\mathbf{a}$.

Although the Information Gain is a good measure to decide the relevance of a particular feature concerning class discriminative power, a notable problem occurs if the examined features can take a large number of distinctive values. In such cases the computed Information Gain value can be small and therefore implies poor discriminative power, even though the feature is actually discriminative. In order to avoid this problem, the *Gain Ratio* (☞ see 3.2.3) can be used instead.

3.2.3 Gain Ratio

The second heuristic discrimination model is the *Gain Ratio* which is based on the mutual information approach and therefore constitutes an impurity function as well. Since the Information Gain $IG(X,Y)$ tends to overestimate multi-valued features, the Gain Ratio have been introduced in [45] as a normalized realization of the Information Gain. The Gain Ratio also measures the uncertainty reduction of the random variable X achieved by getting knowledge about the state of the random variable Y. However, the Information Gain does not correctly estimate the relevance

of variables which have a large value range. In such cases, the obtained value for $IG(X,Y)$ will be always biased to zero and indicates poor dependency between the two variables, even though a strong dependency actually exists. Thus, the estimation concerning the discriminative power of the corresponding feature is not always reliable.

In order to guarantee good estimation of the dependencies of the two random variables X and Y, the obtained value for $IG(X,Y)$ must be additionally normalized to offer robustness regarding to the actual number of distinctive values the variable can be assigned with. The Gain Ratio $GR(X,Y)$ provides this robustness by using the entropy $H(Y)$ as the normalization factor.

In terms of arbitrary classification, the Gain Ratio $GR(\mathbf{a},c)$ of a certain feature $\mathbf{a} \in \mathcal{A}$ and a class $c \in \mathcal{C}$ can be computed by using the explicit formulation

$$GR(\mathbf{a},c) = \frac{\sum_{i\in\{\mathbf{a},\bar{\mathbf{a}}\}} \sum_{j\in\{c,\bar{c}\}} p(i,j) \log \frac{p(i,j)}{p(i)p(j)}}{-\sum_{i\in\{\mathbf{a},\bar{\mathbf{a}}\}} p(i) \log p(i)} \tag{3.12}$$

where $\bar{\mathbf{a}} := \mathcal{A} \setminus \mathbf{a}$ is the set of all features but feature $\mathbf{a}$ and $\bar{c} := \mathcal{C} \setminus c$ constitutes the set of all classes expect of the class c respectively. Alternatively, the Gain Ratio can be reformulated into

$$GR(\mathbf{a},c) = \frac{IG(\mathbf{a},c)}{H(\mathbf{a})} \tag{3.13}$$

where the Information Gain which is described in the previous subsection is directly applied.

To calculate the final value for the discriminative power concerning all included classes, the weighted summarization

$$f(\mathbf{a}) = \sum_{c\in\mathcal{C}} GR(\mathbf{a},c)p(c) \tag{3.14}$$

yields the desired value $f(\mathbf{a})$ which actually estimates how discriminative the feature $\mathbf{a}$ is in order to distinguish between classes. Again, the higher the value for $f(\mathbf{a})$ is, the more discriminative is feature $\mathbf{a}$.

3.2.4 Balanced Information Gain

Another realization of the impurity function is the *Balanced Information Gain* which is basically based on the Information Gain and also originated from information theory. Since the original Balanced Information model only deals with discrete variables, Wu et al. [58] have introduced a specific variation of the original Balanced Information Gain to also handle continuous variables which are often defined in machine learning applications. This definition of the Balanced Information Gain was also applied in this thesis.

As the Information Gain tends to overestimate multi-valued features, the Balanced Information Gain constitutes another heuristic to normalize the original feature contribution computed by the Information Gain. Another comparable approach which also utilizes a specific normalization heuristic with the Information Gain is the Gain Ratio discussed in the previous subsection. In terms of arbitrary classification, the Balanced Information Gain $B_g(\mathbf{a},c)$ of a certain feature

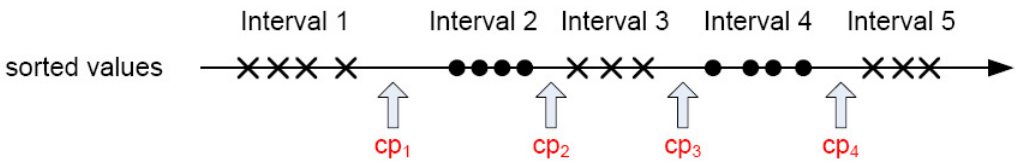


Figure 3.1: A common approach of discretising a numerical variable according to a binary class problem. The cut points cp_i with $1 \leq i \leq 4$ divide two consecutive feature values which are contained in instances labeled with different classes.

$\mathbf{a} \in \mathcal{A}$ and a class $c \in \mathcal{C}$ can be computed by using the explicit formulation

$$B_g(\mathbf{a}, c) = \frac{IG(\mathbf{a}, c)}{\log_2 \kappa} \tag{3.15}$$

where κ is the discretization cardinality of the feature $\mathbf{a}$. The actual value of κ represents a straightforward penalty on the bias of information gain due to multi-valued features and, therefore, it normalizes the feature contribution originated by the Information Gain.

The determination of the discretization cardinality κ of a feature is related to a common variable discretization approach which is illustrated in figure 3.1.

To obtain the discretization cardinality, all feature values are first sorted in ascending order. Afterwards, all adjacent values which are contained in data instances labeled to the same class are grouped together. Actually, this corresponds to the common discretization approach of a numerical variable by defining $t - 1$ cut points cp_i which create t continuous intervals. To turn a continuous feature into a discrete variable, these intervals are assumed to be individual values of the given feature. In that sense the discretization cardinality *kappa* of a feature $\mathbf{a}$ represents the number of individual values obtained by the discretization of the feature $\mathbf{a}$ which is 4 in the example illustrated in figure 3.1.

To calculate the final value for the discriminative power concerning all included classes, the weighted summarization

$$f(\mathbf{a}) = \sum_{c \in \mathcal{C}} B_g(\mathbf{a}, c) p(c) \tag{3.16}$$

yields the desired value $f(\mathbf{a})$ which actually estimates how discriminative the feature $\mathbf{a}$ is in order to distinguish between classes. Again, the higher the value for $f(\mathbf{a})$ is, the more discriminative is feature $\mathbf{a}$.

3.2.5 ReliefF

The *ReliefF* heuristic also measures the interdependencies between two or more random variables and was first introduced in machine learning for determining sufficient relevance orders of features in order to build rule-based learners, e. g. decision trees or random forests, properly. Although the context of measuring and the application domain of ReliefF and Information Gain are similar, ReliefF does not define the dependency of two random variables by the amount of contained

mutual information of those variables. In fact, ReliefF estimates the dependency of those random variables X and Y by how well the realizations of Y actually distinguish the realizations of X. In this context the realization of a random variable means the possible value which the variable can take. Obviously, this correlation between relations of X and Y will exist only if the corresponding distributions are dependent in some way.

Basically, ReliefF is one of several modifications of the core *Relief* algorithm which is presented in [29]. Both algorithms use another approach based on the nearest-neighbor algorithm and aim to estimate the quality of features to distinguish corresponding class values according to a specifically defined neighborhood within the input space. The main difference between the original Relief and its modification ReliefF is that ReliefF can also handle multi-class problems as well as incomplete and noisy data instances. To measure the quality of a certain feature $\mathbf{a} \in \mathcal{A}$, the key idea of both algorithms can be described as follows:

1. Select an instance $\mathbf{x} \in \mathbf{X}$ of the dataset $\mathcal{D}$ randomly
2. Determine two neighboring instances $\mathbf{x_H}$, $\mathbf{x_M} \in \mathbf{X}$ with $(\mathbf{x_H}, y_H)$, $(\mathbf{x_M}, y_M) \in \mathcal{D} : \; y_H \neq y_M$. The first instance $\mathbf{x_H}$ is the nearest neighbor having the same class assignment as $\mathbf{x}$, i. e. $(\mathbf{x}, y)$, $(\mathbf{x_H}, y_H) \in \mathcal{D} : \; y = y_H$, and is called *nearest hit*. The latter instance is the nearest neighbor which does not agree with the class related to $\mathbf{x}$, i. e. $(\mathbf{x}, y)$, $(\mathbf{x_M}, y_M) \in \mathcal{D} : \; y \neq y_M$. Consequently, this instance is denoted as *nearest miss*.
3. The actual quality of feature $\mathbf{a}$ is determined by how well different values of $\mathbf{a}$ result in different class assignments with respect to the instances $\mathbf{x}$, $\mathbf{x_H}$ and $\mathbf{x_M}$. On the one hand, if the feature $\mathbf{a}$ has different values in $\mathbf{x}$ and $\mathbf{x_H}$ then this feature obviously separates two instances with the same class. But this behavior of feature $\mathbf{a}$ is not desirable concerning the distinction of class values based on the feature observations and therefore the quality of feature $\mathbf{a}$ must be reduced. One the other hand, if $\mathbf{x}$ and $\mathbf{x_M}$ possess different values for feature $\mathbf{a}$ then this feature is actually more discriminative according to separate the different classes of $\mathbf{x}$ and $\mathbf{x_M}$. In that case the quality of feature $\mathbf{a}$ has to be increased.

Considering this relation between feature quality and class separation ability, the obtained quality measure of a certain feature correlates with the discriminative power of that feature. Thus, Relief and ReliefF can be employed to estimate how discriminative a feature is in order to distinguish class values. Instead of formulating possible variable discrimination due to interdependencies, the key idea of Relief, and also its extension ReliefF, can be outlined as the approximation of the following probability difference

$$ReliefF(\mathbf{a}) = P(\mathbf{x}^{j} \neq \mathbf{x}_M^{j}) - P(\mathbf{x}^{j} \neq \mathbf{x}_H^{j}) \tag{3.17}$$

where $j = \eta(\mathbf{a})$ denotes the unique index of $\mathbf{a}$. The function $\eta(\cdot)$ has already been defined in the beginning of section 3.2. Consequently, the notation $\mathbf{x}^{\eta(\mathbf{a})}$ constitutes the value of feature $\mathbf{a}$ contained in the given instance $\mathbf{x}$. Actually (3.17) is a more compact formulation of the usual definition of ReliefF. As the instances $\mathbf{x}$, $\mathbf{x_H}$ and $\mathbf{x_M}$ already inherit corresponding class

assignments, the first probability is based on different class assignments and the latter probability assumes same class assignment. Thus, this compact definition is equivalent to the longer definition of the fundamental works [29,48]. The following reformulation

$$ReliefF(\mathbf{a}) = \sum_{c \in \mathcal{C} \setminus \gamma(y)} \frac{p(c)}{1 - p(\gamma(y))} \delta(\mathbf{a}, \mathbf{x}, \mathbf{x}_M) - \delta(\mathbf{a}, \mathbf{x}, \mathbf{x}_H) \tag{3.18}$$

describes an explicit calculation model to estimate the value of ReliefF for a certain feature $\mathbf{a}$ by aggregating the partial estimations based on all classes $c \in \mathcal{C}$ but the class assigned to instance $\mathbf{x}$. The function $\gamma(y)$ returns the corresponding class of the selected instance $\mathbf{x}$ as defined in the beginning of section 3.2. The second function, $\delta(\cdot)$, describes the value difference between the chosen attribute of two given instances. In order to guarantee a more robust estimation of the actual discriminative power of a feature $\mathbf{a}$, the following two modifications of the equation (3.18) must be considered. First, instead of only using a single nearest hit $\mathbf{x}_H$ and nearest miss $\mathbf{x}_M$ with respect to the selected $\mathbf{x}$ a set of k nearest hits $\mathbf{X}_H$ and nearest misses $\mathbf{X}_M$ should be used. Moreover, the entire computation of ReliefF should not be proceeded a single time but l times. This means that the entire computation of $ReliefF(\mathbf{a})$ is separately performed l times and the corresponding partial estimations are aggregated like in equation (3.20). The modified ReliefF definition is given by the following equation

$$ReliefF(\mathbf{a}) = \sum_{c \in \mathcal{C} \setminus \gamma(y)} \frac{\frac{p(c)}{1-p(\gamma(y))} \sum_{i=1}^{k} \delta(\mathbf{a}, \mathbf{x}, \mathbf{X}_M^i)}{l \cdot k} - \sum_{i=1}^{k} \frac{\delta(\mathbf{a}, \mathbf{x}, \mathbf{X}_H^i)}{l \cdot k} \tag{3.19}$$

where $\mathbf{X}_H^i$ and $\mathbf{X}_M^i$ are the i^{th} nearest hit and miss respectively. The equation (3.19) was also employed in this thesis. The function $\delta(\cdot)$ represents a metric to measure the differences of the values of the chosen feature $\mathbf{a}$ from two given instances. The actual definition of $\delta(\cdot)$ depends on the algorithms of Relief or ReliefF but only the definition of the latter algorithm is given here. Contrary to the original Relief algorithm, ReliefF's $\delta(\cdot)$ can also be calculated in the case of one value or even both values being actually unknown. Thus, ReliefF can actually handle incomplete data instances as well. The following four definitions of function $\delta(\cdot)$ is defined to handle all possible situations concerning the availability of the desired attribute values where the unique attribute index of $\mathbf{a}$ is $j = \eta(\mathbf{a}$:

1. If both values of the feature $\mathbf{a}$ are known for both $\mathbf{x}_1$ and $\mathbf{x}_2$ then:

$$\delta(\mathbf{a}, \mathbf{x}_1, \mathbf{x}_2) = \begin{cases} 0 & \mathbf{x}_1^j = \mathbf{x}_2^j \\ 1 & otherwise \end{cases}$$

2. If the value of the feature $\mathbf{a}$ is unknown for $\mathbf{x}_1$ then:

$$\delta(\mathbf{a}, \mathbf{x}_1, \mathbf{x}_2) = 1 - p(\mathbf{x}_2^j | \gamma(y_1))$$

3. If the value of the feature $\mathbf{a}$ is unknown for $\mathbf{x}_2$ then:

$$\delta(\mathbf{a}, \mathbf{x}_1, \mathbf{x}_2) = 1 - p(\mathbf{x}_1^j | \gamma(y_2))$$

4. If the value of the feature $\mathbf{a}$ for both $\mathbf{x}_1$ and $\mathbf{x}_2$ are unknown then[1]:

$$\delta(\mathbf{a}, \mathbf{x}_1, \mathbf{x}_2) = 1 - \sum_{a \in \mathbf{a}} p(a|\gamma(y_1)) \cdot p(a|\gamma(y_2))$$

where $a \in \mathbf{a}$ is a specific value of the attribute $\mathbf{a}$.

As the reference instance $\mathbf{x}$ is randomly selected, a robust estimation of the quality of feature $\mathbf{a}$ can not be guaranteed because only a very small area of the entire feature domain is used for the determination. Consequently, a repeated aggregation of ReliefF introduces a more reliable estimation. The following formulation

$$f(\mathbf{a}) = \sum_{i=1}^{l} ReliefF(\mathbf{a}), \quad l > 0 \tag{3.20}$$

constitutes this repeated aggregation by summarizing the obtained ReliefF values for every iteration. Similar to all previous heuristic models, the higher the value for $f(\mathbf{a})$ is, the more discriminative is feature $\mathbf{a}$.

To summarize, ReliefF has two basic parameters to control robustness of the feature quality estimation which are the number of repeated iterations l and the number of k nearest hits and misses according to the reference instance $\mathbf{x}$.

3.3 Experiments

The following empirical study of the five heuristic discrimination models and the three descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histogram, which are described in section 2.5, is established to conclude possible answers to the main question of this thesis which is: *Can a specific rhythmic feature pattern or patterns be significantly related to a particular musical genre?* This question implies some very interesting conclusions concerning a possible improved rhythmic description of some musical genres or by offering a promising method for feature selection in music genre classification.

The three musical genre collections GTZAN, ISMIR 2004 Genre and Rhythm (☞ see 2.6) have been employed to compute the discriminative power of every included feature but the main part of the analysis focuses on results computed on the base of the GTZAN music collection. GTZAN has been favored because it includes the largest number of popular musical genres with equally distributed musical pieces per genre. For each feature set the key observations of the results based on GTZAN were additionally verified on the ISMIR 2004 Genre collection, since

[1] In this definition of $\delta(\cdot)$ the variable a denotes a specific value of the feature $\mathbf{a}$.

this music collection contains quite similar musical genres. Remarks to the results based on ISMIR 2004 Rhythm are also given for each feature set. Moreover, the evaluation results are also provided in terms of each of the five heuristic discrimination models described in section 3.2. As already discussed, the calculation models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain are based on the concept of mutual information between two random variables and specific computation parameters are not needed. Contrarily, some important computational parameters must be adjusted for using the ReliefF model. Actually, those parameters have been used which are originally defined by the WEKA workbench. In particular all training instances were used to establish sample probes to estimate the attributes. Additionally, the 10 nearest neighbors were considered for relevance estimation.

Moreover, the evaluation is based on one-vs.-rest genre situations. For every music collection all musical pieces are incorporated for computation. To determine those features which actually distinguish a certain genre, the original genres are relabeled to a corresponding binary genre situation. The underlying multidimensional feature space and the actual ranges of possible values of every feature are not altered or reduced and the complete information of every feature may influence the discrimination calculation. The actual discrimination value of every feature is estimated by a multiple-fold calculation approach. The entire dataset is randomly divided into 10 individual folds containing the same number of instances each. The discrimination values according to every feature are computed for every fold separately. After the computation of the discrimination values a feature ranking is determined for every fold based on the discrimination values of every feature. Thus, 10 independent rank estimates are obtained for every feature. In order to take only statistically reliable results into account, a rank correlation test based on Kendall's correlation coefficient τ is successively applied. If the results of a specific fold are significantly different comparing with the other folds, this fold will be recalculated. Consequently, the final discrimination and ranking estimates of every feature are obtained by averaging the corresponding results of the 10 folds.

The empirical study is structured into three sub sections referring to the individual feature sets Rhythm Patterns in 3.3.1, the Statistical Spectrum Descriptor in 3.3.2 and the Rhythm Histogram descriptor in 3.3.3. Four evaluation steps were performed for each feature set respectively.

First, the existence of discriminative features and possible individual feature patterns according to a specific genre is discussed on the results of the Gain Ratio model by illustrating the discriminative power of every feature in a matrix representation. In the second step, the conclusions regarding the discriminative feature patterns computed by the Gain Ratio will be compared with the corresponding results based on the Balanced Information Gain as well as the ReliefF. Third, the numeric distribution of discrimination values related to a specific genre and calculation model are examined by computing important statistical measures and by depicting the discrimination values against the ranking order where the largest discrimination value is always denoted with rank 1. Fourth and last, the definition of a ranking order with respect to the discriminative power of every feature should provide further insight into the ability of every

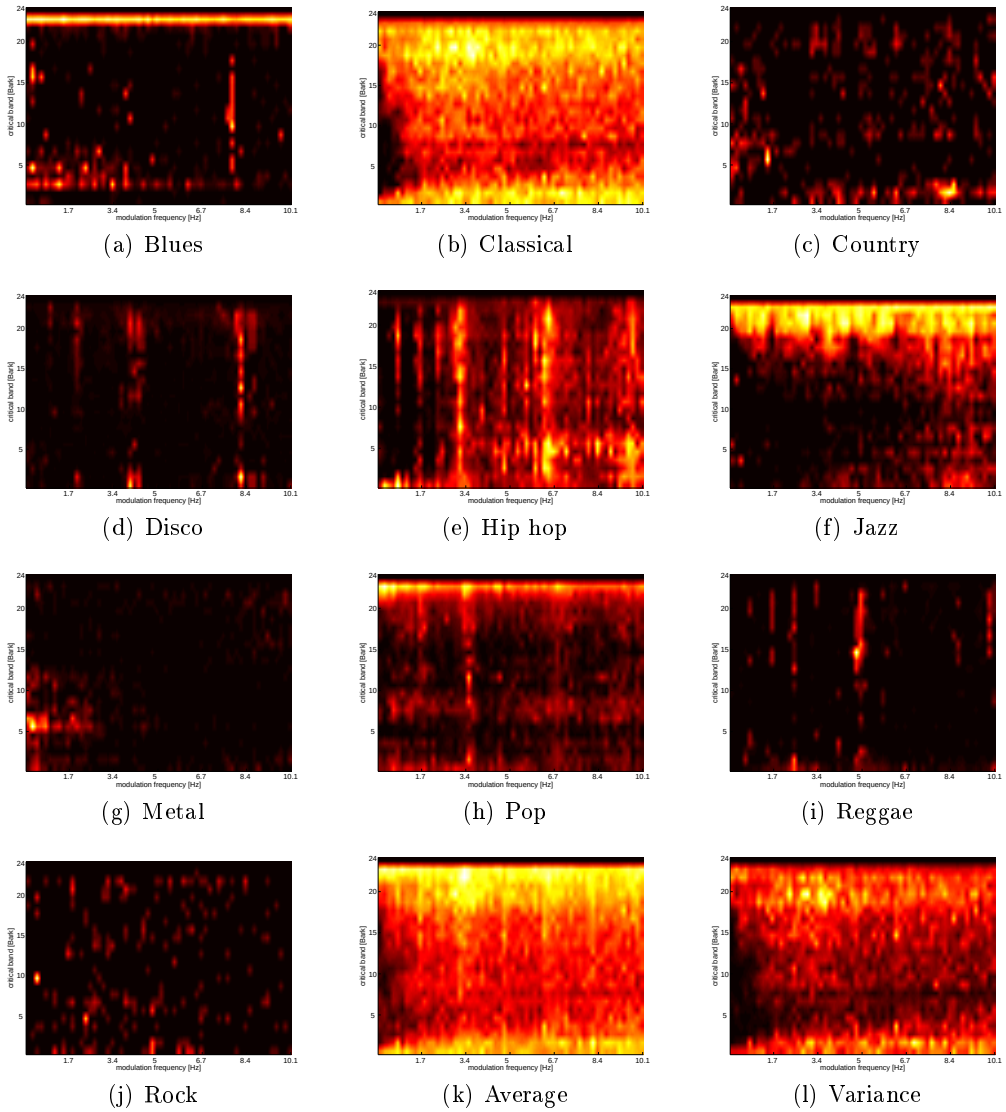
calculation model to exhibit individual discriminative feature patterns for every genre. Such a feature ranking is based on a specific genre and calculation model and also provides a way of representing individual feature patterns. Although the intrinsic meaning of every feature concerning critical band and modulation frequency can not be clearly recognized, the significance of individual feature patterns can be considered for every genre. To obtain such a conclusion of the significance of individual feature patterns based on a specific calculation model, a rank correlation test has been performed to verify the correlation of the respective feature rankings to be compared. The rank correlation test is based on Kendall's correlation coefficient τ and assumes the hypothesis $\mathcal{H}_0$ of non-zero correlation with a significance level $\alpha = 0.05$. On the basis of these test results, genres with non-correlated feature ranking and therefore differently assigned discriminative features can be plausibly identified.

As already mentioned, the key goal of the following experiments is to identify certain rhythmic patterns which significantly represent a single or a small group of particular musical genres. Those rhythmic patterns are always related to a certain rhythmic component and if a significant assignment to a genre can be actually assumed then this will suggest a potentially improved rhythmic description of that genre. Evidently, it can not be assumed in the first place that the perception of information due to the applied heuristic discrimination models actually coincides with the meaning of musical information and therefore stronger proof is needed. Thus, the discriminative feature patterns have to be additionally verified. Chapter 4 focuses on this verification of using discriminative feature patterns, i. e. specific feature subsets, based on the discrimination ranking in terms of musical genre classification. If the classification accuracy will be robust or even increased although only a subset of features has been applied then this empirically confirms the assumption.

3.3.1 Rhythm Pattern

The features defined by the Rhythm Pattern descriptor reflect loudness fluctuations on critical frequency bands with respect to modulation frequencies between 0.2 and 10.1 Hz. The features can be constituted in a matrix representation which also offers a convenient way to visualize and to analyze discriminative features according to the underlying genre. In figure 3.2, the discriminative features computed by the Gain Ratio model are illustrated for all 10 genres of the GTZAN collection. Additionally, the average and the variance of the discrimination values computed over all genres are visualized.

Considering the discriminative features according to every genre, two groups of genres can be recognized which are characterized by the number of discriminative features describing the respective genres. The first group contains the genres Classical, Hip hop, Jazz and Pop where the number of discriminative features is definitely larger compared with the other genres. As figure 3.2(b) illustrates, features describing classical music exhibit high discrimination values which are related to critical bands with Bark numbers larger than 15, i. e. frequencies above 2.7 kHz, and Bark numbers less than 4, i. e. frequencies less than 0.2 kHz. Moreover, these features correspond to various modulation frequencies. Those features which are related to critical bands



(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock (k) Average (l) Variance

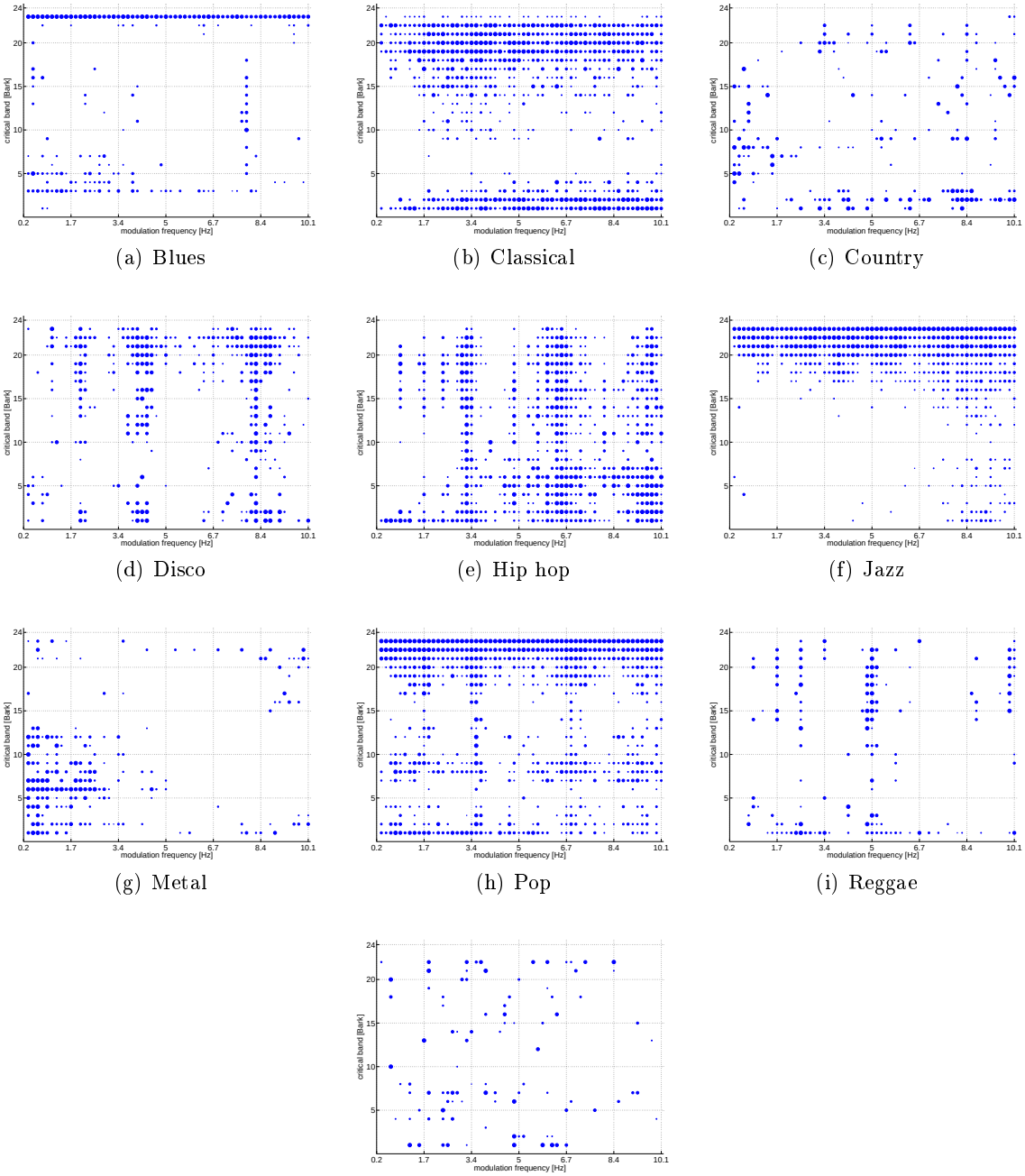
Figure 3.2: Discriminative features of the Rhythm Pattern descriptor according to the music collection GTZAN. The Gain Ratio model was used to compute all discrimination values where less discriminative features are colored with red and black (darker) tones, while more discriminative features are colored with yellow (brighter) tones. Figures (k) and (l) represent the average and variance results over all genres.

between 3 and 15 Bark mostly possess far less discrimination. Discriminative features describing the genres Jazz and Pop are also especially related to very high critical bands, while discriminative features corresponding to Hip hop are actually distributed along specific modulation frequencies instead. In particular those features have relatively high discrimination values which correspond to modulation frequencies close to 3.3, 6.6 and 9.9 Hz along various critical bands. In the case of the genre Pop the modulation frequency close to 3.4 is emphasized. Contrary to Pop and Hip hop, figure 3.2(f) implies for jazz music that features located at high critical bands and along various modulation frequencies are relevant which is a quite similar observation as in the case of

classical music. As jazz music is played in various rhythmic styles while Pop and Hip hop have fewer rhythm variations, this observation appears to be plausible. Despite the discriminative feature patterns describing each of the four genres Classical, Hip hop, Jazz and Pop diverge, the number of selected features being discriminative is quite large. In order to explain this large number of discriminative features the intrinsic musical characteristics of those genres must be considered. For instance classical and pop music involve a high variation of rhythmic styles and loudness. Furthermore, the instrumentation might be more manifold as in the case of the genres Blues and Rock. Moreover, as the evaluation is based on one-vs.-rest comparisons, these strong differences to almost all other genres are even more emphasized during the discrimination computation.

The second group contains the other 6 genres Blues, Country, Disco, Metal, Reggae and Rock which are represented by a relative small number of discriminative features. In fact, most of the features have actually a zero or a very low discrimination value. Consequently, those features can be assumed to be irrelevant in terms of determining the genre and maybe even in terms of classification of the respective genres. Chapter 4 will further evaluate this assumption by performing genre classification. It can be observed clearly that for all genres but Blues the discriminative features are located at a limited but different number of critical bands and modulation frequencies. An interesting fact regarding the genre Blues is that those features are particularly relevant which relate to the critical band of 23 Bark and along almost the entire range of modulation frequencies. The reason why this high bark band is so discriminative may be the usually very restricted variation of rhythmic styles within Blues. Since one-vs.-rest evaluation has been performed, this difference to other genres is more emphasized. The large variation of emphasized modulation frequencies concerning rock music is also not surprising because of the typical broader musical understanding of rock music which implies a higher variation of rhythmic characteristics.

To summarize the observations according to the discrimination results based on the Gain Ratio model, every genre was represented by a considerable number of discriminative features where individual feature patterns were also suggested for every genre. The next step of this discriminant analysis is to expand the examination to the other four heuristic discrimination models. The figures 3.3, 3.4 and 3.5 illustrate the discriminative features according to each genre and based on the calculation models Gain Ratio, Balanced Information Gain and ReliefF respectively. To provide a clear visualizations, only 50 % of all actually discriminative features, i. e. having a non-zero discrimination value, were plotted as filled dots with varying size. A larger size indicates better discrimination of the corresponding feature. The discrimination results computed by the calculation models Chi-square and the Information Gain are not explicitly discussed as those results only marginally differ compared with the results based on the Gain Ratio and Balanced Information Gain model. In fact, this similarity is not surprising as all heuristic discrimination models but the ReliefF implement the impurity function. In particular the Information Gain, Gain Ratio and the Balanced Information Gain are directly related to each other as they utilize the entropy measure to estimate the dependency between a specific



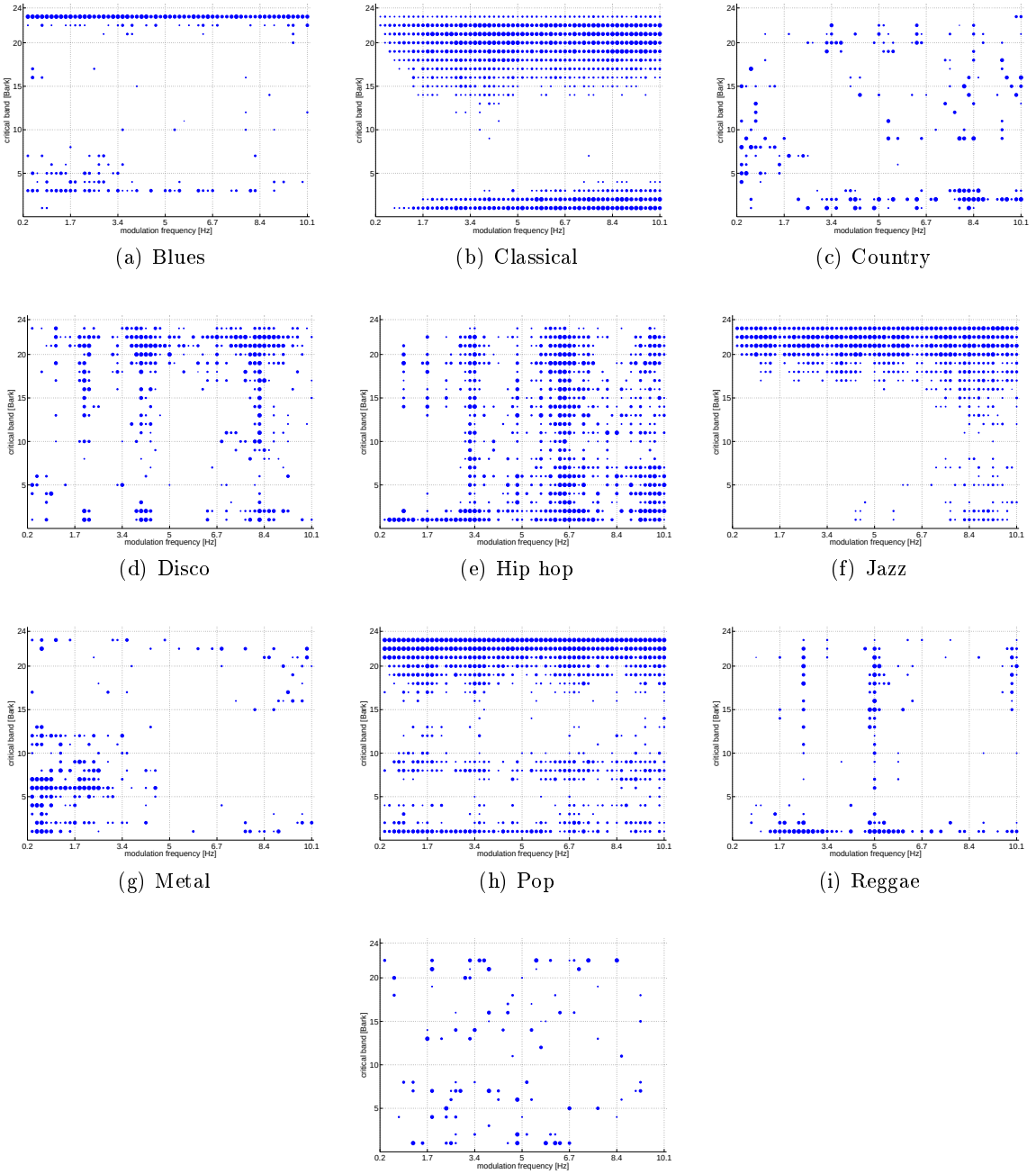
(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock

Figure 3.3: Inter-genre comparison of discriminative features according to the Rhythm Pattern descriptor and the Gain Ratio on the GTZAN collection. In order to provide a clear visualization, only 50 % of those features were taken into account which have a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.



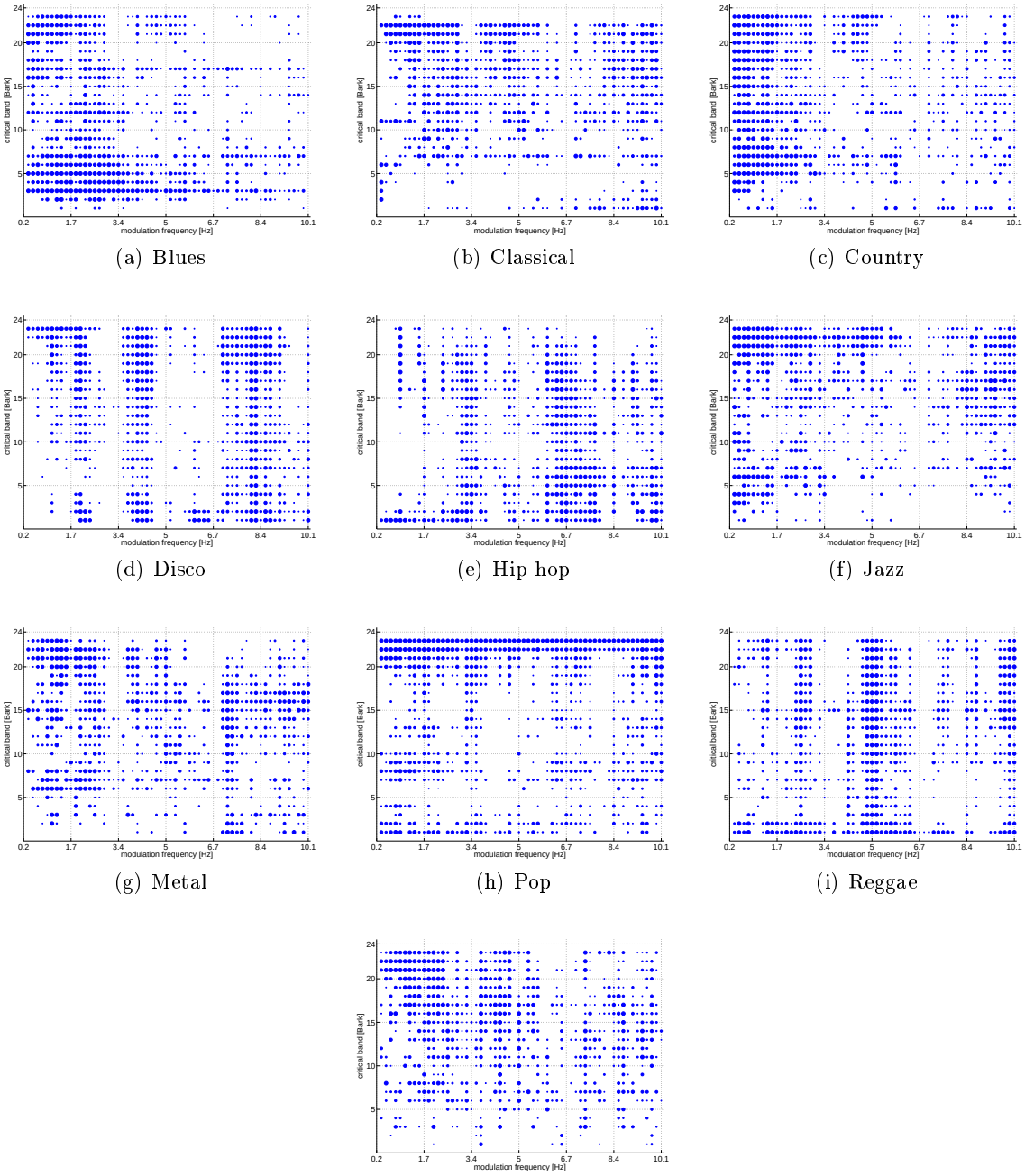
(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock

Figure 3.4: Inter-genre comparison of discriminative features according to the Rhythm Pattern descriptor and the Balanced Information Gain on the GTZAN collection. In order to provide a clear visualization, only 50 % of those features were taken into account which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.



(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock

Figure 3.5: Inter-genre comparison of discriminative features according to the Rhythm Pattern-descriptor and the ReliefF on the GTZAN collection. In order to provide a clear visualization, only 50 % of those features were taken into account which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.

feature and a genre.

Figure 3.3 illustrates the most discriminative features according to the Gain Ratio model. Specific observations and conclusions according to the discrimination results based on the Gain Ratio have already been discussed in context of figure 3.2 and can also be followed from the illustration representing only 50 % of the most discriminative features. In figure 3.4, the corresponding discrimination results according to the Balanced Information Gain are depicted. In fact, the discrimination results are very similar compared with the results based on the Gain Ratio where almost the same features are selected to be discriminative. Yet, the corresponding discrimination values of those discriminative features vary. Nevertheless, very similar discriminative feature patterns can actually be recognized for each of the 10 genres. Consequently, the following observations hold true for both the Gain Ratio and the Balanced Information Gain.

Two genre groups can be determined which differ in the number of features being discriminative. In terms of the genres Classical, Jazz and Pop the most discriminative features correspond to very high critical bands and also along various modulation frequencies. In the case of Jazz and Pop discriminative features are actually related to almost every modulation frequency between 0.2 and 10.1 Hz and to the critical bands 21 to 23. Considering other feature regions within the Rhythm Pattern descriptor, those three genres introduce relatively diverging results. Those features exhibit a non-zero discrimination value for all three genres which are located on critical bands with less than 5 Bark and modulation frequencies larger than 8 Hz. The number of discriminative features according to Hip hop is relatively similar compared with the other three genres. Yet, those features definitely correspond to different critical bands as well as modulation frequencies. In fact, three feature regions of the Rhythm Pattern descriptor are emphasized where features having a high discrimination value. All discriminative features are distributed along the modulation frequencies 3.3, 6.6 and 9.9 Hz at various critical bands between 1 and 23 Bark (almost the entire range). Nevertheless, for all genres of this first group it can be concluded that features corresponding to the highest critical band having the Bark number 24 as well as to some low modulation frequencies are not discriminative at all.

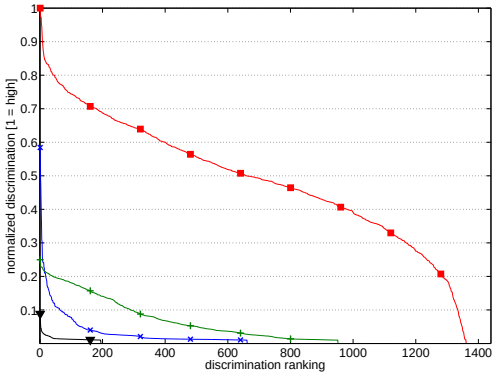
For the other 6 genres of GTZAN the discrimination results based on the Gain Ratio as well as the Balanced Information Gain suggest a definitely smaller number of features being discriminative compared with the genres of the first group. Two conclusions can be made. First, the critical bands and modulation frequencies related to the discriminative features diverge among these genres. Second, a subset of discriminative features are recognizable corresponding to consecutive but very limited intervals of critical bands and modulation frequencies for the genres Blues, Country, Disco, Metal and Reggae. These subsets are illustrated as closed ("blobby") areas within the respective matrix representation. Contrarily, the discriminative features describing Rock are related to varying critical bands as well as modulation frequencies. Moreover, only very few features exhibit non-zero discrimination values.

According to the ReliefF model the discrimination results considerably diverge comparing with the corresponding results based on the calculation models implementing the impurity function. Figure 3.5 visualizes the discrimination values based on the ReliefF. On the one hand, vari-

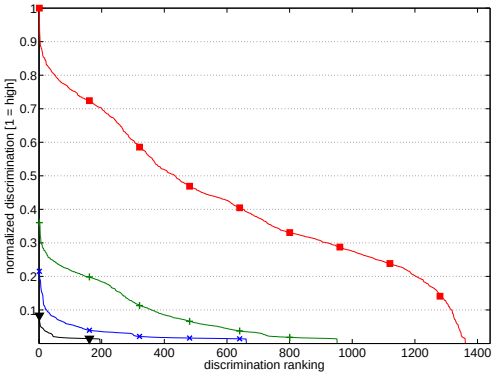
ations of the discrimination values exist among both the ReliefF and the Gain Ratio or Balanced Information Gain. But on the other hand, the ReliefF model estimates non-zero discrimination for features which are not discriminative in terms of the other two calculation models. Actually, these differences are not completely surprising, since the ReliefF model implements an approach of estimating feature discrimination which also incorporates the dependencies between the features and not only the dependencies between a specific feature and a genre. The actual difference between the concepts of calculation models implementing the impurity function and the ReliefF is discussed in [48]. It can be observed that the number of features being discriminative does not vary as much between the genres. In fact, the number of discriminative features is relatively consistent for all genres defined by the GTZAN collection. But this also means that some genres are represented by features related to various modulation frequencies although musical pieces of those genres usually include a less variation of rhythmic characteristics. For instance musical pieces of Blues or Metal do not vary considerably in terms of the rhythmic style. It appears that ReliefF is influenced by the underlying evaluation procedure. According to the one-vs.-rest evaluation, a high discrimination value does not always imply that the corresponding feature characterizes genre. The inverse meaning is also possible in the way that the corresponding feature does not characterize the genre. Despite the difference in number of discriminative features according to the discrimination results based on the Gain Ratio and Balanced Information Gain, an interesting similarity can be observed. Considering the genres Blues, Classical, Jazz and Pop, those features exhibit large discrimination values which are distributed along a broad range of modulation frequencies but only at specific critical bands. On the other hand, the inverse observation can be concluded for the genres Disco, Hip hop and Reggae, namely that a large number of discriminative features is related to specific modulation frequencies but at almost the entire Bark range from 1 to 23. This observation confirms that less rhythmic variations, e. g. varying beats per minute, exist in musical pieces of Disco, Hip hop and Reggae. Again features corresponding to the critical band 24 do not exhibit a large discrimination value in terms of all genres.

The next step of the discriminant analysis is to examine the distribution of the discrimination values according to every genre. Particularly, the scale of the discrimination values according to every genre and the number of features having zero discrimination should be discussed. Figure 3.6 illustrates the relation of the discrimination values against the ranking order according to each of the four genres Classical, Disco, Jazz and Rock and the three calculation models, while a statistical description regarding the underlying distribution of the discrimination values is given in table 3.1. As the range of discrimination values is very large, all values were normalized into the interval $[0, 1]$ considering the discrimination values of all 10 genres. These normalized discrimination values were used both in the illustration regarding the relation of the discrimination values against the ranking order and in the statistical description.

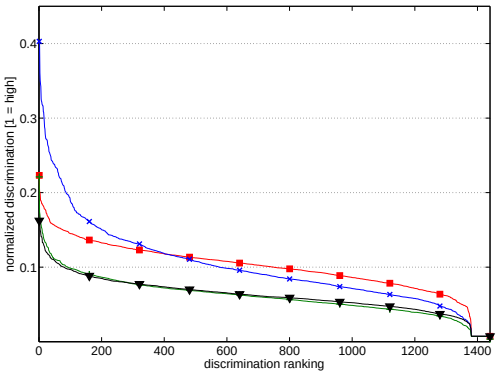
According to figure 3.6, the features describing the genre Classical exhibit far larger discrimination values comparing with the other genres. Also the number of features having zero discrimination is smaller compared with the other genres. Contrarily, in terms of Disco, Jazz and



(a) Gain Ratio



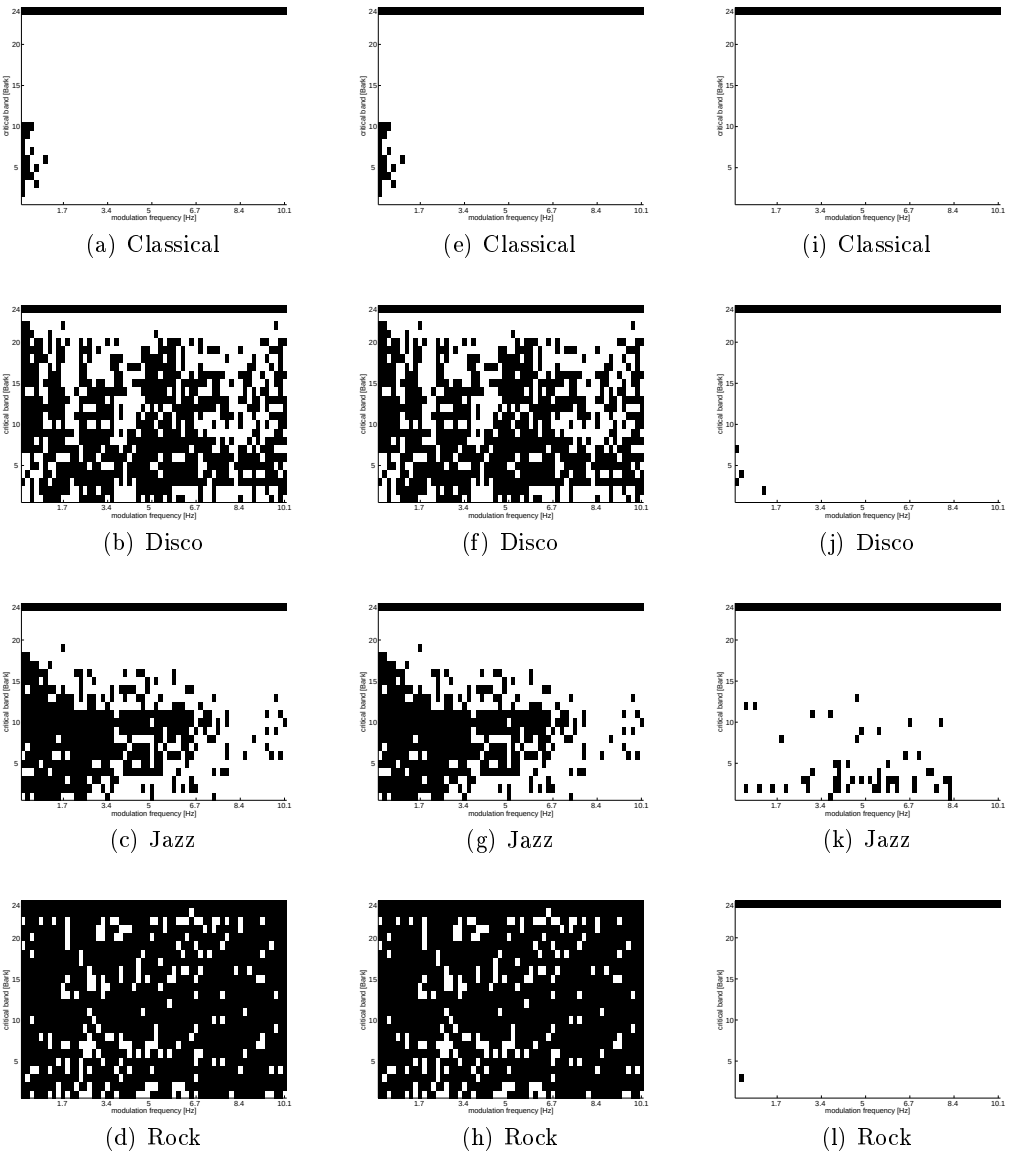
(b) Balanced IG



(c) ReliefF

Figure 3.6: Illustration of the discrimination values against the ranking order according to the Rhythm Pattern descriptor on the GTZAN collection (normalized into the interval $[0, 1]$ for every calculation model separately). Each musical genre is illustrated by an individual color and symbol: ■ for **Classical**, × for **Disco**, + for **Jazz** and ▼ for **Rock**.

Rock a larger number of features suggest zero discrimination with rock music having the smallest number of discriminative features. Also the number of features exhibiting large discrimination values is far smaller in the case of those three genres compared with classical music. According to the ReliefF the relation of the discrimination values against the ranking order implies a small number of features with high discrimination values but the differences between the four genres are not that significant compared with the results of the first two calculation models. An interesting fact is that the largest discrimination values are related to pop music and not to classical music. Figure 3.7 explicitly points out which features are actually irrelevant with respect to the four selected genres and the three calculation models. Although the ReliefF model estimates non-zero discrimination values for every feature and every genre, irrelevant features must be still considered as some features have a discrimination value less than the discrimination value of a random probe feature. This generic feature was added to the original feature set where the feature values were distributed by a Gaussian distribution with $\mu = 0$ and $\sigma = 1$. All features with a zero discrimination value or a discrimination value smaller than the one of the random



(a) Classical (e) Classical (i) Classical

(b) Disco (f) Disco (j) Disco

(c) Jazz (g) Jazz (k) Jazz

(d) Rock (h) Rock (l) Rock

Figure 3.7: Irrelevant features of the Rhythm Pattern descriptor according to the GTZAN collection. All *black* colored features are considered as irrelevant. Gain Ratio: (a) to (d), Balanced Information Gain: (e) to (h), ReliefF: (i) to (l).

feature are depicted as black in figure 3.7. An interesting fact is that all features related to the highest critical band, i. e. frequencies larger than 12 kHz, are actually irrelevant. But this is not surprising considering the encoding of the samples contained in the GTZAN collection where the sampling frequency is 22 kHz. Thus, only frequencies up to 11 kHz were actually encoded in the samples of the GTZAN collection.

Table 3.1 lists a short statistical description of the discrimination results. The statistical measures *mean*, *standard derivation*, *min* and *max value* were computed by using the normalized discrimination value range. The Gain Ratio and the Balanced Information Gain models estimate the highest discrimination value for classical music followed by pop music. Yet, the margin

	Gain Ratio				Balanced IG				ReliefF			
Genre	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**
Blues	0.05	0.06	0.00	0.20	0.05	0.09	0.00	0.31	0.09	0.04	0.02	0.32
Classical	0.49	0.18	0.00	1.00	0.41	0.21	0.00	1.00	0.10	0.03	0.02	0.22
Country	0.01	0.01	0.00	0.09	0.01	0.02	0.00	0.13	0.06	0.02	0.01	0.18
Disco	0.03	0.06	0.00	0.58	0.02	0.03	0.00	0.20	0.10	0.05	0.01	0.40
Hip hop	0.11	0.08	0.00	0.48	0.11	0.07	0.00	0.48	0.18	0.06	0.05	0.42
Jazz	0.07	0.06	0.00	0.24	0.08	0.08	0.00	0.35	0.06	0.02	0.00	0.22
Metal	0.02	0.03	0.00	0.29	0.02	0.03	0.00	0.18	0.06	0.02	0.01	0.15
Pop	0.10	0.10	0.00	0.69	0.11	0.13	0.00	0.90	0.19	0.10	0.03	1.00
Reggae	0.04	0.05	0.00	0.42	0.02	0.03	0.00	0.22	0.10	0.04	0.00	0.29
Rock	0.01	0.01	0.00	0.08	0.01	0.01	0.00	0.07	0.06	0.02	0.01	0.16

Table 3.1: Statistical summarization of the discrimination values according to the Rhythm Pattern descriptor and the GTZAN collection (normalized into the interval $[0, 1]$ for every calculation model separately). Only those discrimination values were considered which were originally non-zero.

between the maximum values regarding classical and pop music considerably varies for both calculation models. This variation of the maximum values is mainly explained by the different approach of normalizing multi-valued features. More details concerning the different normalization of multi-values features are described in section 2.3. Also the highest average discrimination value listed in table 3.1 is related to classical music. This agrees with the observation concluded in figure 3.6 where a large number of features representing classical music have high discrimination values. The maximum and average values according to ReliefF are highest in case of pop music instead. Classical music is not that emphasized as in the case of the calculation models implementing the impurity function.

The next step of the discriminant analysis is to verify the difference of the discriminative feature patterns according to every genre. The possible existence of individual, i. e. statistically significant, feature patterns is to be examined. To effectively conclude individual discriminative feature patterns, genre-based ranking sequences were used. Those ranking sequences were sorted in descending order of the discriminative power of every feature, and therefore the first rank corresponds to the most discriminative feature and the least discriminative feature is denoted with the largest rank position. With the use of this rank-based representation of the discriminative features a statistical rank correlation test was performed to verify whether two different feature rankings are individual or not. In terms of the rank correlation test this means that the p-value indicates a significant non-zero or zero correlation respectively. Unfortunately, this rank correlation test only works well if all features have an individual rank respectively different discrimination values. Since a considerable number of features exhibited zero discrimination according to various genres, the ranks of these features are statistically tied. Thus, those features have the same rank which refers to their average rank. Consequently, the expressiveness of the rank correlation test is limited in that way that not all individual discriminative feature patterns

are actually recognized by the test results. Nevertheless, all discriminative feature patterns which are suggested to be significantly different according to the corresponding test results are actually different.

Table 3.4 lists the p-values of all tested genre pairs according to the three heuristic discrimination models. The significance level is defined with $\alpha = 0.05$ and a p-value greater than α indicates that the corresponding two genres are represented by individual discriminative feature patterns each. Two conclusions can be made in terms of the obtained number of individual discriminative feature patterns. On the one hand, the test results reveal few significantly different feature patterns only for the Gain Ratio and the Balanced Information Gain. Yet, this is not surprising as those calculation models estimate quite many feature having zero discrimination for various genres. According to the Gain Ratio and the Balanced Information Gain the genre Pop is represented by the largest number of individual feature pattern compared with the other genres where the discriminative feature patterns representing Pop are significantly different compared with genres Blues, Disco and Metal. In the case of the Balanced Information Gain the feature patterns representing the genres Reggae and Rock also diverge significantly with respect to Pop. However, from the view point of audio perception and musical styles, those genres are more related to Pop than for instance to Classic or Jazz. The discriminative feature patterns computed by the ReliefF model are far more often correlated to each other. As all features exhibit non-zero discrimination values in terms of the ReliefF the rank correlation test unfolds its full expressiveness. Consequently, it can be assumed that many genres are actually represented by quite similar discriminative feature patterns.

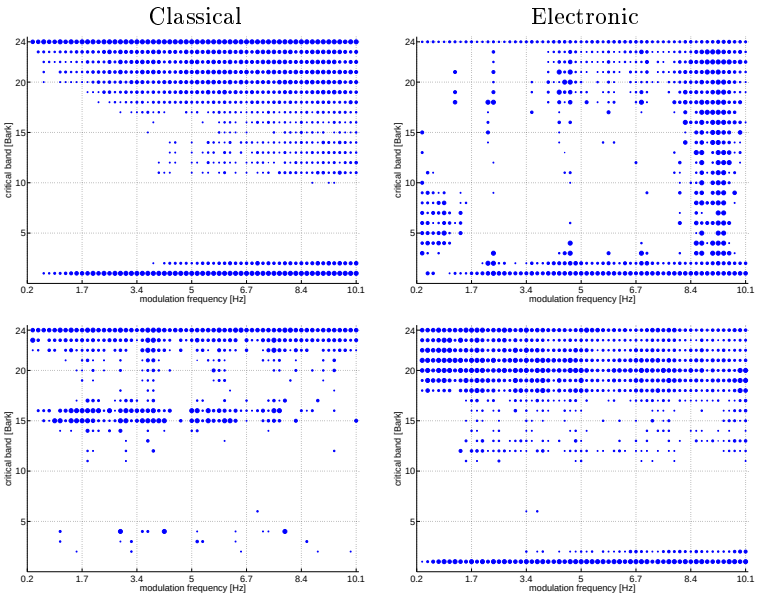
The eventual step of the discrimination analysis is to examine the discrimination results based on the ISMIR 2004 Genre and Rhythm music collections. Similar and diverging conclusions concerning the identification of genre-specific feature patterns based on these collections will be shortly discussed and a comparison to the conclusions regarding the GTZAN collection will be made.

ISMIR 2004 Genre

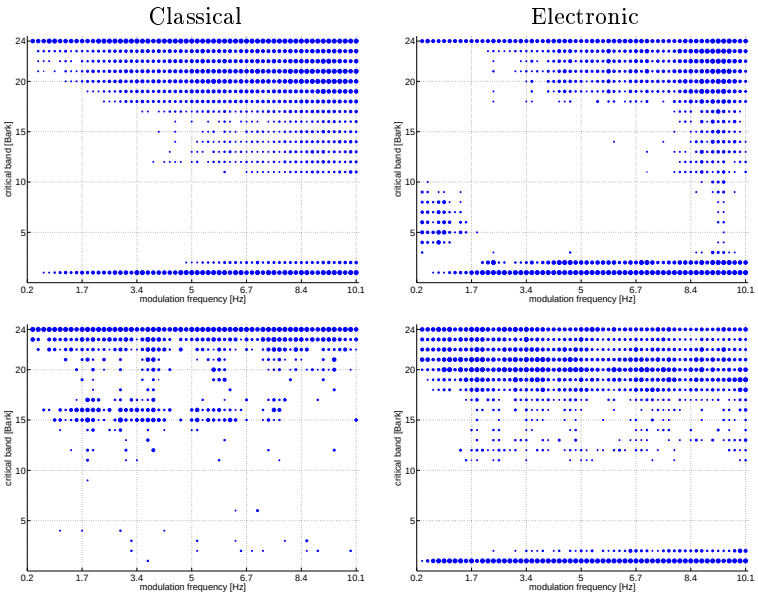
Although the genres of the ISMIR 2004 Genre collection do not explicitly correspond to the genres of the GTZAN collection and, moreover, the number of defined genres differs in both collections, four genres were selected to be used in the following analysis. These genres are Classical, Electronic, Jazz & Blues and Rock & Pop and relate to the GTZAN genres Classical, Disco[2], Jazz and Rock respectively. Since at least a partial correlation can be assumed to the genres Classical, Disco, Jazz and Rock of the GTZAN collection respectively, a comparison of the respective discrimination results was done although only both classical genres coincide sufficiently.

Figure 3.8 illustrates the computed discriminative features according to each of the four selected genres based on the calculation models Gain Ratio, Balanced Information Gain and

[2]It has been assumed that the genre Electronic is based on a more generic musical genre description which actually covers the genre Disco of the GTZAN collection.



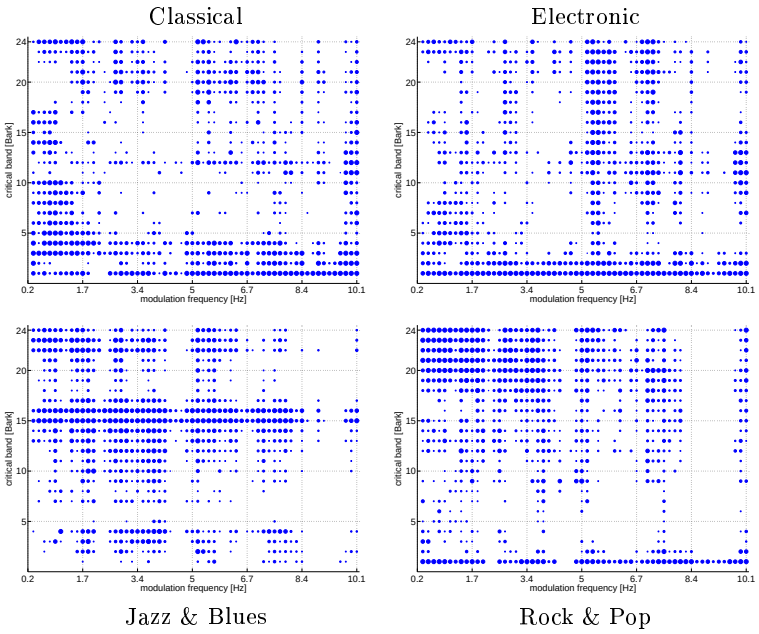
(a) Gain Ratio



(b) Balanced Information Gain

Continued on the next page ...

ReliefF respectively. Again, the two models Chi-square and the Information Gain are not explicitly presented, since the corresponding discrimination results are quite similar compared to the corresponding results based on the Gain Ratio and the Balanced Information Gain. To provide clear visualizations, only 50 % of all actually discriminative features are plotted as filled dots



(c) ReliefF

Figure 3.8: Inter-genre comparison of discriminative features according to the Rhythm Pattern descriptor on the ISMIR 2004 Genre collection. Three calculation models were used where figure (a) represents the Gain Ratio, figure (b) corresponds to the Balanced Information Gain and figure (c) is related to the ReliefF. In order to provide a clear visualization, 50 % of those features were taken into account only which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.

with varying size, while the larger size indicates better discrimination of the corresponding feature. According to the Gain Ratio and the Balanced Information Gain only the discrimination results of the genre Classical are quite similar compared with the corresponding results regarding the genre Classical of the GTZAN collection. Again, those features exhibit high discrimination values which correspond to critical bands with Bark numbers larger than 15 and various modulation frequencies. This observation agrees with the observation based on the GTZAN collection. Moreover, those features related to the genre Classical possess a high discrimination which correspond to critical bands with less than 3 Bark and almost all modulation frequencies but the very lowest. An interesting fact is that features related to the highest critical band (24 Bark) are also estimated being discriminative in the case of ISMIR 2004 Genre collection but not for the GTZAN collection. In fact, this observation is valid for all four selected genres. The critical band denoted by the Bark number 24 represents the frequency range of $[12, 15.5]$ kHz. The reason why those feature are discriminative in the case of the ISMIR 2004 Genre collection is the different encoding (different sampling frequencies) of the musical pieces contained in the two collections. The difference in the encoding regarding these two collections is described in section 2.6. For the other three genres a considerable similarity of the discrimination results according to both

the GTZAN and the ISMIR 2004 Genre collections can not be assumed because the respective discriminative feature patterns strongly differ.

As in the case of the GTZAN collection, the ReliefF calculation model yields a large number of discriminative features which is quite equal among the four selected genres. Those features related to the very low critical bands possess a high discrimination value in the case of all genres but Jazz & Blues. On the other hand, many discriminative features are related to the critical bands having a Bark number of at least 20. Some partial similarities can be concluded in the results based on ReliefF compared to the corresponding results based on the GTZAN collection. In the case of the genres Rock and Rock & Pop, which are contained in the collections GTZAN and ISMIR 2004 Genre respectively, features are discriminative according to both collections which correspond to higher critical bands and lower modulation frequencies. Also a partial similarity of the discrimination results can be recognized for the genres Disco and Electronic. On the other hand, the discrimination results of both classical genres considerably differ although these two genres are assumed to correlate more than the other genres. Thus, a consistent performance over correlated music collections can not be concluded although the difference in the discrimination values between the two collections appears to be more limited in the case of 2 of the 4 genre comparisons. According to the Gain Ratio and the Balanced Information Gain the discrimination results based on the two collections actually coincide more compared to the results computed by the ReliefF. In particular the discrimination results of both classical genres imply a considerable similarity. Also for the genres Jazz and Jazz & Blues the difference in the discriminative feature patterns is quite limited, as for both genres features related to critical bands 15 Bark and above are particularly discriminative. The results of genres Pop and Rock & Pop also show quite many features which are discriminative for both genres. Thus, a more consistent performance of the Gain Ratio and the Balanced Information Gain is affirmed.

Table 3.5 lists the p-values of the rank correlation tests according to the ISMIR 2004 Genre collection. Two key agreements are notable comparing with the respective test results based on the two music collections. First, the performances of the three calculation models can be divided into two groups again. On the one hand, the models Gain Ratio and Balanced Information Gain offer few genre-specific feature patterns only. As already mentioned in terms of the rank correlation test based on the GTZAN collection, these two calculation models estimates zero discrimination for quite many features which reduce the effectiveness of the entire correlation test. On the other hand, a large number of genres are represented by individual feature patterns according to the ReliefF. An interesting fact is that the corresponding genres do not coincide with the results on the GTZAN collection at all because very few individual discriminative feature patterns are concluded in the case of the GTZAN collection.

ISMIR 2004 Rhythm

Eventually, key conclusions which have been made on both the GTZAN and the ISMIR 2004 Genre collections should be verified on the ISMIR 2004 Rhythm collection. This collection contains Latin and Ballroom dance music only and a correlation to genres of the other two music

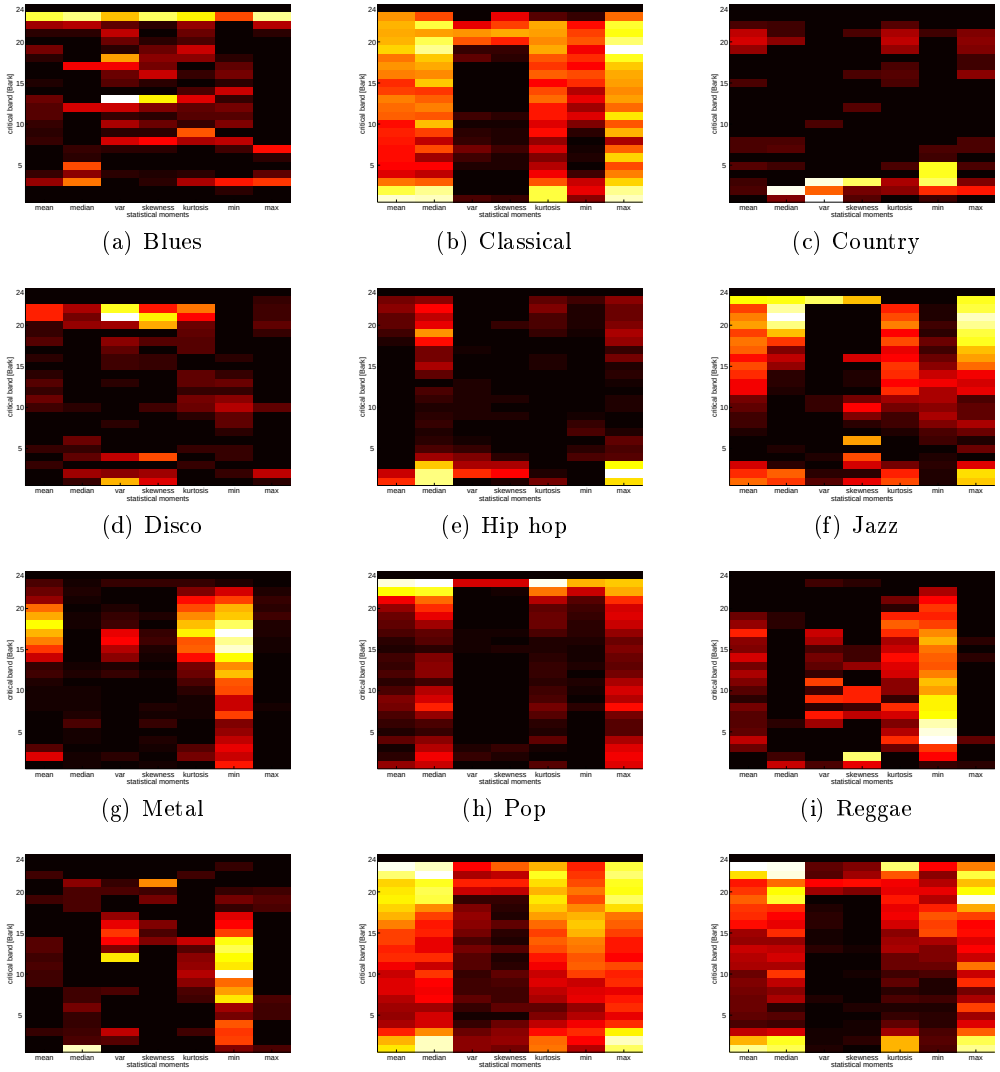
collection can not be assumed. Thus, a genre-to-genre comparison of the respective discriminative feature patterns computed by the three heuristic discrimination models Gain Ratio, Balanced Information Gain and ReliefF is not meaningful. Nevertheless, some basic characteristics of the three calculation models were observed.

The discrimination results according to the ISMIR 2004 Rhythm collection are obviously difficult to compare as completely different genres are included comparing with the other two collections. Nevertheless, it can also be concluded that respective discrimination results based on the Gain Ratio and the Balanced Information Gain only marginally diverge. From this follows that different approaches of normalizing multi-valued features do actually only slightly influence the computation of discriminative features. As this conclusion is also valid in the case of the other two collections it can be assumed that the conclusion is generally valid. For many genres like for instance ChaChaCha, Jive or Samba few features are estimated to be discriminative. Another interesting fact is that according to many genres the discriminative features correspond to specific modulation frequencies as well as critical bands. Thus, the discriminative feature patterns often appear as "blobby" areas where all contained features exhibit high discrimination values. The genres ChaChaCha, Jive and Samba are examples for this observation. The ISMIR 2004 Rhythm collection only contains Latin and Ballroom dance music and musical pieces of each dance genre are specifically related to a small variation of rhythm styles. As the discrimination results actually emphasize few discriminative modulation frequencies with a small range of critical bands, this assumption appears to be affirmed.

The ReliefF model also estimates a limited number of discriminative features which is considerably smaller compared to the other two collections. Nevertheless, more features are actually discriminative as in the case of the Gain Ratio and the Balanced Information Gain. This observation is also valid for both the other two music collections. An interesting fact is that for some genres, e. g. ChaChaCha and Jive, the discrimination results of all three calculation models are strongly similar. This is a contradictory observation according to both the GTZAN and the ISMIR 2004 Genre collection. An explanation may be the stronger correlation of the genres to specific rhythm styles as the dance music is particularly represented by a specific rhythm or a small variation of rhythm. Thus, some features corresponding to specific critical bands and modulation frequencies are particularly correlated to a single genre and have no or a small correlation with other genres. It appears that this strong correlation between specific features and a genre is responsible for the higher agreement of the discrimination results based on the three calculation models.

3.3.2 Statistical Spectrum Descriptor

The features of the Statistical Spectrum Descriptor are constituted by the computation of seven statistical measures for each of the 24 available critical bands. Basically, the Statistical Spectrum Descriptor describes the distributions of modulations frequencies per critical band and includes the following statistical measures: *mean*, *median*, *variance*, *skewness*, *kurtosis* and *min- and max-value* which are abbreviated with the numbers 1 to 7. The feature set can also be represented by a



(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock (k) Average (l) Variance

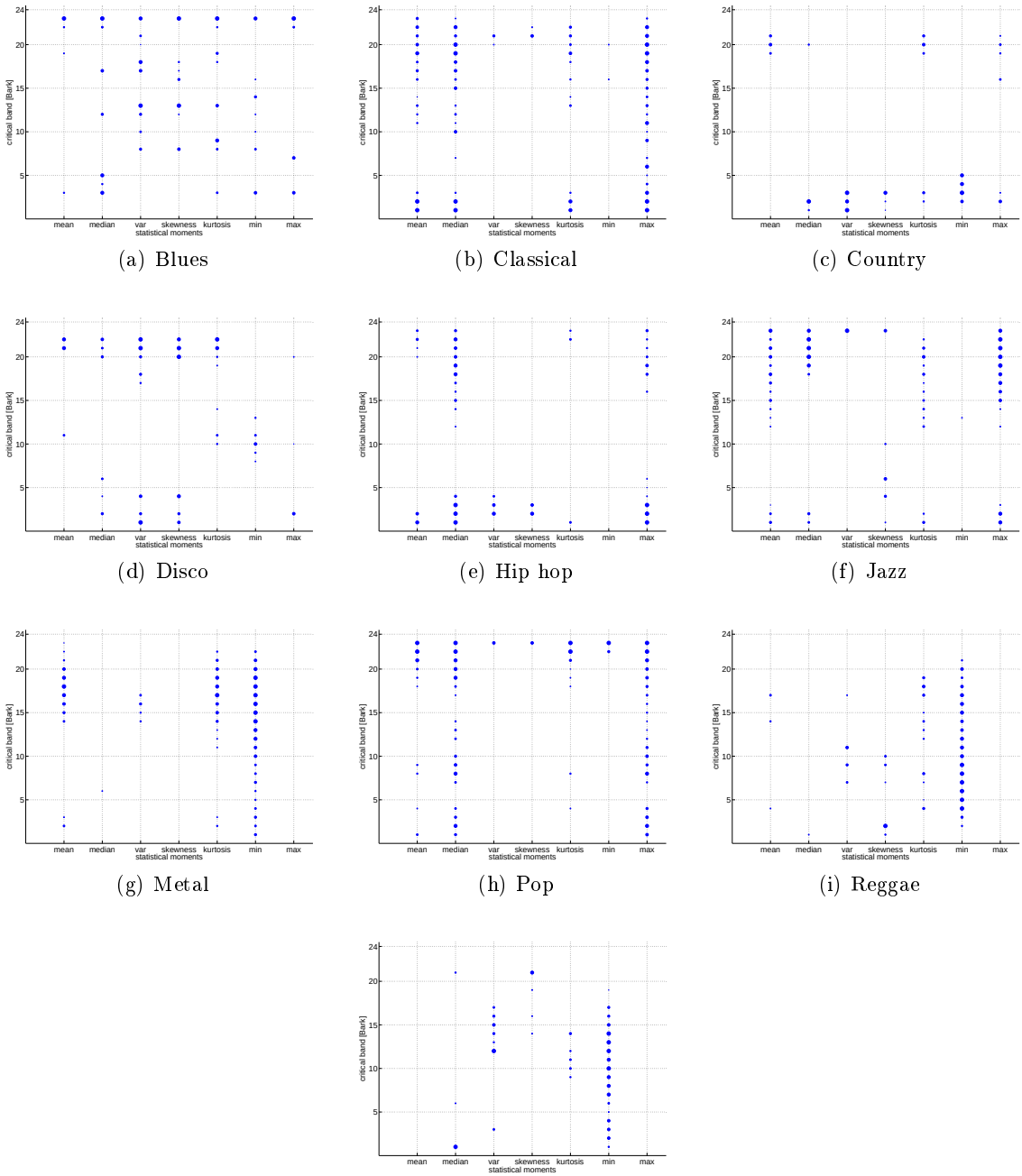
Figure 3.9: Discriminative features of the Statistical Spectrum Descriptor according to the music collection GTZAN. The Gain Ratio model was used to compute all discrimination values where less discriminative features are colored with red and black (darker) tones, while more discriminative features are colored with yellow (brighter) tones. Figures (k) and (l) represent the average and variance results over all genres.

matrix representation which provides a convenient way to visualize the discriminative distribution according to all features.

As in the case of the Rhythm Pattern descriptor the discriminant analysis of the Statistical Spectrum Descriptor begins with an introductory discussion of the discrimination results based on the Gain Ratio. Figure 3.9 visualizes the discrimination results based on the Gain Ratio according to all 10 genres of the GTZAN collection. Moreover, two illustrations presenting the average and the variance of the discrimination values computed over all genres are depicted in this figure. Similar to the Rhythm Pattern descriptor all genres are constituted by quite different feature patterns. In particular, two genres are represented by a large number of discriminative features which are the genre Classical and Jazz, while in the case of the Rhythm Pattern descriptor Hip hop and Pop are represented by a large number discriminative features. According to both Classical and Jazz the statistical measures mean, median and max value are emphasized, while features related to the measures variance and skewness exhibit very low discrimination values or even zero discrimination. Also all critical bands with a Bark number between 15 and 23 are emphasized. The critical band 15 is defined by the frequency interval $[2.32, 2.7]$ kHz, while the critical band 23 corresponds to the interval $[9.5, 12]$ kHz. In fact, in the case of Classical and Jazz music these critical bands are also emphasized according to the Rhythm Pattern descriptor. Additionally, in terms of the genre Classical features possess high discrimination values which are related to critical bands below the Bark band 3, i. e. frequencies less than 0.2 kHz, and measures mean, median, kurtosis and max value. Thus, a key difference in discriminative feature patterns of classical and jazz music appears to be that features corresponding to lower critical frequency bands have high discrimination values according to classical music only.

Another group can be observed containing genres which are represented by discriminative features corresponding to few or even single statistical measures. The genres Metal, Reggae and Rock are such genres where the min value appears to be particularly important. Many features being highly discriminative correspond to this very measure over various critical bands, while features related to the other measures are considerably less discriminative. In fact, only specific features related to the measures mean, median and kurtosis also exhibit relatively high discrimination values. The other 4 genres Blues, Country, Disco and Hip hop are constituted by a considerably smaller number of discriminative features. Nevertheless, an interesting observation is that according to all these 4 genres almost only features are actually discriminative which are distributed along either high critical bands with more than 20 Bark or low critical bands with less than 6 Bark. This suggests that a quite large number of critical bands does not include any discriminative information about the respective underlying genre.

According to figure 3.9, two very promising observations can be concluded for a wider range of genres. First, for a considerably large number of genres the statistical measures variance and skewness are highly irrelevant in terms of suggesting discriminative features. Only in the case of the genres Country and Disco features related to these two measures exhibit high discrimination values. This observation also agrees with the illustration of the average discrimination values in figure (k) where features can be assumed as relatively irrelevant which correspond to these



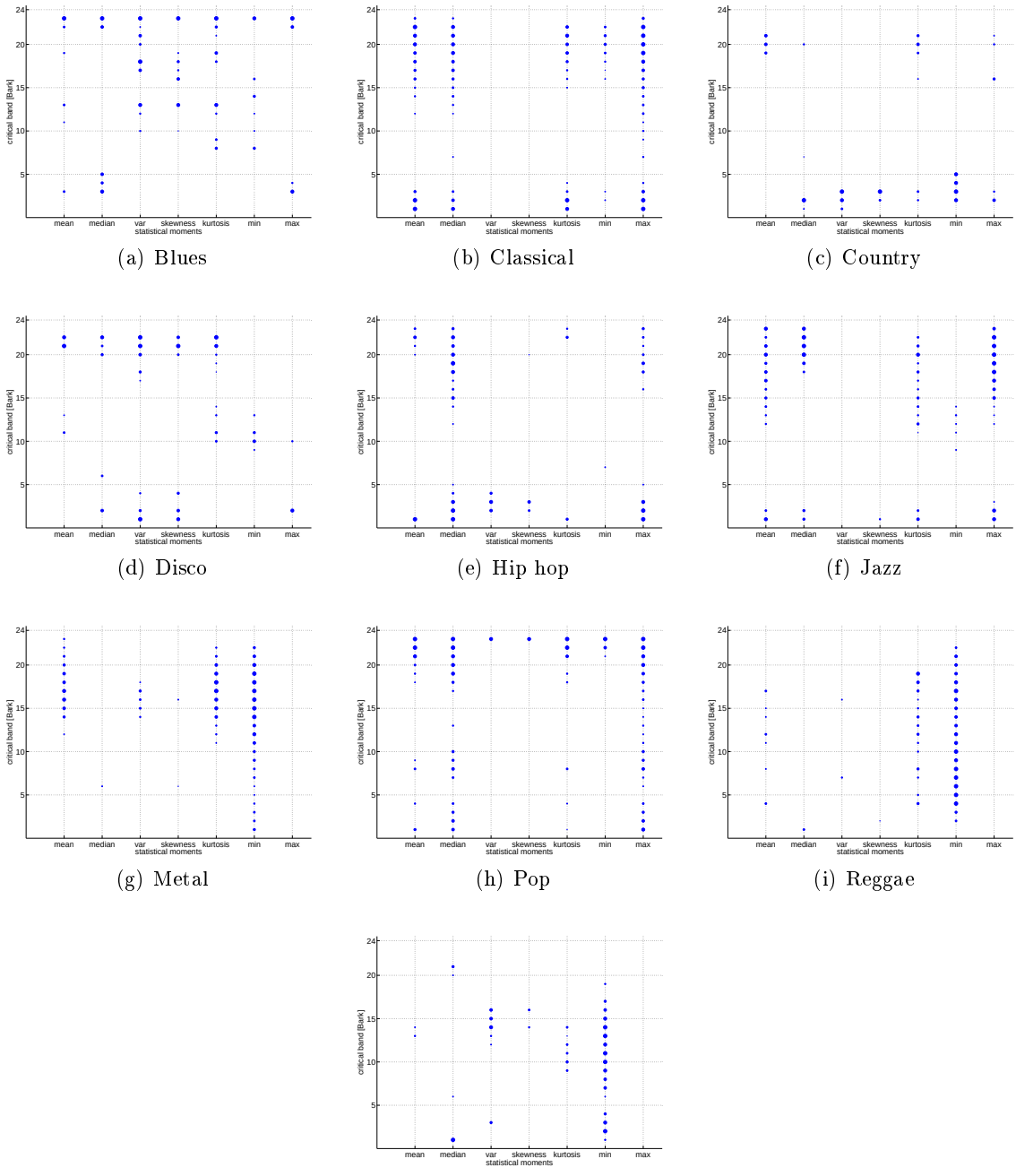
(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock

Figure 3.10: Inter-genre comparison of discriminative features according to the Statistical Spectrum Descriptor and the Gain Ratio on the GTZAN collection. In order to provide a clear visualization, only 50 % of those features were taken into account which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.



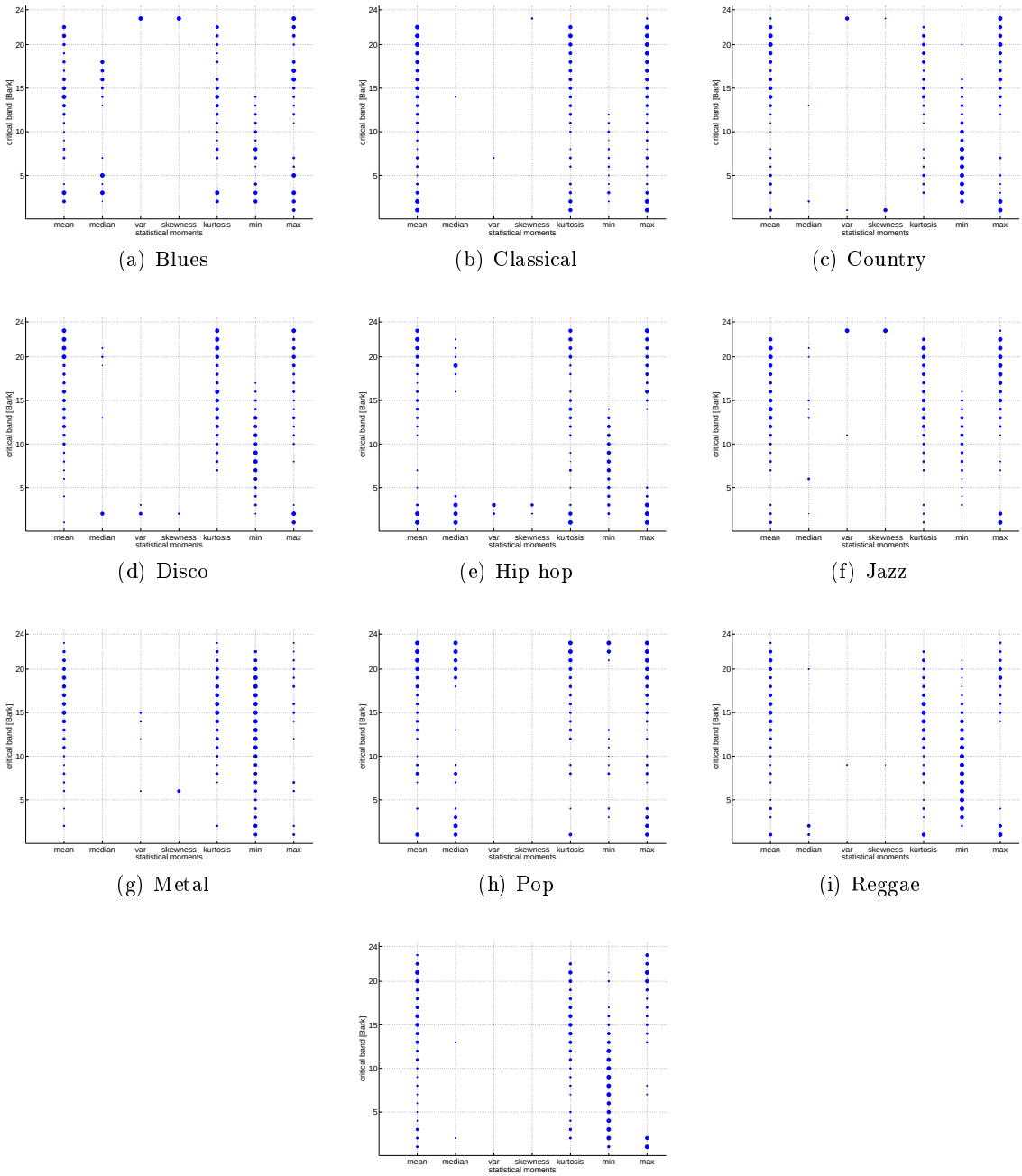
(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock

Figure 3.11: Inter-genre comparison of discriminative features according to the Statistical Spectrum Descriptor and the Balanced Information Gain based on the GTZAN collection. In order to provide a clear visualization, only 50 % of those features were taken into account which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.



(a) Blues (b) Classical (c) Country

(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

(j) Rock

Figure 3.12: Inter-genre comparison of discriminative features according to the Statistical Spectrum Descriptor and the ReliefF on the GTZAN collection. In order to provide a clear visualization, only 50 % of those features were taken into account which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.

two measures and all but very high critical bands. Moreover, the variances according to the measures variance and skewness, which are depicting in figure (l), are very low or even zero for those features. This mean that it can be consistently assumed that these features are quite irrelevant over all genres. Like in the case of the Rhythm Pattern descriptor in 3.3.1, the critical band 24 representing the frequency interval [12, 15.5] kHz is completely irrelevant for all examined genres.

The next step of the discriminant analysis is to expand the examination to other heuristic discrimination models. In order to compare the discrimination results of each of the three calculation model, the figures 3.10, 3.11 and 3.12 illustrate the discriminative results according to each genre based on the Gain Ratio, Balanced Information Gain and the ReliefF respectively. The discrimination results of the calculation models Chi-square and the Information Gain are not explicitly presented because of the high degree of similarity regarding those discrimination results comparing with the results of both he Gain Ratio and the Balanced Information Gain. To provide a clear visualizations, again 50 % of all actually discriminative features are taken into account only. Every discriminative feature is depicted by a dot of varying size. The size depends on the actual discrimination value of that feature where larger a size denotes a higher discrimination value.

The figures 3.10 and 3.11 confirm clearly that the discriminative feature patterns computed by both the Gain Ratio and the Balanced Information Gain are quite similar. As the discrimination results according to the Gain Ratio calculation model have already been discussed before and the discrimination results based on the Balanced Information Gain only marginally differ, the discrimination results based on the ReliefF calculation model will be examined in the following. In figure 3.12, the discriminative features of every genre are visualized according to the ReliefF. Contrary to the results based on the other two calculation models, all genres are represented by an almost similar number of discriminative features. The genre Blues is the only exception as it is represented by a slightly smaller number of discriminative features. This is a different conclusion comparing with the Gain Ratio and the Balanced Information Gain where differences in the number of discriminative features could be recognized. In fact, this different performance of the ReliefF has also been concluded in terms of the Rhythm Pattern descriptor. Thus, it appears that the ReliefF consistently estimates a large number of features to be discriminative according to all genres. An important fact also holds true for the ReliefF which actually coincides with the Gain Ratio and the Balanced Information Gain. The statistical measures variance and skewness are quite irrelevant for the majority of genres, since features corresponding to those two measures exhibit low discrimination values or even zero discrimination. In fact, these two measures even appear to be more irrelevant compared with the Gain Ratio and the Balanced Information Gain. As already seen before and in context of the Rhythm Pattern descriptor, all features related to the critical band 24 Bark possess no discrimination at all.

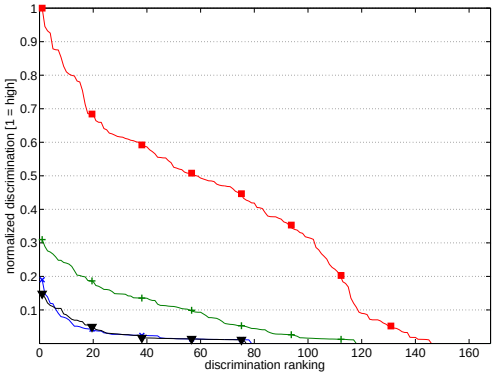
The comparison of the results according to the three calculation models reveals a surprising fact which could not be concluded in terms of the Rhythm Pattern descriptor. It appears that the difference in the discriminative feature patterns is limited and a considerable degree of

similarity can be recognized for many genres. For instance the discriminative feature patterns describing the genres Classical, Metal or Pop imply such a considerable degree of similarity because all three calculation models estimated quite a large number of the same features to be discriminative, while the corresponding discrimination values vary. As this degree of similarity regarding the discrimination results of the three calculation models could not be seen for both the Rhythm Pattern descriptor and the Rhythm Histogram descriptor, specific characteristics of the Statistical Spectrum Descriptor must be the reason of this interesting observation. In the case of the ISMIR 2004 Rhythm collection and the Rhythm Pattern descriptor such a similarity of the discrimination results based on the three calculation models has also been concluded for many genres where the strong feature-genre dependency has been assumed to be the reason. This might also be a plausible explanation for the same observation regarding the Statistical Spectrum Descriptor. Lidy et al. showed in [38] that the Statistical Spectrum Descriptor outperforms both the Rhythm Pattern descriptor and the Rhythm Histogram descriptor according to the GTZAN collection in terms of genre classification. Thus, it can be assumed that the features of the Statistical Spectrum Descriptor possess more information regarding genre discrimination. This higher correlation of the feature to a respective genre may be the reason for the similar performances.

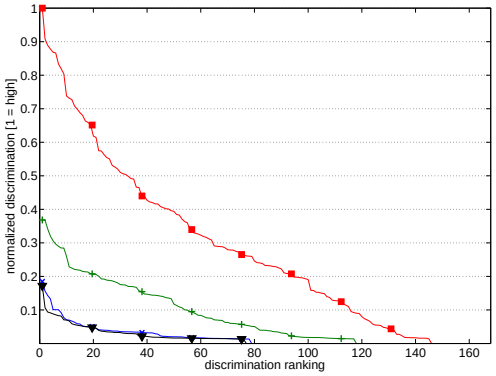
The next step of the discriminant analysis is to examine the distribution of the discrimination values according to every genre. Particularly, the scale of the discrimination values according to every genre and the number of features having zero discrimination should be discussed. Figure 3.13 illustrates the relation of the discrimination values against the ranking order according to each of the four genres Classical, Disco, Jazz and Rock and the three calculation models, while a statistical description regarding the underlying distribution of the discrimination values is given in table 3.2. As the range of discrimination values is very large, all values were normalized into the interval $[0, 1]$ considering the discrimination values of all 10 genres. These normalized discrimination values were used both in the illustration regarding the relation of the discrimination values against the ranking order and in the statistical description.

Considering the figure 3.13, considerably larger discrimination values were estimated according to classical music and all three calculation models comparing with the other genres. In terms of the Gain Ratio and Balanced Information Gain the following conclusion can be made. The number of features representing classical music having zero discrimination is smaller comparing with the other genres. In fact, quite a large number of features suggest zero discrimination in terms of Disco, Jazz and Rock with rock music having the smallest number of discriminative features. The number of features exhibiting large discrimination values is far smaller in the case of those three genres comparing with classical music. According to the ReliefF, features having a zero discrimination value do not exist for any of the four selected genres. An interesting fact is that the same order based on the highest discrimination value per genre is given for all calculation models.

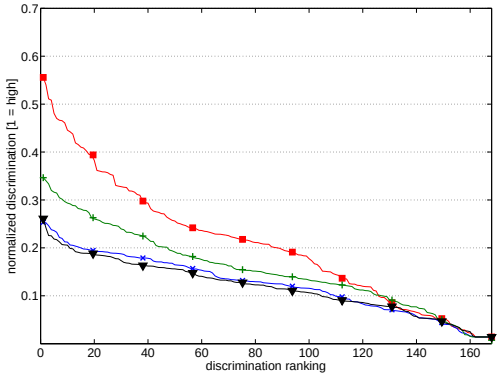
Figure 3.14 illustrates those features which are irrelevant with respect to the four selected genres of figure 3.13 and the three calculation models Gain Ratio, Balanced Information Gain



(a) Gain Ratio



(b) Balanced IG

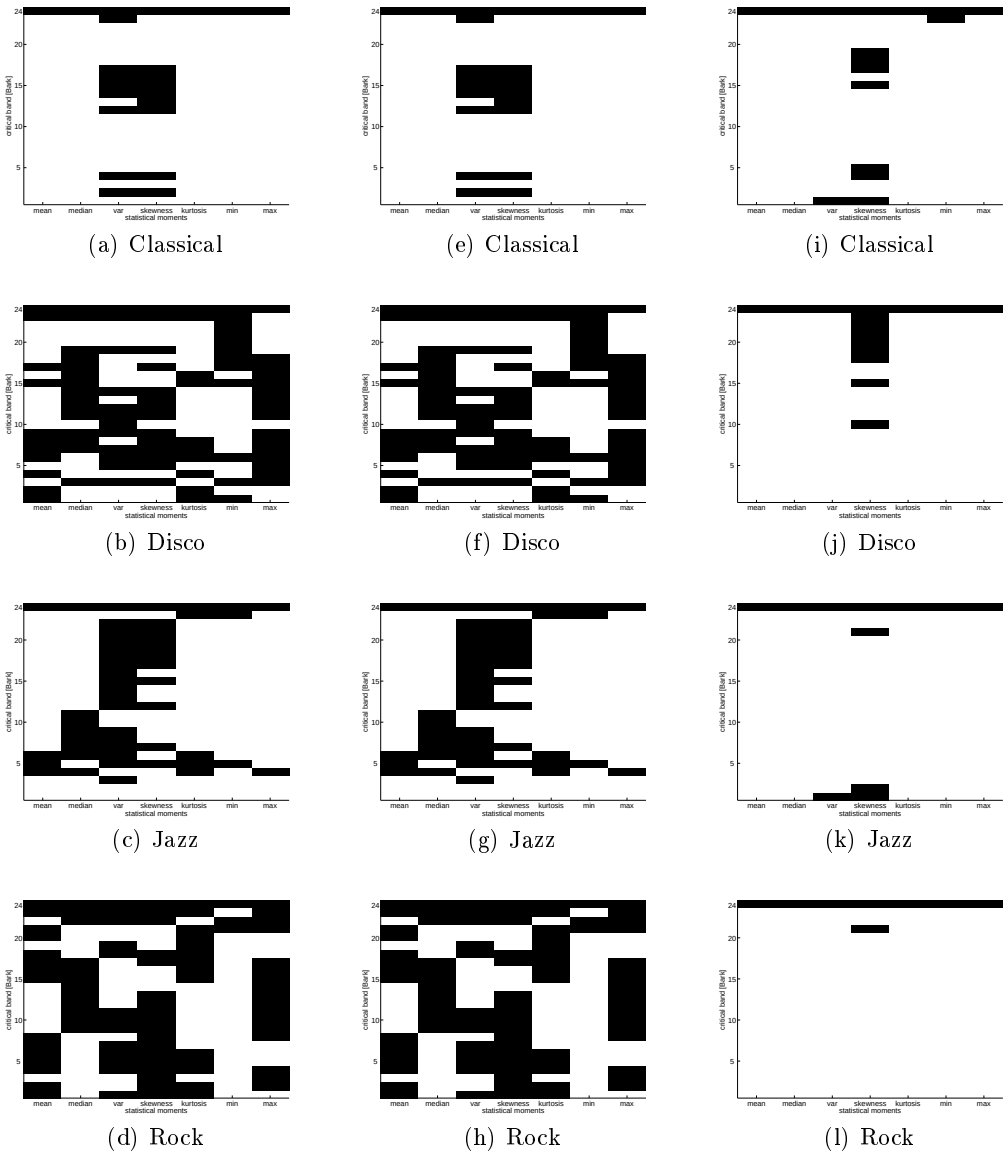


(c) ReliefF

Figure 3.13: Illustration of the discrimination values against the ranking order according to the Statistical Spectrum Descriptor on the GTZAN collection (normalized into the interval $[0, 1]$ for every calculation model separately). Each musical genre is illustrated by an individual color and symbol: ■ for **Classical**, × for **Disco**, + for **Jazz** and ▼ for **Rock**.

and ReliefF. As in the case of the Rhythm Pattern descriptor, features are also assumed to be irrelevant which have a discrimination value less than the discrimination value of a generic feature. This random feature is equally defined as in the case of the Rhythm Pattern descriptor. In fact, this generic feature only plays a role for the ReliefF model, as this model basically estimates non-zero discrimination values for every feature. As some features exhibit a smaller discrimination value than the value of the random feature in terms of the ReliefF model, it can be followed that some features related to the genres Disco, Jazz and Rock are actually irrelevant.

Table 3.2 presents a short statistical description of the discrimination results. The estimations of the statistical measures *mean*, *standard derivation*, *min* and *max value* were computed by using the normalized discrimination value range. Similar to the Rhythm Pattern descriptor the genre Classical is related the highest maximum and average discrimination values according to the Gain Ratio and the Balanced Information Gain, while the genre Pop is related the highest maximum and average discrimination values in terms of the ReliefF. Also an interesting fact is that the standard derivation of the discrimination values based on all three calculation models



(a) Classical (e) Classical (i) Classical

(b) Disco (f) Disco (j) Disco

(c) Jazz (g) Jazz (k) Jazz

(d) Rock (h) Rock (l) Rock

Figure 3.14: Irrelevant features of the Statistical Spectrum Descriptor according to the GTZAN collection. All *black* colored features are considered as irrelevant. Gain Ratio: (a) to (d), Balanced Information Gain: (e) to (h), ReliefF: (i) to (l).

is quite similar. Consequently, all three calculation models are quite robust in order to estimate the discriminative power of specific features to distinguish genres.

Like in the case of the Rhythm Pattern descriptor the next step of the discriminant analysis is to verify the difference of the discriminative feature patterns according to every genre. The possible existence of individual, i.e. significantly different, feature patterns was evaluated by using Kendall's statistical rank correlation test where p-value indicates a significant non-zero or zero correlation respectively. It must be reminded that the occurrence of tied ranks within the ranking sequence limits the expressiveness of the rank correlation test. This problem has already been discussed in 3.3.1.

	Gain Ratio				Balanced IG				ReliefF			
Genre	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**
Blues	0.04	0.05	0.00	0.22	0.04	0.06	0.00	0.27	0.16	0.06	0.00	0.32
Classical	0.42	0.25	0.00	1.00	0.31	0.24	0.00	1.00	0.22	0.13	0.02	0.56
Country	0.02	0.04	0.00	0.13	0.03	0.05	0.00	0.20	0.12	0.05	0.02	0.27
Disco	0.02	0.04	0.00	0.18	0.03	0.04	0.00	0.18	0.13	0.06	0.02	0.25
Hip hop	0.06	0.09	0.00	0.40	0.06	0.07	0.00	0.38	0.17	0.08	0.02	0.45
Jazz	0.09	0.08	0.00	0.30	0.10	0.09	0.00	0.36	0.16	0.08	0.01	0.35
Metal	0.14	0.17	0.00	0.66	0.14	0.16	0.00	0.59	0.18	0.12	0.01	0.58
Pop	0.17	0.20	0.00	0.96	0.18	0.19	0.00	0.92	0.23	0.17	0.01	1.00
Reggae	0.06	0.05	0.00	0.21	0.06	0.06	0.00	0.23	0.14	0.06	0.01	0.30
Rock	0.03	0.04	0.00	0.14	0.02	0.03	0.00	0.16	0.12	0.05	0.02	0.26

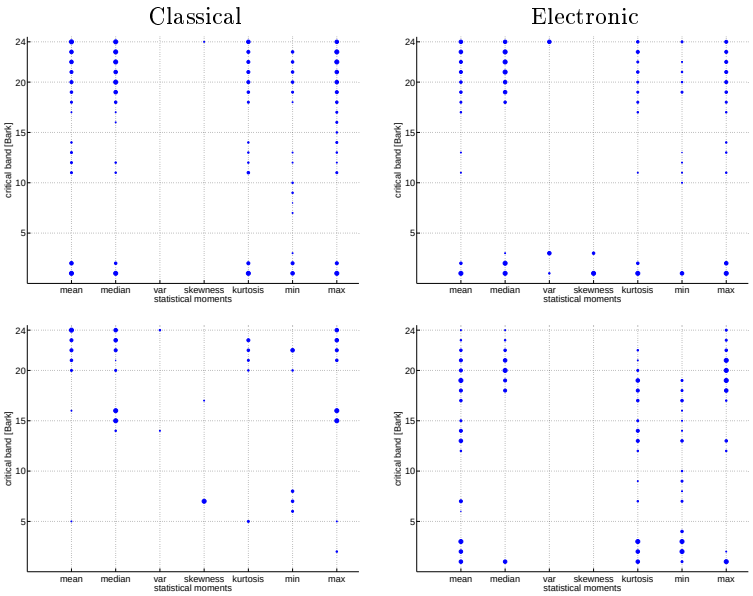
Table 3.2: Statistical summarization of the discrimination values according to the Statistical Spectrum Descriptor and the GTZAN collection (normalized into the interval $[0, 1]$ for every calculation model separately). Only those discrimination values were considered which were originally non-zero.

In table 3.4, the p-values of the rank correlation tests based on every tested genre pair are listed. As the significance level of all rank correlation tests was defined by $\alpha = 0.05$ every p-value grater than α indicates a zero correlation of the underlying discriminative feature patterns. In the case of zero correlation individual feature patterns were concluded. According to the Gain Ratio and the Balanced Information Gain the number of actual individual feature patterns is almost similar. Moreover, for both calculation models holds true that considerably more individual feature patterns are recognized than in terms of the Rhythm Pattern descriptor. As these two calculation models estimate a large number of features having zero discrimination, the test results unfortunately do not represent all possible individual feature ranks. Nevertheless, in terms of the genres Classical, Jazz and Pop the feature patterns significantly differ to most of the other genres. The test results according to the discriminative patterns estimated by the ReliefF suggest considerably more individual feature patterns. This means that many genres are actually represented by individual feature patterns which significantly differ to those of other genres. Even feature patterns of genres like Blues and Jazz as well as Pop and Rock significantly differ although often similar rhythm styles are related to each of these two genre pairs.

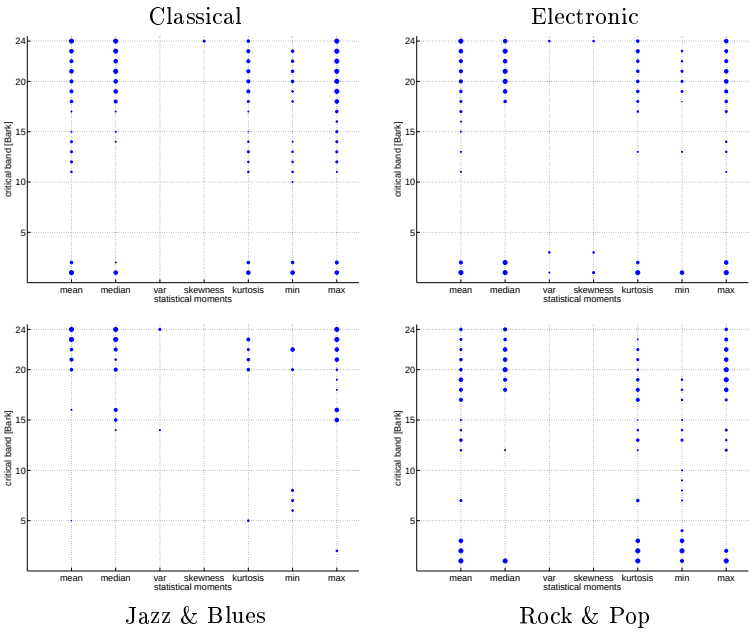
Since all previous conclusions only refer to the GTZAN collection, the question raise whether and how the performances of the three heuristic discrimination models vary in the case of the ISMIR 2004 Genre and Rhythm collections. Thus, similar and diverging conclusions concerning the identification of genre-specific feature patterns based on these collections will be shortly discussed comparing with those conclusions based on the GTZAN collection.

ISMIR 2004 Genre

In order to compare the discriminative feature patterns according to both the GTZAN and the ISMIR 2004 Genre collections, four specific genres were selected which are more or less



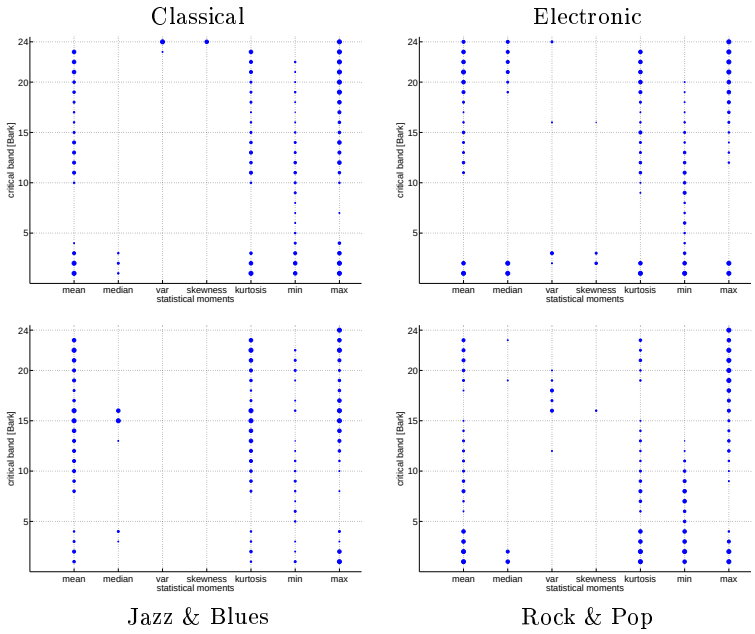
(a) Gain Ratio



(b) Balanced Information Gain

Continued on the next page ...

represented in both collections. According to the ISMIR 2004 Genre collection these genres are Classical, Electronic, Jazz & Blues and Rock & Pop. Since at least a partial correlation can be assumed to the genres Classical, Disco Jazz and Rock of the GTZAN collection respectively, a basic comparison of the respective discriminative feature pattern was performed although only



(c) ReliefF

Figure 3.15: Inter-genre comparison of discriminative features according to the Statistical Spectrum Descriptor on the ISMIR 2004 Genre collection. Three calculation models were used where figure (a) represents the Gain Ratio, figure (b) corresponds to the Balanced Information Gain and figure (c) is related to the ReliefF. In order to provide a clear visualization, 50 % of those features were taken into account only which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.

in the case of the genre Classical a sufficient correlation genre can be assumed.

Figure 3.15 illustrates the discrimination values according to the four selected genres of the ISMIR 2004 Genre collection based on the three heuristic discrimination models Gain Ratio, Balanced Information Gain and ReliefF. As already done in similar illustration, only 50 % of all actually discriminative features were plotted as filled dots with varying size, while a larger size indicates better discrimination of the corresponding feature. According to the Gain Ratio and the Balanced Information Gain, the discrimination results diverge between almost all related genres of the GTZAN and the ISMIR 2004 Genre collection. Only for the genre Classical a high degree of similarity can be observed. In particular the features related to the measures mean, median and max value were estimated to be discriminative according to both collections. As already seen before, features related to the critical band 24 Bark were also estimated to be discriminative in the case of ISMIR 2004 Genre collection but not in the case of the GTZAN collection.

According to the ReliefF model, similar discrimination results regarding the both classical genres can also be concluded where the degree of similarity is even higher as in the case of the other two calculation models. Contrarily, a considerably degree of similarity can also be

concluded according to the other three genres as well. In fact, each comparison reveals a large number of features being discriminative in the feature patterns according to each of the correlated genres. Considering the comparison Jazz & Blues versus Jazz for instance, a relatively high consistency exists for those discriminative features which are related to the statistical measures kurtosis, min and max value. Thus, it can be followed that the discrimination results computed by the ReliefF are actually more consistent on partially correlated music collections. A reason explaining the consistency of the ReliefF model on partially correlated music collections may be the point of view how ReliefF estimates the feature-genre dependency. In [48], calculation models implementing the impurity function are referred as the global point of view to estimate feature-genre dependency, while for the ReliefF model a local point of view was concluded as it also takes the context of other features into account. It appears that the global point of view disregarding local feature dependencies depends on the characteristics of the underlying music collection. In particular partially correlated genres like for instance Disco and Electronic could be a source of different discriminative feature patterns, as musical pieces also affect the computation of the genre discrimination which are not consistently assigned to both genres. Also possible outliers according to the two correlated genres are crucial. For example the genres Disco and Electronic of the GTZAN and the ISMIR 2004 Genre collections respectively are only partial correlated and the corresponding discriminative feature patterns differ more in the case of the Gain Ratio and the Balanced Information Gain comparing with the ReliefF. The ReliefF estimates the feature-genre dependency over a smaller part of the input space as it implements a nearest-neighbor algorithm. The dependencies between the features also play a role in the computation of the genre discrimination. Consequently, the ReliefF appears to be more consistent on partially correlated music collections because the data instances might be more correlated within local areas of the corresponding input spaces as well as less outliers might also be included. It can be followed that the ReliefF calculation model guarantees more consistent discrimination results according to different music collections.

Another important fact can be concluded on the discrimination results regarding the ISMIR 2004 Genre collection which is also valid for the GTZAN collection. It can be followed that the discriminative feature patterns computed by the three heuristic calculation models Gain Ratio, Balanced Information Gain and ReliefF imply a notable degree of similarity. On the on hand, the Gain Ratio and the Balanced Information Gain estimate almost the same features to be discriminative and also the discrimination values of those features only marginally vary. As the similar performance of these two calculation models has already been concluded in the case of the Rhythm Pattern descriptor, the different approaches of normalizing multi-valued features implemented in those two calculation models do not decisively affect the computation. Section 3.2 reviews the problem of multi-valued features in terms of the Information Gain and the two normalization heuristics representing the Gain Ratio and the Balanced Information Gain. On the other hand, the ReliefF model estimates more differing discriminative feature patterns but nonetheless a certain degree of similarity to the feature patterns computed by the other two calculation models can be recognized. The corresponding discrimination results according

to the GTZAN collection also imply this conclusion. Thus, the discriminative feature patterns computed by each of the three calculation models appear to be more consistent and the different approaches of estimating the feature-genre dependency do not affect the computation that strong comparing with both the Rhythm Pattern descriptor and the Rhythm Histogram descriptor. A possible explanation has already been given in terms of the discussion regarding the results based on the GTZAN collection.

Table 3.5 lists the p-values of the rank correlation test due to pair wise genre tests according to the ISMIR 2004 Genre collection. It can be followed that more individual discriminative feature patterns exist regarding the Statistical Spectrum Descriptor than in the cases of the Rhythm Pattern descriptor. The number of individual discriminative feature patterns is quite balanced but not similar according to the Gain Ratio and the Balanced Information Gain as well as the ReliefF. In particular the feature pattern estimated by each of the three calculation models and representing the genres Classical and World appears to be individual comparing with most of the other genres.

ISMIR 2004 Rhythm

As in the discriminant analysis of the Rhythm Pattern descriptor the performances of the three heuristic discrimination models should be shortly discussed on the ISMIR 2004 Rhythm collection. Since this music collection contains Latin and Ballroom dance music only, genre-to-genre comparisons regarding the respective feature discriminative patterns of the other music collections are not possible. Nevertheless, some conclusions can also be verified on this collection.

The discriminative feature patterns computed by the Gain Ratio and the Balanced Information Gain are also very similar with respect to the ISMIR 2004 Rhythm collection. In particular almost the same features were defined to be discriminative, while the discrimination values of those features differ because of the different approaches of normalizing multi-valued features utilized in the Gain Ratio and the Balanced Information Gain. As the differences in the discrimination values are considerably limited according to all three music collections, it can definitely be assumed that both the Gain Ratio and the Balanced Information Gain compute sufficiently similar discriminative feature patterns. Actually, this conclusion is also valid for the Chi-square and the Information Gain which both also implement the impurity function.

Contrarily, the ReliefF calculation model computes discriminative feature patterns diverging from those computed by the Gain Ratio and the Balanced Information Gain. Different discriminative features as well as varying discrimination values could be recognized. This conclusion is only partially valid in terms of the other two collections where a certain degree of similarity could be concluded for some genres like Classical and Electronic. It also disagrees with the observation according to both the Rhythm Pattern descriptor and the Rhythm Histogram descriptor where at least for some genres a notable similarity could be concluded among to the discrimination results of the three calculation models. Another interesting fact is that the features corresponding to the statistical measures variance and skewness mostly exhibit very low discrimination values. This conclusion could also be made in terms of the other two music collections. Consequently,

a general irrelevance of the measures variance and skewness can be assumed as almost all features related to these two measure exhibit low or zero discrimination according to almost all one-vs.-rest genre situations on all three music collection.

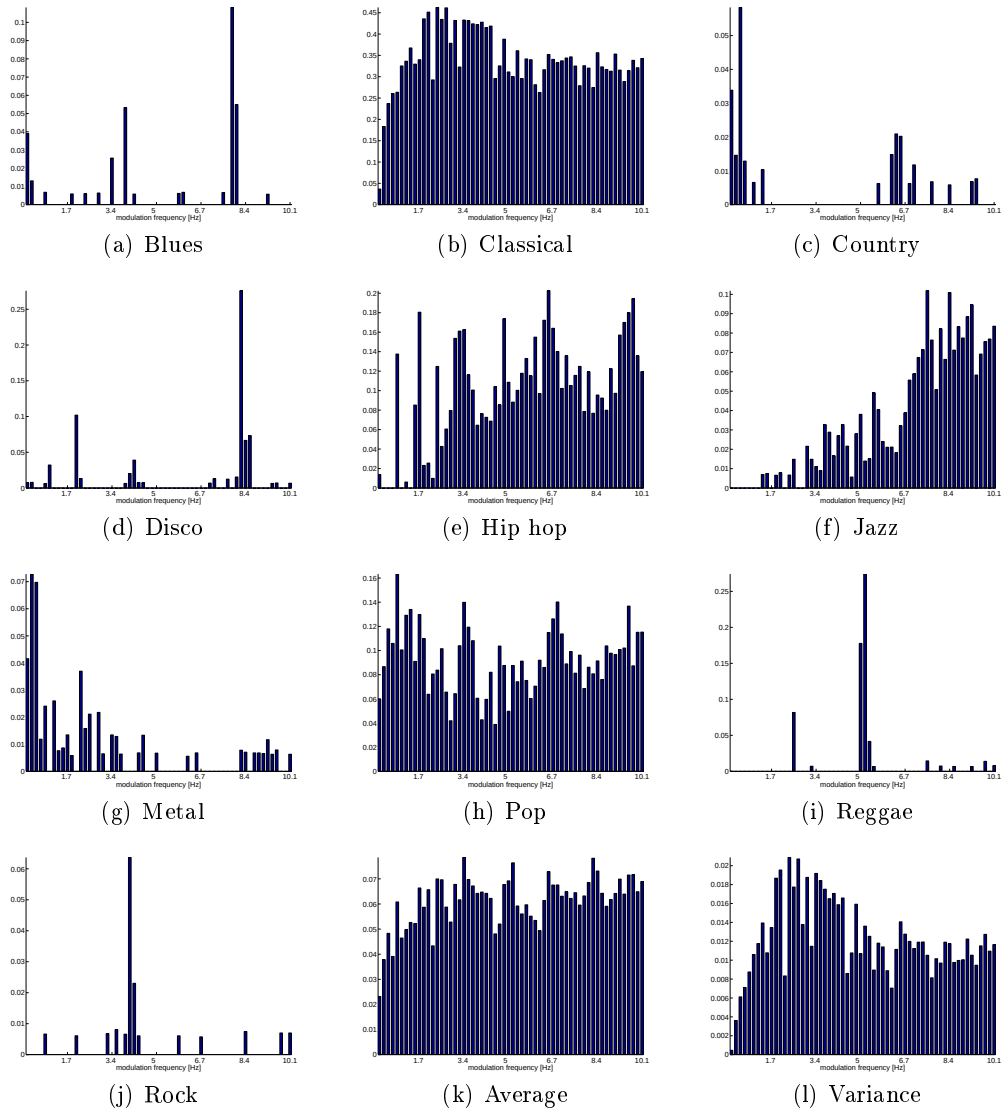
3.3.3 Rhythm Histogram

Although the extraction of the Rhythm Histogram descriptor utilizes the same psycho-acoustic transformation as both the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor, the features are not directly related to the 24 critical bands but represent 60 modulation frequency bins instead. Further details concerning the extraction of the Rhythm Histogram descriptor can be reviewed in section 2.5. Usual bar plots will be used to visualize the discrimination values of every feature. Beginning with the Gain Ratio calculation model, figure 3.16 illustrates the discrimination results according to all 10 genres of the GTZAN collection. Additionally, the average and the variance of the discrimination values computed over all genres are illustrated in this figure.

As already observed in context of both the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor, two groups of genres can be identified which differ in the number of features being discriminative. The genres Classical, Hip hop, Jazz and Pop are represented by a large number of discriminative features. It can be followed that a broad range of of modulation frequencies characterizes the underlying genres. Those four genres are also related to a large number of discriminative features according to the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor. The second group contains the genres Blues, Country, Disco, Metal, Reggae and Rock. Each genre of this group is related to a significantly smaller number of discriminative features. Moreover, a considerably large number of features are actually irrelevant as they exhibit a zero discrimination value. A possible explanation of this interesting observation is that genres like Classical or Pop usually cover a wider range of different rhythmic styles and tempo variations comparing with very specific genres like Blues, Disco and Metal for instance. In fact, musical pieces related to these more specific genres are particularly characterized by a limited variation of rhythmic styles. Another key difference to the genres of the first group is that some very specific features exhibit very high discrimination values comparing with the other features having also a non-zero discrimination values. This difference in the discrimination values between the most discriminative features and less discriminative feature is large for according to all genres of the second group.

In fact, the partitioning of genres based on the number of discriminative features can also be observed in terms of the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor. Since the three descriptors are strongly related to each other as they utilize both the same computation of the frequency spectrum and the same psycho-acoustic transformation, this analog observation is not surprising. Moreover, the considerable numerical difference in the discrimination values of the features among the genres of the second group is also revealed in the corresponding discrimination results of the other two descriptors.

An interesting fact concerning the group containing genres with a large number of discrim-



(a) Blues (b) Classical (c) Country

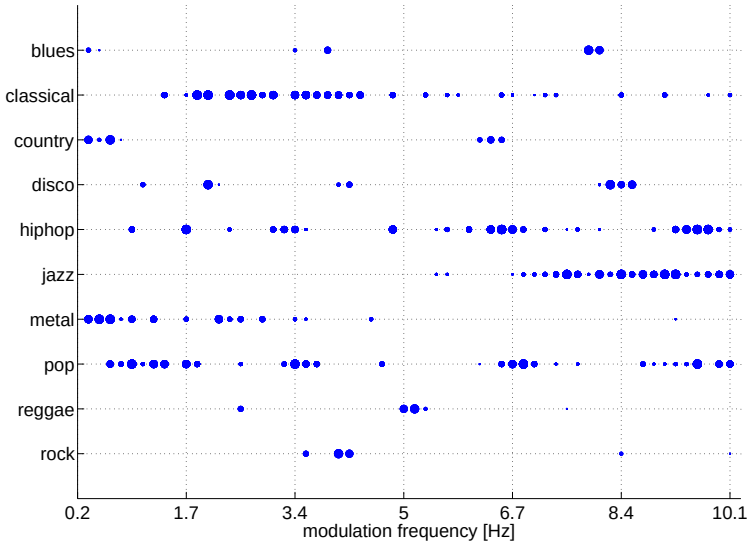
(d) Disco (e) Hip hop (f) Jazz

(g) Metal (h) Pop (i) Reggae

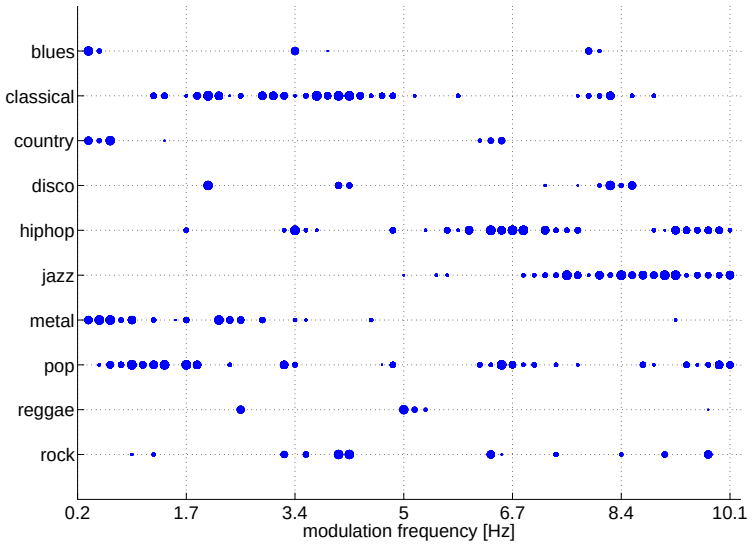
(j) Rock (k) Average (l) Variance

Figure 3.16: Discriminative features of the Rhythm Histogram descriptor according to the music collection GTZAN. The Gain Ratio model was used to compute all discrimination values where a more discriminative feature is denoted by a higher bar. Figures (k) and (l) represent the average and variance results over all genres.

inative features is that the scale of discrimination values is quite balanced between the features. This means that a considerable number of features and, respectively, modulation frequencies exhibiting similar discrimination values actually exist. As tempo characteristics and in particular rhythmic styles vary among musical pieces according to these genres it appears that a clear decision which particular modulation frequencies characterizing a specific genre can not be made. According to the genres Classical, Hip hop and Pop this number of features possessing similar discrimination values is quite large, while for the genre Jazz this number of features is smaller. Contrarily, in the case of the second group of genres very few features can be identified to uniquely characterize the corresponding genre. According to Blues and Disco a single but dif-



(a) Gain Ratio

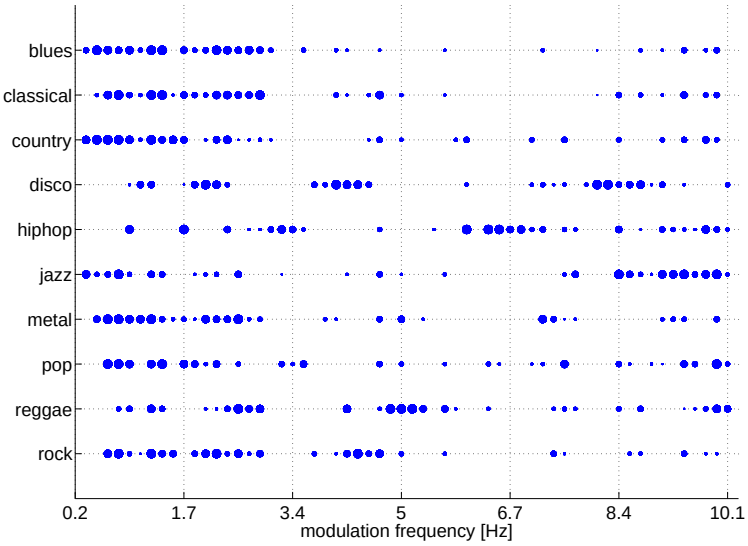


(b) Balanced Information Gain

Continued on the next page ...

ferent modulation frequency close to 8 Hz achieves such a remarkable genre characterization. In the case of the genres Country and Metal the lowest modulation frequencies appear to be particularly discriminative, while single modulation frequencies close to 5 Hz and 4 Hz are particularly discriminative for the genres Reggae and Rock, respectively.

To summarize the observations according to the discrimination results based on the Gain Ratio calculation model, the number of discriminative features considerably varies between the genres, while individual features could be identified for some genres which particularly char-



(c) ReliefF

Figure 3.17: Inter-genre comparison of discriminative features according to the Rhythm Histogram based on the GTZAN collection. Three calculation models were used where figure (a) represents the Gain Ratio, figure (b) corresponds to the Balanced Information Gain and figure (c) is related to the ReliefF. In order to provide a clear visualization, only 50 % of those features were taken into account which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.

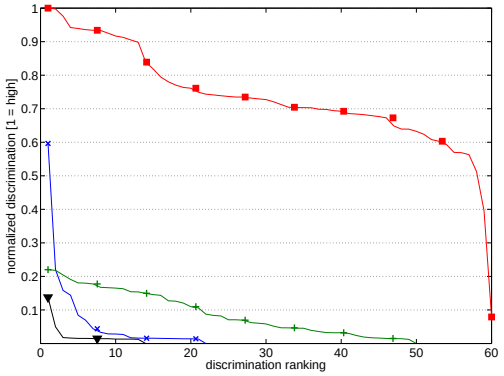
acterize the corresponding genres. The next step of the discriminant analysis is to verify the conclusions according to the Gain Ratio on the corresponding discrimination results based on the Balanced Information Gain and the ReliefF. Since the discrimination results according to the two heuristic discrimination models Chi-Square and Information Gain only marginally differs to the results based on the Gain Ratio or the Balanced Information Gain, these two calculation models will not be considered any further. Figure 3.17 illustrates the discriminative features computed by the Gain Ratio, the Balanced Information Gain and the ReliefF for each genre respectively. In order to provide better visualization, only 50 % of all actually discriminative features were plotted as filled dots with varying size. A larger size indicates better discrimination of corresponding feature.

The discrimination results based on the Gain Ratio, which are illustrated in (a), are already discussed in terms of the figure 3.16 and therefore the same conclusions are valid. The discrimination results according to the Balanced Information Gain are depicted in figure (b). Similar to the Gain Ratio the Balanced Information Gain also implements the impurity function which means basically that both calculation models utilize the entropy measure to estimate the dependency between a specific feature and a genre. Therefore it is not surprising that the corresponding discrimination results are very similar where almost the same discriminative features are recognized. In fact, only the corresponding discrimination values of same feature vary. The

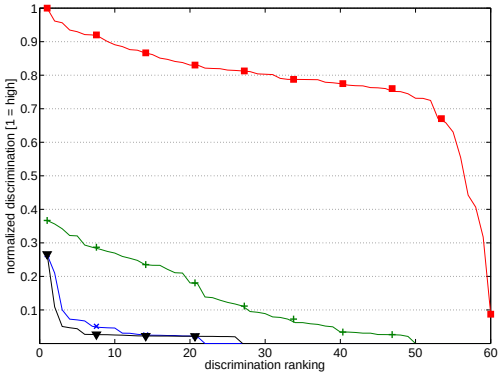
genre Blues represents a good example in which the same features were actually selected to be discriminative but the discrimination values of these features differ considerably. The reason of such differences in the discrimination values is due to the applied normalization regarding multi-valued features which differs among the Gain Ratio and the Balanced Information Gain. As both calculation models basically utilize the Information Gain which tends to overestimate multi-valued features, this normalization is necessary although different approaches actually exist. The normalization of multi-valued features implemented in each of these two calculation models is described in section 3.2. Nevertheless, the partitioning of genres into two groups depending on the number of discriminative features is also clearly observable in the discrimination results based on the Balanced Information Gain.

According to the ReliefF calculation model illustrated in figure (c) diverging discrimination results were computed comparing with the results based on the Gain Ratio and the Balanced Information Gain. This difference is not only constituted in terms of numerical variations regarding the estimated discrimination values. Considering the discriminative features of each genre according to the three calculation models, it can be followed that even different features are selected to be discriminative. Only in the case of the genres Pop and Metal the selected features coincide by a notable degree. It clearly holds true that the number of discriminative features does not vary as much between the genres as in the case of the results based on the calculation models implementing the impurity function. In fact, 8 of 10 genres are related to a large number of discriminative features which is even quite similar among these genres. Only in the case of the genres Disco and Hip hop a smaller number of discriminative features were estimated although the number is only slightly smaller. Nevertheless, the number of discriminative features related to Disco is still considerably larger than in the case of the other two calculation models. This means that some genres are also represented by a large number of discriminative features although musical pieces of those genres usually include a slight variation of rhythmic styles and beats. Another interesting fact is that for 6 of 10 genres the most discriminative features refer to modulation frequencies from 0.2 to 3 Hz. On the other hand, the modulation frequency range from 3 to 6 Hz appears to be relevant for 4 genres only. The different discrimination results of the ReliefF are not completely surprising, since this model follows a diverging concept of measuring the dependency between a feature and a genre.

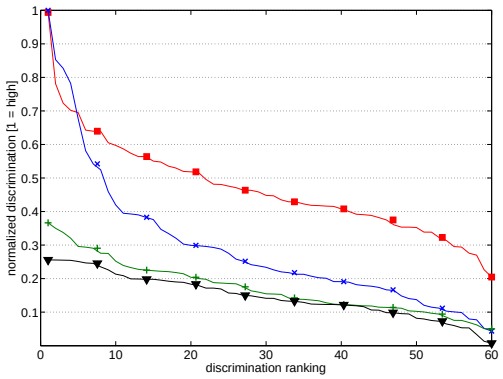
It could already be concluded that the scale of the discrimination values varies according to every genre and also that the degree of difference in the discriminative values depends on the underlying genre. To emphasize the difference in the scale of the discrimination values among the genres, both an illustration regarding the relation of the discrimination values against the ranking order for the genres Classical, Disco, Hip hop and Rock as well as a statistical description of the discrimination values are given in the following. Figure 3.18 visualizes the relation of the discrimination values against the ranking order according to each of the four genres Classical, Disco, Jazz and Rock based on the three heuristic discrimination models Gain Ratio, Balanced Information Gain and ReliefF respectively. Table 3.3 lists the estimations of four key statistical descriptors in order to give a better insight into the relation of the discrimination



(a) Gain Ratio



(b) Balanced IG

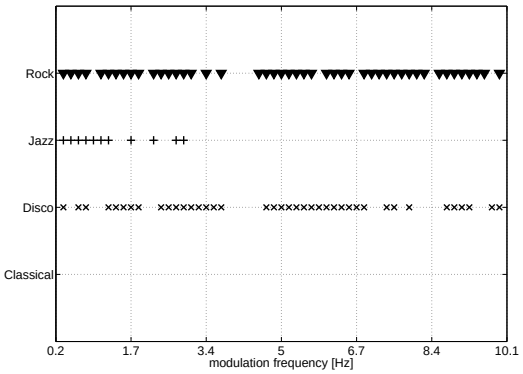


(c) ReliefF

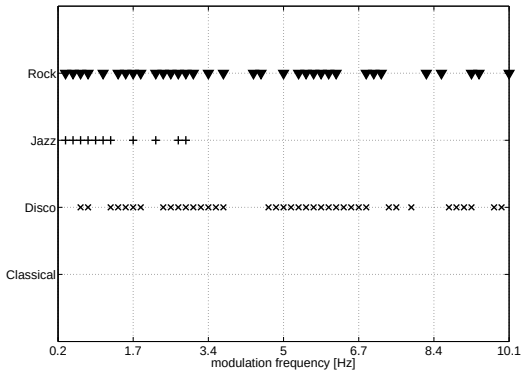
Figure 3.18: Illustration of the discrimination values against the ranking order according to the Rhythm Histogram Descriptor on the GTZAN collection (normalized into the interval $[0, 1]$ for every calculation model separately). Each musical genre is illustrated by an individual color and symbol: ■ for **Classical**, × for **Disco**, + for **Jazz** and ▼ for **Rock**.

values to corresponding genres. As the range of discrimination values is very large, all values were normalized into the interval $[0, 1]$ considering the non-zero discrimination values of all 10 genres. These normalized discrimination values were used to establish both the figure 3.18 and the table 3.3.

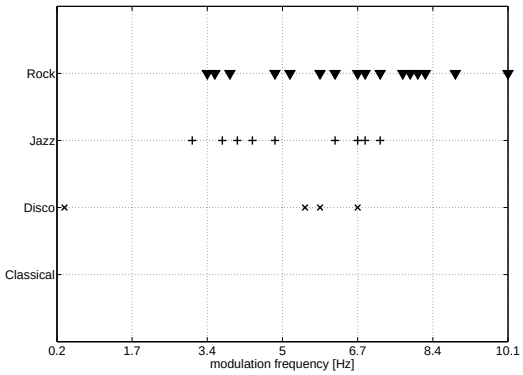
As already observed in terms of the visualization of the discriminative features, all features and, respectively the entire range of modulation frequencies are discriminative according to the genre Classical. Moreover, a larger number of features exhibit similar discrimination values according to the Gain Ratio and Balanced Information Gain, while the distribution according to the ReliefF decreases less rapidly. Another interesting observation is the number of features possessing a zero discrimination values which actually coincides with the corresponding results based on the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor. In the case of the Gain Ratio as well as the Balanced Information Gain quite a large number of features have a zero discrimination value according to the genres Disco, Jazz and Rock. It was assumed that zero discrimination refers to a irrelevance of distinguishing genres. Additionally, figure 3.19



(a) Gain Ratio



(b) Balanced Information Gain



(c) ReliefF

Figure 3.19: Irrelevant features of the Rhythm Histogram descriptor according to the GTZAN collection. Gain Ratio (a), Balanced Information Gain: (b), ReliefF: (c).

illustrates these irrelevant features for each genre and calculation model. According to Disco and Rock the irrelevant features are quite the same but only in terms of the calculation models Gain Ratio and the Balanced Information Gain. Also the irrelevant features according to the ReliefF calculation model are illustrated in figure 3.19. Although every feature has actually a non-zero discrimination value, some features are considered to be irrelevant because they exhibit a discrimination value less than the discrimination value of a random probe feature. The random probe feature was equally generated as in the case of the other two descriptors. Since some features were less discriminative than the random probe feature in terms of the ReliefF model, it was followed that those features related to the genres Disco, Jazz and Rock were irrelevant respectively.

In the next step of the discrimination analysis, the discriminative feature patterns according to every genre were evaluated whether the difference in the feature patterns is significant. The verification was done for each of the three calculation models individually. Like in the case of the Rhythm Pattern descriptor the possible existence of individual, i. e. significantly different, feature patterns was evaluated by using Kendall's statistical rank correlation test where p-value indicates

	Gain Ratio				Balanced IG				ReliefF			
Genre	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**	$\hat{\mu}$	$\hat{\sigma}$	**min**	**max**
Blues	0.04	0.06	0.00	0.22	0.02	0.04	0.00	0.13	0.16	0.09	0.02	0.40
Classical	0.73	0.16	0.07	1.00	0.77	0.16	0.07	1.00	0.47	0.14	0.20	0.99
Country	0.02	0.03	0.00	0.12	0.04	0.06	0.00	0.22	0.15	0.08	0.00	0.49
Disco	0.06	0.13	0.00	0.59	0.04	0.06	0.00	0.25	0.29	0.20	0.04	1.00
Hiphop	0.23	0.10	0.00	0.43	0.31	0.12	0.00	0.50	0.59	0.17	0.31	0.95
Jazz	0.08	0.07	0.00	0.21	0.14	0.11	0.00	0.35	0.17	0.08	0.05	0.37
Metal	0.02	0.04	0.00	0.15	0.04	0.06	0.00	0.26	0.15	0.08	0.02	0.38
Pop	0.19	0.06	0.07	0.35	0.33	0.07	0.17	0.56	0.55	0.17	0.26	0.91
Reggae	0.11	0.19	0.00	0.59	0.07	0.09	0.00	0.26	0.27	0.16	0.08	0.88
Rock	0.01	0.04	0.00	0.13	0.02	0.05	0.00	0.25	0.15	0.06	0.01	0.26

Table 3.3: Statistical summarization of the discrimination values according to the Rhythm Histogram descriptor and the GTZAN collection (normalized into the interval $[0, 1]$ for every calculation model separately). Only those discrimination values were considered which originally were non-zero.

a significant non-zero or zero correlation respectively. It must be reminded that the occurrence of tied ranks within the ranking sequence limits the expressiveness of the rank correlation test. This problem has already been discussed in 3.3.1.

Table 3.4 lists the p-values of all tested genre pairs according to the three heuristic discrimination models Gain Ratio, Balanced Information Gain and ReliefF on the GTZAN collection. The significance level is defined with $\alpha = 0.05$ and a p-value greater than α indicates that the corresponding two genres are represented by individual discriminative feature patterns each. It is not surprising that the genres Classical and Pop which are related to the largest number of discriminative features are also represented by the largest number of individual feature patterns. For these two genres the number of potential individual feature patterns is similar to those based on the Statistical Spectrum Descriptor. The test results based on the Gain Ratio and the Balanced Information Gain are very similar. Almost all recognized individual feature patterns are being individual for these two calculation models. According to the ReliefF the test results offer far more individual feature patterns. As all features exhibit a non-zero discrimination value in terms of the ReliefF, which can be concluded in the figure 3.18, the rank correlation test unfolds its full expressiveness. Thus, far more individual discriminative feature patterns can be concluded for various genres. In particular some very interesting genre pairs are actually represented by significantly different feature patterns. For instance Blues vs. Jazz, Metal vs. Rock, Country vs. Rock or Disco vs. Hip hop are actually among those genre pairs, although each of those pairs contains genres which are somehow related to each other in terms rhythmic styles and number of beats. The discriminative feature patterns according to the genres Hip hop and Reggae are also individual with respect to all other genres.

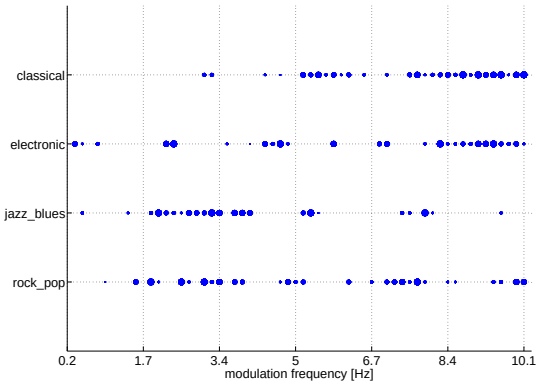
ISMIR 2004 Genre

Since the genres of the ISMIR 2004 Genre and the GTZAN collection only partially correlate, four specific genres were selected which are more or less represented in both collections. According to the ISMIR 2004 Genre collection these genres are Classical, Electronic, Jazz & Blues and Rock & Pop. As at least a partial correlation can be assumed to the genres Classical, Disco Jazz and Rock of the GTZAN collection respectively, a comparison of the respective discrimination results was done although only both classical genres coincide sufficiently.

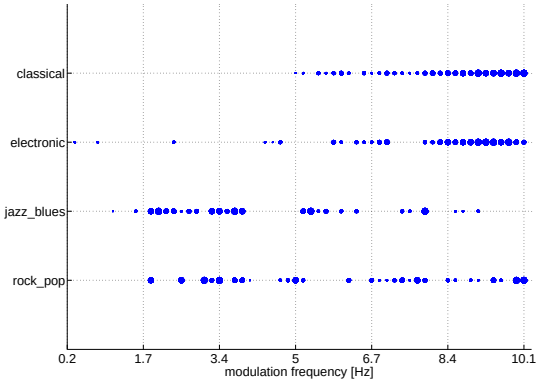
According to the ISMIR 2004 Genre collection the discrimination results based on the Gain Ratio and the Balanced Information Gain on the one hand, and the ReliefF on the other hand also diverge. This means that the different concepts of estimating the dependency between a specific feature and a genre, namely the impurity function versus an approach based on the nearest-neighbor algorithm, actually influence the computation of the discrimination values. It can be followed that this difference in the discriminative patterns according to a specific genre will be observed in further music collections either. As the difference in the discrimination results based on the Gain Ratio and the Balanced Information Gain are also limited in the case of the ISMIR 2004 Genre collection, these two calculation models appear to be consistent with a high degree regarding the estimation of discriminative feature patterns. Since this conclusion is also valid for the ISMIR 2004 Rhythm collection, this equal performance can be concluded generally.

Figure 3.20 illustrates the discrimination results of the selected genres based on the three heuristic discrimination models Gain Ratio, Balanced Information Gain and ReliefF. Only 50 % of all actually discriminative features were plotted as filled dots with varying size. Again, a larger size indicates better discrimination of the corresponding feature. At first, the results confirm the high similarity of the discriminative feature patterns computed by the Gain Ratio and the Balanced Information Gain. It can be observed that the estimated discriminative features of both calculation models mostly coincide although the actual discrimination values of a large number of features vary. In the case of the genre Classical a quite large number of features was selected to be discriminative which actually affirmed the corresponding observation according to the GTZAN collection. Also the number of discriminative features according to the other three genres is almost as large as for classical music. Comparing the results of the related genres of the GTZAN collection, this is a contradictory observation but might be related to the compared genres. In fact, the genres Electronic, Jazz & Blues and Rock & Pop are more general and contain more variation of rhythmic styles as the more specific genres Disco, Jazz and Rock of the GTZAN collection. Thus, the diverging observation regarding the actual number of selected features is not surprising.

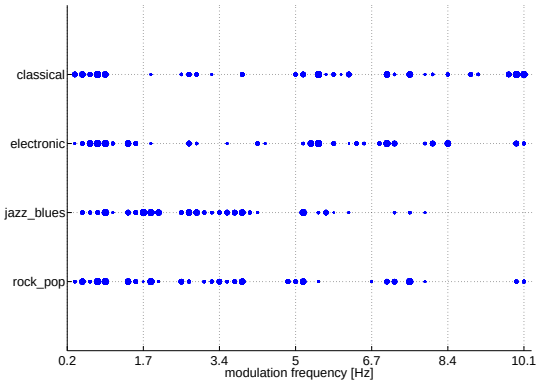
An interesting fact is that although the discrimination results based on the Gain Ratio and the Balanced Information Gain diverge with respect to the corresponding results according to the GTZAN collection, in the case of the ReliefF this difference in the discrimination values between the two collections appears to be more limited. In fact, many features are selected to be discriminative according to both collections, although the corresponding discrimination values



(a) Gain Ratio



(b) Balanced IG



(c) ReliefF

Figure 3.20: Inter-genre comparison of discriminative features according to the Rhythm Histogram descriptor on the ISMIR 2004 Genre collection. In order to provide a clear visualization, 50 % of those features were taken into account only which had a non-zero discrimination value. The size of every dot indicates the degree of discrimination the corresponding feature has where a large size implies higher discrimination.

do vary. In particular for the genre Classical or Rock & Pop a partial similarity of selected discriminative features can be concluded with respect to the genres Classical and Rock of the GTZAN collection. Considering the corresponding results based on the Gain Ratio and the Balanced Information Gain, such a degree of similarity can be hardly seen at all and even orthogonal discriminative feature patterns exist like those for the two classical genres for instance. A reason explaining the consistency of the ReliefF model may be the point of view how ReliefF estimates the feature-genre dependency. As already mentioned before, a calculation model implementing the impurity function follows the global point of view to estimate feature-genre dependency. In particular, this means that they use the entire input space for the computation. Contrarily, the ReliefF model implements a nearest-neighbor algorithm and therefore the estimation of feature-genre dependencies are more focused on local areas of the input space. Additionally, it also takes the context of other features into account. As already discussed for the Statistical Spectrum Descriptor, these different concepts of dependency estimation highly depend on the characteristic of the underlying music collection where a global point of view is much more influenced by diverging characteristics of music collections. Thus, the ReliefF calculation model guarantees more consistent discrimination results on partially correlated music collections.

Table 3.5 lists the p-values of the rank correlation test due to pair wise genre tests according to the ISMIR 2004 Genre collection. It can be followed that more individual discriminative feature patterns exist regarding the Rhythm Histogram descriptor than in the case of the Rhythm Pattern descriptor. According to the Statistical Spectrum Descriptor this number of individual feature patterns is quite balanced. Again, the number of individual discriminative feature patterns is considerably smaller according to the Gain Ratio and the Balanced Information Gain. As already mentioned in terms of the rank correlation test based on the GTZAN collection, these two calculation models estimates zero discrimination for quite many features which reduce the effectiveness of the entire correlation test. Nevertheless, the genres Jazz & Blues as well as Rock & Pop are represented by individual feature patterns with respect to almost all genres. According to the ReliefF, most of the genres are related to individual feature patterns and this is a similar conclusion as in the case of the GTZAN collection.

ISMIR 2004 Rhythm

The discrimination results according to the ISMIR 2004 Rhythm collection are obviously difficult to compare as completely different genres are included than with the other two collections. Nevertheless, it can also be concluded that respective discrimination results based on the Gain Ratio and the Balanced Information Gain only marginally diverge. This conclusion is also actually valid in terms of the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor. In particular both calculation models estimated almost the same features to be discriminative. Since this conclusion is also valid according to the other two descriptors and all three examined music collection, a consistent similarity regarding the estimation of discriminative features can be expected for all the models implementing the impurity function. Possible different approaches of normalizing multi-valued features, which are implemented in the Gain Ratio and the Balanced

Information Gain, do not sufficiently affect the estimation of feature-genre discrimination. Only variations of the actual discrimination values were observed. But this is not surprising considering the fact that the original feature contribution based on the Information Gain is differently normalized. More details on the specific approaches of normalizing multi-valued features and the actual implementations within the calculation models used in this thesis can be reviewed in section 2.3. Another interesting fact is that both the Gain Ratio and the Balanced Information Gain estimated quite a large number of features exhibiting zero genre discrimination, while all features possess non-zero discrimination values according to the ReliefF calculation model.

The discrimination results computed by the ReliefF model implied a surprising similarity to the results based on the other two calculation models. In fact, a considerable similarity was recognizable for 5 of the 8 genres. All three calculation models delivered quite similar estimates for the most discriminative features. In particular the most discriminative feature was always identical according to all calculation models. On the other hand, the discrimination results according to the genres Rumba and Slow Waltz diverged considerably. This partial correlation of the discrimination results based on the three calculation models has already been observed in the case of the Rhythm Pattern descriptor. Contrarily, the discrimination results based on the Gain Ratio and Balanced Information Gain on the one hand, and the ReliefF on the other hand diverged according to the Statistical Spectrum Descriptor.

Gain Ratio

	Blues			Classical			Country			Disco			Hip hop			Jazz			Metal			Pop			Reggae			Rock		
	RP	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**
Blues		—		.00	**.51**	**.24**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.06**	**.41**	**.16**	.00	.00	.00	.00	.00	.00
Classical	.00	**.51**	**.24**		—		.00	**.98**	**.07**	.00	**.62**	**.06**	.00	**.21**	**.08**	**.43**	**.49**	**.08**	.01	**.95**	**.61**	**.12**	**.68**	**.13**	.00	**.55**	**.08**	.00	**.22**	**.23**
Country	.00	.00	.00	.00	**.98**	**.07**		—		.00	.00	.00	.00	.00	.00	.00	**.88**	.00	.00	.00	.00	.00	.01	**.61**	.00	.00	.00	.00	.00	.00
Disco	.00	.00	.00	.00	**.62**	**.06**	.00	.00	.00		—		.00	.00	.00	.00	**.63**	.04	.00	.00	.03	**.48**	**.18**	.05	.00	.00	.00	.00	.00	.00
Hip hop	.00	.00	.00	.00	**.21**	**.08**	.00	.00	.00	.00	.00	.00		—		.00	**.37**	.00	.00	.00	.00	.00	**.91**	.03	.00	.00	.00	.00	.00	.00
Jazz	.00	.00	.00	**.43**	**.49**	**.08**	.00	**.88**	.00	.00	**.63**	.04	.00	**.37**	.00		—		.00	.01	.00	**.05**	.00	**.36**	.00	.01	.00	.00	**.68**	.00
Metal	.00	.00	.00	.01	**.95**	**.61**	.00	.00	.00	.00	.00	.03	.00	.00	.00	.00	.01	.00		—		**.41**	.01	**.43**	.00	.00	**.08**	.00	.00	.01
Pop	**.06**	**.41**	**.16**	**.12**	**.68**	**.13**	.00	.01	**.61**	**.48**	**.18**	.05	.00	**.91**	.03	**.05**	.00	**.36**	**.41**	.01	**.43**		—		.01	**.07**	**.11**	.02	**.08**	**.28**
Reggae	.00	.00	.00	.00	**.55**	**.08**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	**.08**	.01	**.07**	**.11**		—		.00	.00	.00
Rock	.00	.00	.00	.00	**.22**	**.23**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.68**	.00	.00	.00	.01	.02	**.08**	**.28**	.00	.00	.00		—	

Balanced Information Gain

	Blues			Classical			Country			Disco			Hip hop			Jazz			Metal			Pop			Reggae			Rock		
	RP	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**
Blues		—		.00	**.36**	**.49**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.10**	.00	.00	.00	.00	**.23**	**.21**	.05	.00	.00	.00	.00	.00	**.05**
Classical	.00	**.36**	**.49**		—		.00	**.96**	**.15**	.00	**.30**	.04	.00	**.88**	**.96**	**.29**	**.61**	**.11**	.00	**.14**	**.87**	.01	**.55**	**1.00**	.00	**.62**	**.25**	.00	**.33**	**.05**
Country	.00	.00	.00	.00	**.96**	**.15**		—		.00	.00	.00	.00	.00	.00	.00	**.62**	.00	.00	.00	.00	.00	**.21**	**.29**	.00	.00	.00	.00	.00	.01
Disco	.00	.00	.00	.00	**.30**	.04	.00	.00	.00		—		.00	.00	.00	.00	**.34**	.01	.00	.00	**.19**	**.17**	**.19**	**.36**	.00	.00	.00	.00	.00	.00
Hip hop	.00	.00	.00	.00	**.88**	**.96**	.00	.00	.00	.00	.00	.00		—		.00	**.58**	.00	.00	.00	.00	.00	.05	**.62**	.00	.00	.00	.00	.00	.02
Jazz	.00	**.10**	.00	**.29**	**.61**	**.11**	.00	**.62**	.00	.00	**.34**	.01	.00	**.58**	.00		—		.00	.01	.00	.03	.02	**.78**	.00	**.29**	.00	.00	**.25**	**.90**
Metal	.00	.00	.00	.00	**.14**	**.87**	.00	.00	.00	.00	.00	**.19**	.00	.00	.00	.00	.01	.00		—		**.26**	.04	**.43**	.00	.00	**.08**	.00	.00	.04
Pop	**.23**	**.21**	.05	.01	**.55**	**1.00**	.00	**.21**	**.29**	**.17**	**.19**	**.36**	.00	.05	**.62**	.03	.02	**.78**	**.26**	.04	**.43**		—		**.29**	**.07**	.01	**.08**	**.10**	**.08**
Reggae	.00	.00	.00	.00	**.62**	**.25**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	**.29**	.00	.00	.00	**.08**	**.29**	**.07**	.01		—		.00	.00	.01
Rock	.00	.00	**.05**	.00	**.33**	**.05**	.00	.00	.01	.00	.00	.00	.00	.00	.02	.00	**.25**	**.90**	.00	.00	.04	**.08**	**.10**	**.08**	.00	.00	.01		—	

ReliefF

	Blues			Classical			Country			Disco			Hip hop			Jazz			Metal			Pop			Reggae			Rock		
	RP	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**
Blues		—		.01	**.53**	**.09**	.00	.05	.04	.02	**.60**	**.07**	.01	.00	**.10**	.00	**.92**	**.14**	.00	**.41**	**.05**	.01	**.07**	.03	**.15**	**.26**	**.28**	.00	.02	.00
Classical	.01	**.53**	**.09**		—		.00	**.45**	.01	.02	**.83**	.03	.01	**.90**	**.75**	.00	**.72**	**.43**	.00	**.59**	**.19**	.00	.02	**.13**	.04	**.24**	**.45**	**.10**	.02	**.08**
Country	.00	.05	.04	.00	**.45**	.01		—		**.68**	**.38**	**.65**	**.14**	.02	**.81**	.00	.00	.05	.00	.01	.00	.00	**.14**	**.89**	**.31**	**.13**	**.29**	.00	.00	**.15**
Disco	.02	**.60**	**.07**	.02	**.83**	.03	**.68**	**.38**	**.65**		—		.00	**.49**	**.69**	**.53**	.02	**.74**	**.09**	.04	**.92**	.00	**.14**	**.60**	**.35**	**.44**	**.91**	.01	.01	**.57**
Hip hop	.01	.00	**.10**	.01	**.90**	**.75**	**.14**	.02	**.81**	.00	**.49**	**.69**		—		**.84**	**.78**	**.43**	.00	**.63**	**.91**	.00	.03	**.97**	.02	**.16**	**.46**	**.07**	**.10**	**.67**
Jazz	.00	**.92**	**.14**	.00	**.72**	**.43**	.00	.00	.05	**.53**	.02	**.74**	**.84**	**.78**	**.43**		—		.00	.05	**.48**	.00	**1.00**	.02	**.21**	**.36**	**.13**	.00	.04	**.54**
Metal	.00	**.41**	**.05**	.00	**.59**	**.19**	.00	.01	.00	**.09**	.04	**.92**	.00	**.63**	**.91**	.00	.05	**.48**		—		.02	**.64**	**.07**	**.49**	.00	**.63**	.00	.00	**.15**
Pop	.01	**.07**	.03	.00	.02	**.13**	.00	**.14**	**.89**	.00	**.14**	**.60**	.00	.03	**.97**	.00	**1.00**	.02	.02	**.64**	**.07**		—		.03	**.09**	**.61**	.00	**.41**	**.17**
Reggae	**.15**	**.26**	**.28**	.04	**.24**	**.45**	**.31**	**.13**	**.29**	**.35**	**.44**	**.91**	.02	**.16**	**.46**	**.21**	**.36**	**.13**	**.49**	.00	**.63**	.03	**.09**	**.61**		—		.00	**.29**	**.29**
Rock	.00	.02	.00	**.10**	.02	**.08**	.00	.00	**.15**	.01	.01	**.57**	**.07**	**.10**	**.67**	.00	.04	**.54**	.00	.00	**.15**	.00	**.41**	**.17**	.00	**.29**	**.29**		—	

Table 3.4: Rank correlation tests with Kendall's τ and a significance of $\alpha = 0.05$ on the GTZAN collection. The features ranks were computed according to one-vs.-rest genre situations where the features are related to the three descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histogram respectively. The three calculation models Gain Ratio, Balanced Information Gain and ReliefF were used to compute the ranking. The feature ranking of every genre situation were tested on the hypothesis $\mathcal{H}_0$ of a zero correlation against the alternative $\mathcal{H}_1$ that there is a non-zero correlation. Thus, a p-value greater than α suggests a zero ranking correlation and therefore a significant ranking variation.

	Classical			Electronic			Jazz & Blues			Metal & Punk			World			Rock & Pop		
	RP	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**	**RP**	**SSD**	**RH**
Gain Ratio																		
Classical		—		.00	**.59**	.00	.00	**.85**	**.32**	.00	**.57**	.00	.00	.01	.01	.05	**.76**	.04
Electronic	.00	**.59**	.00		—		**.63**	.04	**.66**	.00	.02	.00	**.77**	**.62**	**.06**	.00	.00	**.07**
Jazz & Blues	.00	**.85**	**.32**	**.63**	.04	**.66**		—		.00	**1.00**	**.24**	.00	**.59**	**.23**	.00	.00	**.99**
Metal & Punk	.00	**.57**	.00	.00	.02	.00	.00	**1.00**	**.24**		—		.04	**.18**	**.06**	**.54**	**.19**	**.46**
Rock & Pop	.00	.01	.01	**.77**	**.62**	**.06**	.00	**.59**	**.23**	.04	**.18**	**.06**		—		**.26**	**.51**	**.13**
World	.05	**.76**	.04	.00	.00	**.07**	.00	.00	**.99**	**.54**	**.19**	**.46**	**.26**	**.51**	**.13**		—	
Balanced Information Gain																		
Classical		—		.00	**.21**	.00	.00	**.36**	**.74**	.00	**.66**	.00	.02	**.09**	.00	**.63**	**.46**	**.10**
Electronic	.00	**.21**	.00		—		**.55**	**.24**	**.17**	.00	.00	.00	**.71**	**.82**	**.07**	.00	**.29**	**.65**
Jazz & Blues	.00	**.36**	**.74**	**.55**	**.24**	**.17**		—		.00	.02	**.63**	.00	**.10**	**.08**	.00	.00	**.07**
Metal & Punk	.00	**.66**	.00	.00	.00	.00	.00	.02	**.63**		—		.02	.04	.00	.00	**.70**	**.31**
Rock & Pop	.02	**.09**	.00	**.71**	**.82**	**.07**	.00	**.10**	**.08**	.02	.04	.00		—		**.10**	.03	.01
World	**.63**	**.46**	**.10**	.00	**.29**	**.65**	.00	.00	**.07**	.00	**.70**	**.31**	**.10**	.03	.01		—	
ReliefF																		
Classical		—		**.38**	.00	**.88**	**.12**	**.26**	**.57**	**.36**	**.10**	**.23**	**.74**	**.56**	**.22**	**.75**	**.10**	**.36**
Electronic	**.38**	.00	**.88**		—		.00	**.19**	**.22**	**.56**	**.96**	**.65**	.04	**.31**	**.17**	**.12**	**.95**	**.79**
Jazz & Blues	**.12**	**.26**	**.57**	.00	**.19**	**.22**		—		.04	.01	**.62**	.00	.03	.04	.00	**.51**	.01
Metal & Punk	**.36**	**.10**	**.23**	**.56**	**.96**	**.65**	.04	.01	**.62**		—		**.29**	**.69**	**.95**	**.39**	**.41**	**.49**
Rock & Pop	**.74**	**.56**	**.22**	.04	**.31**	**.17**	.00	.03	.04	**.29**	**.69**	**.95**		—		.00	**.31**	.00
World	**.75**	**.10**	**.36**	**.12**	**.95**	**.79**	.00	**.51**	.01	**.39**	**.41**	**.49**	.00	**.31**	.00		—	

Table 3.5: Rank correlation tests with Kendall's τ and a significance of $\alpha = 0.05$ on the ISMIR 2004 Genre collection. The features ranks were computed according to one-vs.-rest genre situations where the features are related to the three descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histogram respectively. The three calculation models Gain Ratio, Balanced Information Gain and ReliefF were used to compute the ranking. The feature ranking of every genre situation were tested on the hypothesis $\mathcal{H}_0$ of a zero correlation against the alternative $\mathcal{H}_1$ that there is a non-zero correlation. Thus, a p-value greater than α suggests a zero ranking correlation and therefore a significant ranking variation.

	ChaChaCha			Jive			Quickstep			Rumba			Samba			Slow Waltz			Tango			Viennese Waltz		
	RP	SSD	RH	RP	SSD	RH	RP	SSD	RH	RP	SSD	RH	RP	SSD	RH	RP	SSD	RH	RP	SSD	RH	RP	SSD	RH
Gain Ratio																								
ChaChaCha		—		.00	.00	.00	**.07**	.00	.00	.00	.00	.00	.00	.00	.04	.00	.00	**.69**	.00	.00	**.13**	.00	.00	**.13**
Jive	.00	.00	.00		—		.00	.00	**.08**	.00	.00	.00	.00	.00	**.09**	.00	.00	**.25**	.00	.00	**.77**	.00	.00	**.70**
Quickstep	**.07**	.00	.00	.00	.00	**.08**		—		.00	.00	.01	.00	.00	.00	.00	**.12**	.00	.00	.00	.00	.00	.00	.00
Rumba	.00	.00	.00	.00	.00	.00	.00	.00	.01		—		.00	.00	.01	.00	.00	**.36**	.00	.00	.01	.00	.00	**.12**
Samba	.00	.00	.04	.00	.00	**.09**	.00	.00	.00	.00	.00	.01		—		.00	.04	**.41**	.00	.00	**.06**	.00	.00	.04
Slow Waltz	.00	.00	**.69**	.00	.00	**.25**	.00	**.12**	.00	.00	.00	**.36**	.00	.04	**.41**		—		.00	**.07**	.00	.00	.02	.00
Tango	.00	.00	**.13**	.00	.00	**.77**	.00	.00	.00	.00	.00	.01	.00	.00	**.06**	.00	**.07**	.00		—		.00	.00	.00
Viennese Waltz	.00	.00	**.13**	.00	.00	**.70**	.00	.00	.00	.00	.00	**.12**	.00	.00	.04	.00	.02	.00	.00	.00	.00		—	
Balanced Information Gain																								
ChaChaCha		—		.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00	.02	.00	.00	**.07**	.00	.00	**.23**	.00	.00	**.58**
Jive	.00	.00	.01		—		.00	.00	**.05**	.00	.00	.01	.00	.00	.00	.00	.02	**.13**	.00	.00	**.95**	.00	.00	**.76**
Quickstep	.00	.00	.00	.00	.00	**.05**		—		.00	.00	.01	.00	.00	.00	.00	.04	.00	.00	.00	.00	.00	.00	.00
Rumba	.00	.00	.00	.00	.00	.01	.00	.00	.01		—		.00	.00	.01	.00	.00	**.33**	.00	.00	**.09**	.00	.00	**.15**
Samba	.00	.00	.02	.00	.00	.00	.00	.00	.00	.00	.00	.01		—		.00	.00	**.07**	.00	.00	.04	.00	.00	.02
Slow Waltz	.00	.00	**.07**	.00	.02	**.13**	.00	.04	.00	.00	.00	**.33**	.00	.00	**.07**		—		.00	.00	.00	.00	.04	.02
Tango	.00	.00	**.23**	.00	.00	**.95**	.00	.00	.00	.00	.00	**.09**	.00	.00	.04	.00	.00	.00		—		.00	.00	.00
Viennese Waltz	.00	.00	**.58**	.00	.00	**.76**	.00	.00	.00	.00	.00	**.15**	.00	.00	.02	.00	.04	.02	.00	.00	.00		—	
ReliefF																								
ChaChaCha		—		**.10**	**.30**	**.20**	**.44**	**.93**	**.12**	**.54**	**.79**	**.74**	**.56**	**.55**	**.85**	**.40**	**.55**	**.97**	**.26**	**.86**	**1.00**	**.44**	**.44**	**.64**
Jive	**.10**	**.30**	**.20**		—		.04	**.31**	**.26**	.00	**.27**	**.22**	**.20**	**.93**	**.23**	**.29**	**.86**	**.75**	.01	**.55**	**.30**	**.63**	**.29**	**.55**
Quickstep	**.44**	**.93**	**.12**	.04	**.31**	**.26**		—		**.31**	**.12**	**.70**	**.38**	**.26**	**.54**	**.41**	**.57**	**.10**	**.17**	**.67**	**.47**	.02	**.89**	**.76**
Rumba	**.54**	**.79**	**.74**	.00	**.27**	**.22**	**.31**	**.12**	**.70**		—		**.79**	.05	**.94**	.04	**.79**	**.45**	**.51**	**.69**	**.23**	**.31**	**.81**	**.08**
Samba	**.56**	**.55**	**.85**	**.20**	**.93**	**.23**	**.38**	**.26**	**.54**	**.79**	.05	**.94**		—		**.82**	**.16**	**.28**	**.39**	**.78**	**.45**	.03	**.38**	**.32**
Slow Waltz	**.40**	**.55**	**.97**	**.29**	**.86**	**.75**	**.41**	**.57**	**.10**	.04	**.79**	**.45**	**.82**	**.16**	**.28**		—		**.98**	**.47**	**.94**	**.33**	**.98**	**.54**
Tango	**.26**	**.86**	**1.00**	.01	**.55**	**.30**	**.17**	**.67**	**.47**	**.51**	**.69**	**.23**	**.39**	**.78**	**.45**	**.98**	**.47**	**.94**		—		**.19**	**.63**	**.19**
Viennese Waltz	**.44**	**.44**	**.64**	**.63**	**.29**	**.55**	.02	**.89**	**.76**	**.31**	**.81**	**.08**	.03	**.38**	**.32**	**.33**	**.98**	**.54**	**.19**	**.63**	**.19**		—	

Table 3.6: Rank correlation tests with Kendall's τ and a significance of $\alpha = 0.05$ on the ISMIR 2004 Rhythm collection. The features ranks were computed according to one-vs.-rest genre situations where the features are related to the three descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histogram respectively. The three calculation models Gain Ratio, Balanced Information Gain and ReliefF were used to compute the ranking. The feature ranking of every genre situation were tested on the hypothesis $\mathcal{H}_0$ of a zero correlation against the alternative $\mathcal{H}_1$ that there is a non-zero correlation. Thus, a p-value greater than α suggests a zero ranking correlation and therefore a significant ranking variation.

3.4 Conclusion

This chapter presented a discriminant analysis of audio-based rhythmic descriptors in order to distinguish musical genres. The analysis was based on five different heuristic discrimination models where each model estimates the dependency of a feature's discrimination value to a specific genre. This feature-genre dependency was basically considered as the genre discrimination of that feature. The three different descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histogram were evaluated on each of the music collections GTZAN, ISMIR 2004 Genre and Rhythm. The key goal of this chapter was to examine how the different calculation models perform on different music collections and whether a consistency of the discrimination values according to correlated genres can be concluded among the heuristic discrimination models. Another important goal of this analysis was whether specific features referring to each of the three descriptors express consistent genre discrimination or not. All computations according to the discrimination analysis were performed on one-vs.-rest genre situations, i. e. binary class situations, only.

Basically, the discrimination results revealed specific discriminative feature patterns for almost all examined musical genres. Considerably diverging feature patterns could be observed and an individual relation of feature patterns to specific genres could also be concluded in many cases. Consequently, a feature ranking based on genre discrimination might be effective for feature selection and will be discussed in chapter 4. The discrimination analysis showed clearly according to all three music collections that the heuristic discrimination models implementing the impurity function estimated very consistent feature patterns. In particular, this means that the same features were recognized to be discriminative, while the actual discrimination values slightly varied among the calculation models. It has been pointed out in section 3.2 that entropy-based calculation models tend to overestimate multi-valued features. Different approaches of normalizing the estimates exist in order to reduce the distortion of such multi-valued features. Thus, it is not surprising that the discrimination values varied. Nevertheless, it could be concluded that the different approaches of normalizing have a limited influence in the calculation of discriminative features according to all three discussed music collections.

Considering the performances of the Gain Ratio, the Balanced Information Gain and the ReliefF, the computed discriminative feature patterns according to the Statistical Spectrum Descriptor were considerable more similar among the three calculation models as in the case of the other two descriptors. This conclusion was valid for both the GTZAN and the ISMIR 2004 Genre collection but not for the ISMIR 2004 Rhythm collection. Regarding the genre-to-genre comparisons according to the partially related music collections GTZAN and ISMIR 2004 Genre, it could also be concluded in the case of the Statistical Spectrum Descriptor that the three calculation models estimated quite similar discriminative feature patterns for each of the four examined genre comparisons. In that sense the ReliefF calculation model performed better compared with the Gain Ratio and the Balanced Information Gain. According to the Rhythm Pattern descriptor and the Rhythm Histogram descriptor the discrimination results

based on the three calculation models diverged more for both the GTZAN and the ISMIR 2004 Genre collection. Only for few genres the corresponding discriminative feature patterns revealed a notable degree of similarity. For instance the discrimination results were considerably similar regarding the Classical genre of both the GTZAN and the ISMIR 2004 Genre collection. According to the ISMIR Rhythm 2004 collection the three calculation models only performed consistently with respect to specific genres. But this partially consistent performance of the three heuristic discrimination models could only be concluded for the two descriptors Rhythm Pattern and Rhythm Histogram. In the case of the Statistical Spectrum Descriptor the discriminative feature patterns according to the three calculation models diverge. Also the genre-to-genre comparisons of these two partially related collections revealed diverging discriminative feature patterns. Also in the case of both classical genres, which do correlate most, a certain degree of similarity could not be concluded for both descriptors as only in the case of the Rhythm Pattern descriptor a considerable similarity was shown. Thus, a similarity regarding the performances of three heuristic discrimination models could not be concluded for all three descriptors. The highest degree of similarity was observed in the case of the Statistical Spectrum Descriptor but only for two of three music collections.

Another very important fact was concluded in terms of the Statistical Spectrum Descriptor. The majority of features related to the statistical measures variance and skewness appeared to be irrelevant over all three music collections. In fact, for all three music collections it was shown that only few features were estimated to be discriminative and even only for a small number of genres. A large number of features corresponding to these two measures consistently exhibited zero or very low genre discrimination.

Chapter 4

Evaluation of feature selection

In the previous chapter the features of the three rhythmic descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histogram were investigated in terms of genre discrimination. Additionally, a feature ranking based on the discriminative power of every feature was introduced. This chapter examines the question of whether this feature ranking can be actually used to reduce the original feature set. In terms of music genre classification feature selection is particularly crucial, since most descriptors include a large number of features to represent a certain musical content. One key advantage of a feature selection approach is that it is a self-contained selection process which is entirely independent from the main classification process. As a consequence, available but usually very limited training sets need not to be further split into separate sets to evaluate and select certain features. This chapter empirically examines the potential of this feature selection approach by performing extensive genre classification evaluations based on different music collections, calculation models, e. g. the five heuristic discrimination models introduced in the previous chapter, and three different learning algorithms.

Section 4.1 summarizes key assumptions of feature selection and why machine learning theory strictly recommends its application. The feature selection approach which has been employed in the evaluation is described in section 4.2. The experiment environment and all obtained evaluation results are discussed in section 4.3 for each of the three rhythmic descriptors separately. At last, section 4.4 finalizes this chapter by summarizing the observed performances of the feature selection and by concluding whether this approach is actually suitable in terms of music genre classification with the described setting. Drawbacks of this selection method are also reconsidered in this section.

4.1 Overview

It has been shown for usual classification systems that feature selection is crucial to further increase the classification accuracy. Therefore feature selection is actually independent from the application context. Contrary to the intuitive assumption that the classification accuracy increases the larger the set of features is, too many features can definitely deteriorate the results in practical systems. This phenomenon is known as *curse of dimensionality* and is extensively described in [6]. Although more features may actually incorporate more information about the underlying class structures, classification systems usually do not benefit from larger feature sets unless the according number of training samples is not enlarged as well. In fact, the size of the training set must exponentially grow with the size of the feature set in order to avoid influences due to curse of dimensionality.

In MIR, often a large number of features is required to represent certain musical facet or content on song-level. This assumption applies to audio-based descriptors in particular, but as symbolic-based descriptors may include many features. Although a comprehensive and distinctive description of musical content, e. g. rhythm, chords, instrumentation, appears to require large sets of features, it is possible that a specific machine learning algorithm achieves low classification performance unless the applied feature set will not be reduced before. Thus, a problem-based feature selection approach is desirable for a proper implementation of musical classification systems. Various different approaches have been examined for effective feature selection in MIR including wrapper techniques as well as filter-bases techniques. Fiebrink et al. [20] introduced a wrapper approach based on the computation of feature weights with a genetic algorithm. In machine learning, many generic feature selection techniques are known and an evaluation of such techniques concerning music classification is presented in [23]. A very interesting work has been published by Fiebrink and Fujinaga [19] in which general pitfalls of feature selection in music classification are pointed out.

This thesis focuses on the three rhythmic descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histogram only. While the Rhythm Pattern descriptor defines 1440 features to describe the rhythmic content of a piece of music, the Statistical Spectrum Descriptor consists of 168 features and the Rhythm Histogram descriptor contains the smallest set of 60 features. As the three music collections GZTAN, ISMIR 2004 Genre and Rhythm (☞ see

section 2.6) offer a limited number of musical pieces for separate training and evaluation of genre classification systems, the classification based on those three descriptors may suffer from the problem of curse of dimensionality. In this chapter, a feature selection approach based on the genre discrimination of every feature is examined regarding genre classification. The two key goals of this evaluation are the effectiveness of the feature selection approach as well as the influence of the curse of dimensionality problem according to each of the three descriptors.

4.2 Feature Selection Approach

The previous chapter already pointed out the possibility to rank features according to their suitability for genre discrimination. An obvious assumption is to use that feature ranking for feature selection. Since chapter 3 showed that the five heuristic discrimination models yield substantially different feature ranking results, a separate feature selection evaluation is done for every calculation model. Alternatively, the genre discrimination values can also be used to weight the features instead, but this approach is not discussed within this thesis.

The feature selection approach based on the corresponding discrimination values is actually quite simple and can be described as follows:

1. Based on the computed discrimination values of every feature, a decreasing rank order is established where the first rank is related to the most discriminative feature.

2. All features are discarded which have been assigned with a zero discrimination value, since no discrimination of any genre can be measured for those features. Depending on the employed heuristic discrimination model, the removal of such "irrelevant" feature already yields a substantial feature reduction.

3. In order to determine the best feature subset, a successive evaluation is applied with a certain learning algorithm. The evaluation is repeated n times where in every evaluation step i, with $1 \leq i \leq n$, the learning algorithm is trained with the feature subset candidate C_i. A feature subset candidate C_i includes all those features having a corresponding rank of $1, 2, \ldots \lceil i \cdot d \rceil$ where $d = \frac{|\mathcal{A}|}{n}$ denotes the difference in the number of features according to two successive feature selection candidates C_{i-1} and C_i. Consequently, every feature subset candidate C_i is related to C_{i-1} by $C_{i-1} \subset C_i$ and therefore a linearly growing number of feature subsets must be evaluated only, while the number of feature subsets grows exponentially in the case of the "usual" feature subset evaluation [31] which also considers the combination of features. Common feature selection approaches introduced in [29, 31] incorporate considerably more feature subsets. The feature subset candidate C_i is considered as "optimal" regarding to the maximum accuracy according to the underlying learning algorithm.

While many feature selection approaches often require separate training sets to evaluate the original feature set, a feature ranking based on genre discrimination can be directly computed on

the very same training set which is also used for the successive training of the learning algorithm. Because of the statistical independence of the feature evaluation algorithm and the successive learning algorithm, a partitioning of the original training set is not required which certainly intensifies the problem of generally small training sets otherwise. Another advantage of using discrimination values for feature selection is that the computation time is significantly lower on average than usual wrapper-based selection techniques. The reason of the lower calculation time is the limited number of feature subsets, since for all subset candidates C_i holds true that C_{i-1} by $C_{i-1} \subset C_i$ which is not the case in usual subset selection.

4.3 Experiments

This section provides benchmark tests to determine how effective feature selection based on genre discrimination ranking really is in terms of musical genre classification. The tests have been applied in such manner that separate results are provided for each of two music collections GTZAN and ISMIR 2004 Genre on the one hand, and for each of the five heuristic discrimination models Chi-square, Information Gain, Gain Ratio, Balanced Information Gain as well as ReliefF on the other hand. In order to compare the effectiveness and the generalization of this feature selection approach, all feature selection candidates were evaluated by three learning algorithms separately. These learning algorithms are a probabilistic Naive Bayes learner, a rule-based Decision tree learner and the frequently used Support Vector Machine which estimates a discrimination function for classification. The choice of these three learning algorithm is deliberate because they represent thee quite different concepts of learning and they are frequently used in various machine learning applications.

Descriptor	$\lvert\mathcal{A}\rvert$	**n**	**d**
Rhythm Patterns	1440	30	48
Statistical Spectrum	168	30	≈ 5
Rhythm Histogram	60	30	2

Table 4.1: Evaluation settings depending on the respective rhythmic descriptor. The original feature dimension is indicated with $|\mathcal{A}|$, n is the number of employed feature selection candidates C_i with $1 \leq i \leq n$ and $d = \frac{|\mathcal{A}|}{n}$ represents the difference in the number of features for the successive feature selection candidates C_i and C_{i+1}.

The evaluation was done by applying each of the three learning algorithms with 10-fold cross validation. For every feature selection candidate, the eventual classification accuracy was obtained by averaging the partial accuracy of all 10 independent folds. To limit the number of feature selection candidates and therefore to reduce the overall calculation time, the accuracy of 30 candidates were computed to verify the feature selection performance. The size of a specific feature selection candidate C_i with $1 \leq i \leq n$ is defined by $|C_i| = \lceil i \cdot d \rceil$ with $d = \frac{|\mathcal{A}|}{n}$. Thus, C_i includes all features with a discrimination rank $1, 2, \ldots \lceil i \cdot d \rceil$. It is important to note that the assembly of the feature selection candidates, i. e. the computation of the feature ranking

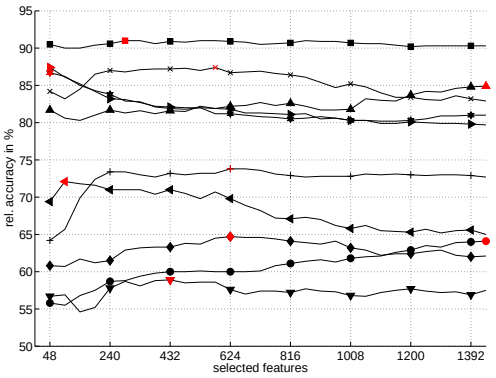
according to the estimated discrimination values of every feature, was done independently from the actual feature selection evaluation. The underlying method to compute the discrimination values of every feature is described in chapter 3. Table 4.1 gives an overview of all parameters regarding the feature selection evaluation.

The entire evaluation was performed with the Java-based machine learning workbench WEKA as it already provides a complete implementation for each of the three learning algorithms as well as the cross validation procedure. Additionally, the *Experimenter* tool of WEKA was used. The Experimenter tool provides convenient access to all required learning models and allows defining large scale experiments including cross validation to be run automatically. The WEKA class `NaiveBayes` was used for the Naive Bayes learner. To represent the Decision tree learner and the Support Vector Machine, the WEKA classes `J48` and `SMO` were chosen, respectively. The Naive Bayes and Decision tree learning algorithms were applied with standard options defined by WEKA. In particular this means that pruning was set with confidence threshold 0.25 for the Decision tree. Two different settings were used for the Support Vector Machine. Regarding the Rhythm Pattern descriptor a kernel function with the exponent $E = 1.0$, i. e. linear kernel function, was selected. For the Statistical Spectrum Descriptor and the Rhythm Histogram descriptor a polynomial kernel function with the exponent $E = 2.0$, i. e. quadratic kernel function, was employed as these descriptors contain a considerably smaller number of features.

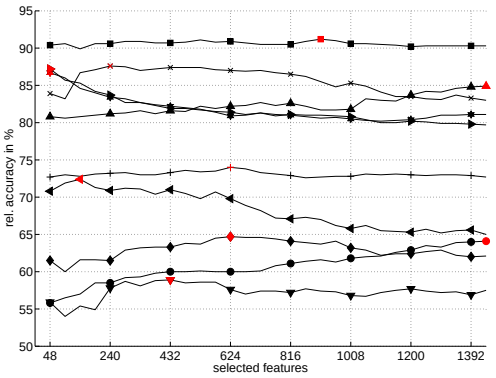
4.3.1 Rhythm Pattern

The figures 4.1, 4.2 and 4.3 illustrate the achieved classification accuracy for every selection candidate according to the Rhythm Pattern descriptor and the three different learning algorithms on the GTZAN music collection. Each figure represents the accuracy results based on different heuristic discrimination models where the Chi-square and the Information Gain were used to compute the results of figure 4.1, the Gain Ratio and the Balanced Information Gain were employed to compute the results of figure 4.2 and the ReliefF was applied to compute the results of figure 4.3. Every feature selection evaluation was performed on a specific one-vs.-rest genre situation by computing the accuracy of 30 specifically selected feature selection candidates. The actual approach of selecting the 30 feature selection candidates is described in the beginning of this section and in the table 4.1 in particular.

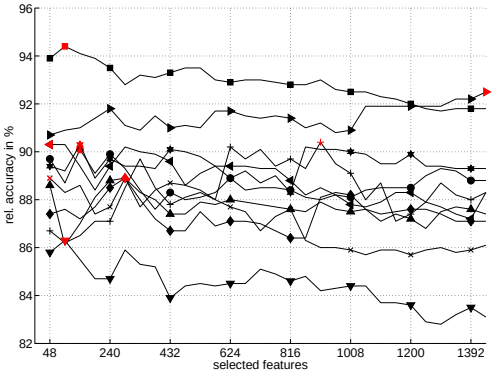
The results of the feature selection evaluation based on the heuristic discrimination models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain are very similar for each of the three learning models, while the results based on the ReliefF actually diverge. This observation is not surprising as the discriminant analysis of chapter 3 concludes that the calculation models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain yield similar feature rankings. In fact, only marginal differences of the classification accuracy according to all selection candidates related to the same genre are recognizable. Therefore the further discussion of the feature selection evaluation will be limited to the Balanced Information Gain and the ReliefF.



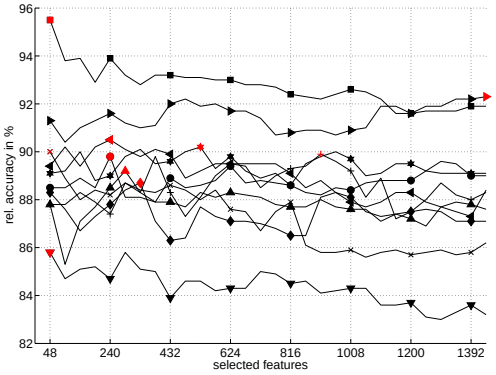
(a) Naive Bayes and Chi-square



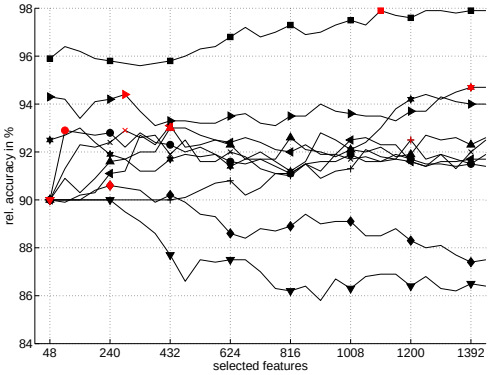
(b) Naive Bayes and Information Gain



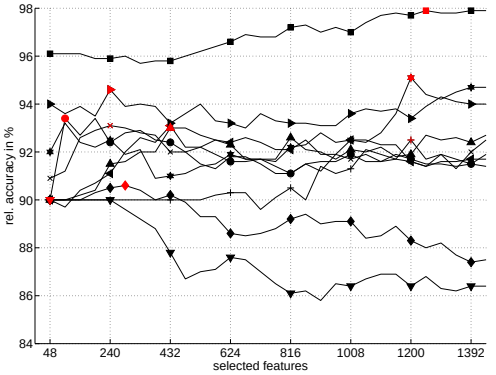
(c) J48 and Chi-square



(d) J48 and Information Gain

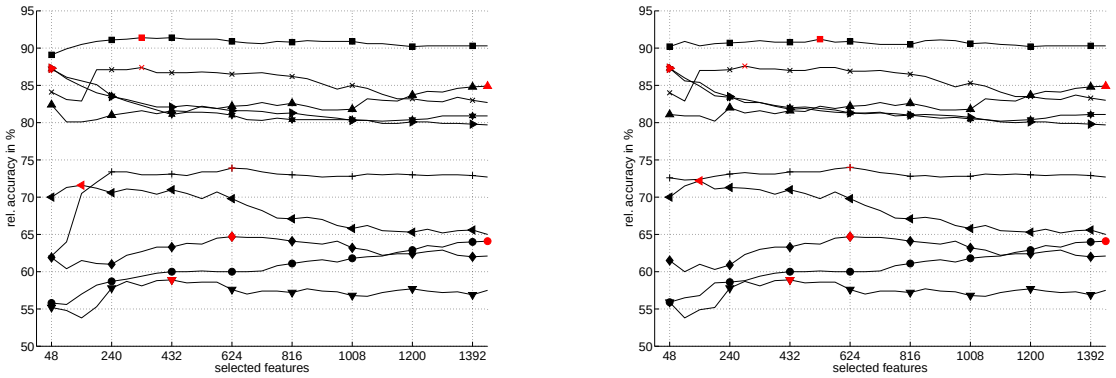


(e) SMO and Chi-square



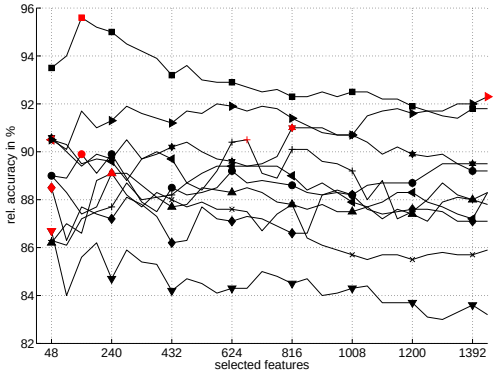
(f) SMO and Information Gain

Figure 4.1: Classification accuracy results according to the Rhythm Pattern descriptor on the GTZAN collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $i \cdot 48$ most discriminative features respectively. Three different learning algorithms were employed with the calculation models Chi-square and Information Gain each. The following symbols constitute every individual genre: ● for Blues, ■ for Classical, ◆ for Country, × for Disco, ✱ for Hip hop, + for Jazz, ◀ for Metal, ▶ for Pop, ▲ for Reggae and ▼ for Rock. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre.

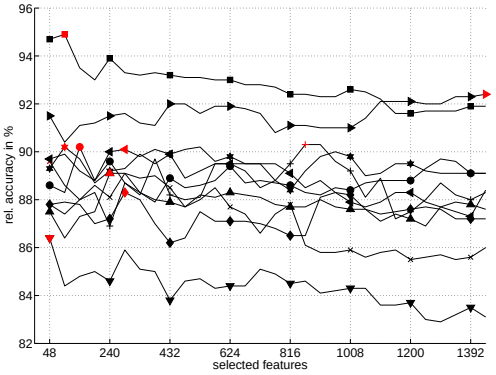


(a) Naive Bayes and Gain Ratio

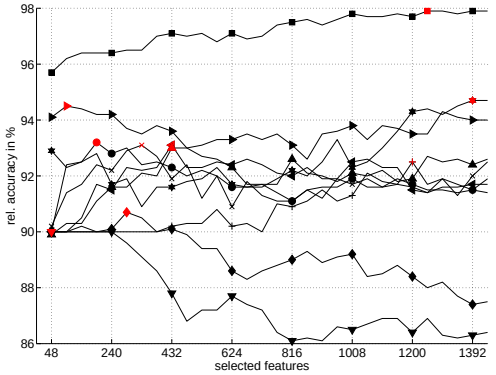
(b) Naive Bayes and Balanced IG



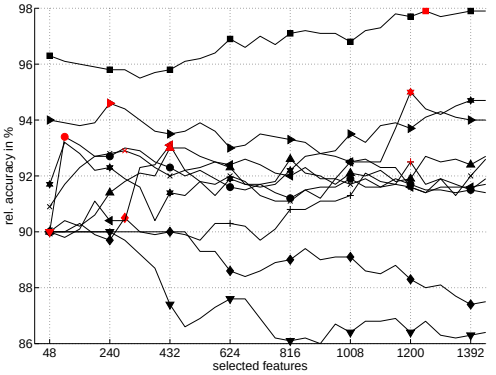
(c) J48 and Gain Ratio



(d) J48 and Balanced IG

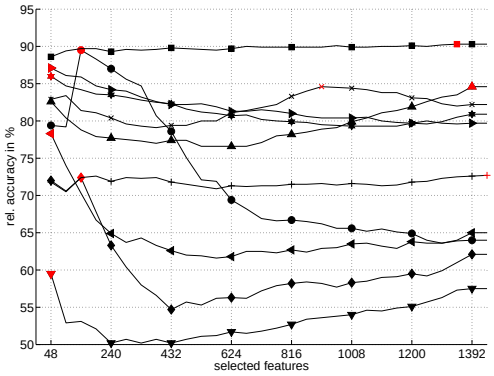


(e) SMO and Gain Ratio

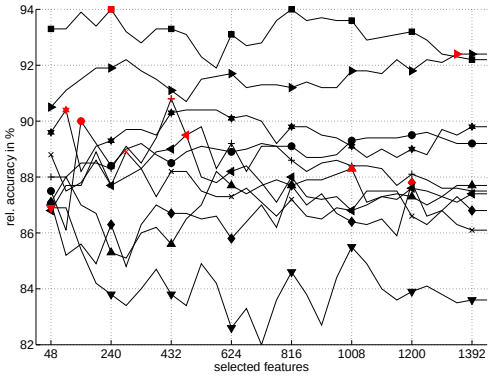


(f) SMO and Balanced IG

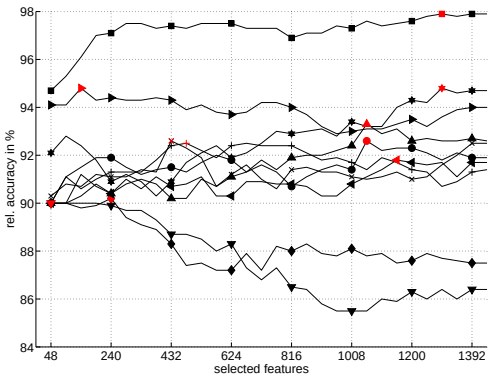
Figure 4.2: Classification accuracy results according to the Rhythm Pattern descriptor on the GTZAN collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $i \cdot 48$ most discriminative features respectively. Three different learning algorithms were employed with the calculation models Gain Ratio and Balanced Information Gain each. The following symbols constitute every individual genre: ● for Blues, ■ for Classical, ◆ for Country, × for Disco, ✸ for Hip hop, + for Jazz, ◄ for Metal, ► for Pop, ▲ for Reggae and ▼ for Rock. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre.



(a) Naive Bayes and ReliefF

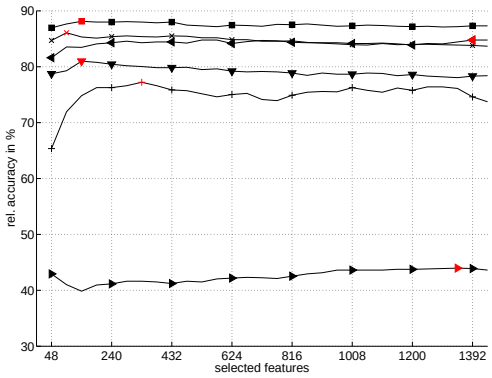


(b) J48 and ReliefF

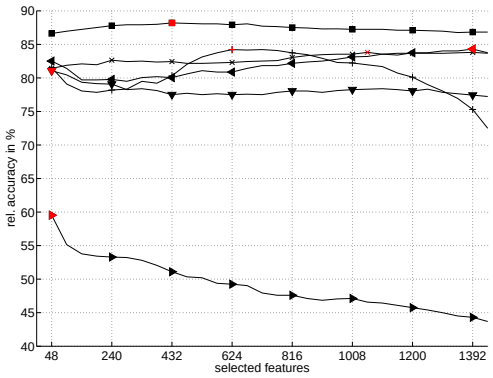


(c) SMO and ReliefF

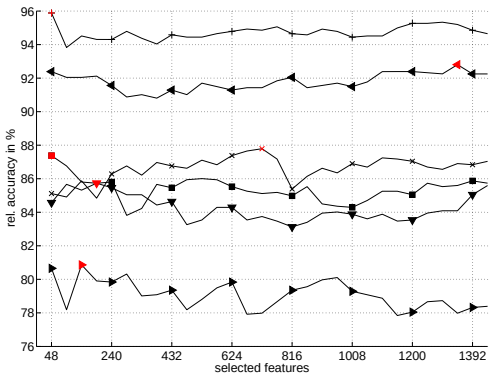
Figure 4.3: Classification accuracy results according to the Rhythm Pattern descriptor on the GTZAN collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $i \cdot 48$ most discriminative features respectively. Three different learning models were employed with the calculation model ReliefF. The following symbols constitute every individual genre: ● for Blues, ■ for Classical, ◆ for Country, × for Disco, ★ for Hip hop, + for Jazz, ◄ for Metal, ► for Pop, ▲ for Reggae and ▼ for Rock. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre.



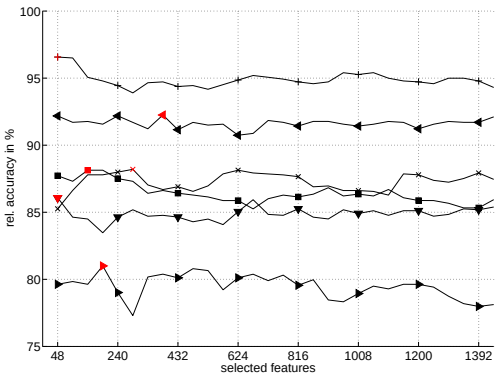
(a) Naive Bayes and Balanced IG



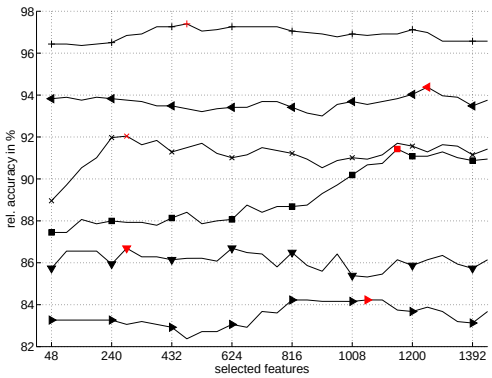
(b) Naive Bayes and ReliefF



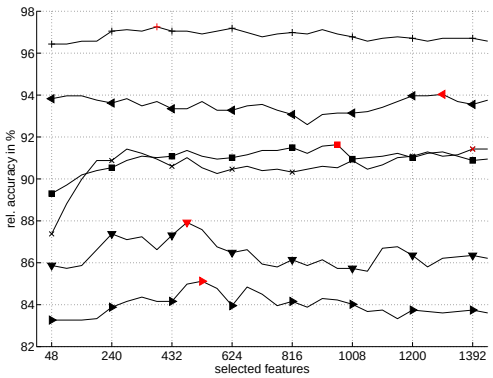
(c) J48 and Balanced IG



(d) J48 and ReliefF



(e) SMO and Balanced IG



(f) SMO and ReliefF

Figure 4.4: Classification accuracy results according to the Rhythm Pattern descriptor on the ISMIR 2004 Genre collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $i \cdot 48$ most discriminative features respectively. Three different learning algorithms were employed with the calculation models Balanced Information Gain and ReliefF each. The following symbols constitute every individual genre: ■ for Classical, × for Electronic, + for Jazz & Blues, ◀ for Metal & Punk, ▼ for Rock & Pop and ▶ for World. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre.

Considering the results of the Naive Bayes learning algorithm and the Balanced Information Gain in figure (b), the classification accuracy of the feature selection candidates related to the same genre is quite consistent for most of the 10 one-vs.-rest genre situations. In fact, for 6 of maximum 10 genres the accuracy of the corresponding feature selection candidates varies with less than 5 %. Only the genres Blues, Hip hop, Metal and Pop imply stronger variations where the accuracy according to both Hip hop and Pop significantly decreases with the use of more discriminative features, while the classification accuracy of Blues as well as Metal improves when more discriminative features were used. Another interesting observation is that the best classification accuracy of 5 genres is significantly worse in comparison to the other genres. This suggests that these genres can not be sufficiently represented by the selected features if the contributions of those features will be used independently during the classification as the Naive Bayes does.

In the case of the J48 learning algorithm and the Balanced Information Gain illustrated in figure (d) the classification accuracy varies considerably more among the feature selection candidates related to the same genre. Particularly, a highly fluctuating progression of the classification accuracy can be observed for successive selection candidates. This accuracy fluctuation among the feature selection candidates is a typical observation regarding the J48 learning algorithm, as the contributions of the feature dependencies also influence the classification performance. Nevertheless, for the majority of genres the variations concerning the classification accuracy are always within an approximate margin of 2 %. This implies a promising performance of the feature selection. It also suggests that the variation might not even be significant at all which represents a strong affirmation for the feature selection approach. Again, the worst accuracy results can be observed for the genres Country and Rock, while the best accuracy of 95 % is related to classical music.

Also in the case of the feature selection based on the Balanced Information Gain and the SMO learning algorithm illustrated in figure (f), the classification accuracy variation is limited by 2 % according to the feature selection candidates of 7 of 10 genres. This confirms the usefulness of the selection approach, as the classification accuracy is quite consistent although a considerable reduction of the original feature set is performed. Another interesting observation concerns the genres Country and Rock. The accuracy for those feature selection candidates containing a smaller number of discriminative features is considerable better in comparison to selection candidates possessing many discriminative features. In fact, the accuracy significantly deteriorates for feature selection candidates having more than 480 and 240 most discriminative features according to the genres Country and Rock, respectively. However, since the SMO learning algorithm refers to a Support Vector Machine, it should not be influenced by additional, even irrelevant, features because a Support Vector Machine utilizes an intrinsic feature selection process based on the feature space transformation. Thus, this accuracy decline is a very unusual phenomenon. As a confirmation, the classification accuracy according to the other 8 genres actually increases when more discriminative feature were selected although of this accuracy improvement might not be statistically significant.

	Chi-square					
	Naive Bayes		J48		SMO	
Genre	**Accuracy**	**# Features**	**Accuracy**	**# Features**	**Accuracy**	**# Features**
Blues	64.10 ± 3.63	1440 (100 %)	90.10 ± 3.75	144 (10 %)	92.90 ± 1.97	96 (07 %)
Classical	91.00 ± 3.23	288 (20 %)	94.40 ± 2.07	96 (07 %)	97.90 ± 1.29	1104 (77 %)
Country	64.70 ± 7.29	624 (43 %)	88.90 ± 2.18	288 (20 %)	90.60 ± 1.26	240 (17 %)
Disco	87.40 ± 3.06	576 (40 %)	88.90 ± 1.10	48 (03 %)	92.90 ± 0.88	288 (20 %)
Hip hop	86.70 ± 3.62	48 (03 %)	90.30 ± 3.30	144 (10 %)	94.70 ± 1.64	1392 (97 %)
Jazz	73.80 ± 3.68	624 (43 %)	90.40 ± 2.67	912 (63 %)	92.50 ± 2.37	1200 (83 %)
Metal	72.10 ± 4.20	96 (07 %)	90.30 ± 1.77	48 (03 %)	93.10 ± 1.66	432 (30 %)
Pop	87.40 ± 3.78	48 (03 %)	92.50 ± 2.01	1440 (100 %)	94.40 ± 1.90	288 (20 %)
Reggae	84.90 ± 4.63	1440 (100 %)	88.90 ± 2.85	288 (20 %)	93.00 ± 1.83	432 (30 %)
Rock	58.90 ± 3.70	432 (30 %)	86.30 ± 1.89	96 (07 %)	90.00 ± 0.00	48 (03 %)
Information Gain						
Blues	64.10 ± 3.63	1440 (100 %)	89.80 ± 2.74	240 (17 %)	93.40 ± 2.41	96 (07 %)
Classical	91.20 ± 1.40	912 (63 %)	95.50 ± 2.07	48 (03 %)	97.90 ± 1.10	1248 (87 %)
Country	64.70 ± 7.29	624 (43 %)	88.70 ± 2.41	336 (23 %)	90.60 ± 1.71	288 (20 %)
Disco	87.60 ± 3.44	240 (17 %)	90.00 ± 1.63	48 (03 %)	93.10 ± 0.74	240 (17 %)
Hip hop	86.70 ± 3.80	48 (03 %)	90.20 ± 3.05	528 (37 %)	95.10 ± 1.73	1200 (83 %)
Jazz	74.00 ± 3.77	624 (43 %)	89.90 ± 2.69	912 (63 %)	92.50 ± 2.37	1200 (83 %)
Metal	72.40 ± 4.30	144 (10 %)	90.50 ± 2.12	240 (17 %)	93.10 ± 1.66	432 (30 %)
Pop	87.20 ± 4.18	48 (03 %)	92.30 ± 2.16	1440 (100 %)	94.60 ± 1.35	240 (17 %)
Reggae	84.90 ± 4.63	1440 (100 %)	89.20 ± 2.78	288 (20 %)	93.00 ± 1.83	432 (30 %)
Rock	58.90 ± 3.70	432 (30 %)	85.80 ± 3.39	48 (03 %)	90.00 ± 0.00	48 (03 %)
Gain Ratio						
Blues	64.10 ± 3.63	1440 (100 %)	89.90 ± 2.33	144 (10 %)	93.20 ± 2.78	192 (13 %)
Classical	91.40 ± 2.67	336 (23 %)	95.60 ± 1.78	144 (10 %)	97.90 ± 1.10	1248 (87 %)
Country	64.70 ± 7.29	624 (43 %)	88.50 ± 2.92	48 (03 %)	90.70 ± 1.64	288 (20 %)
Disco	87.40 ± 3.47	336 (23 %)	89.10 ± 2.60	240 (17 %)	93.10 ± 0.88	336 (23 %)
Hip hop	87.20 ± 3.46	48 (03 %)	91.00 ± 1.63	816 (57 %)	94.70 ± 1.64	1392 (97 %)
Jazz	73.90 ± 3.67	624 (43 %)	90.50 ± 1.96	672 (47 %)	92.50 ± 2.37	1200 (83 %)
Metal	71.60 ± 4.58	144 (10 %)	90.50 ± 1.96	48 (03 %)	93.10 ± 1.66	432 (30 %)
Pop	87.30 ± 4.24	48 (03 %)	92.30 ± 2.11	1440 (100 %)	94.50 ± 2.01	96 (07 %)
Reggae	84.90 ± 4.63	1440 (100 %)	89.10 ± 3.41	240 (17 %)	93.00 ± 1.83	432 (30 %)
Rock	58.90 ± 3.70	432 (30 %)	86.70 ± 3.40	48 (03 %)	90.00 ± 0.00	48 (03 %)

Continued on the next page ...

In figure 4.3, the results of the feature selection evaluation according to the ReliefF calculation model are depicted which introduce slight divergences to the results according to the Balanced Information Gain or the other three calculation models based on the impurity function. Basically, similar conclusions concerning the strong limitation of the classification accuracy among the feature selection candidates related to the same genre can be made. In fact, a strong limitation of the variations regarding the accuracy of the corresponding feature selection candi-

	Balanced Information Gain					
	Naive Bayes		J48		SMO	
Genre	**Accuracy**	**# Features**	**Accuracy**	**# Features**	**Accuracy**	**# Features**
Blues	64.10 ± 3.63	1440 (100 %)	90.20 ± 2.78	144 (10 %)	93.40 ± 2.12	96 (07 %)
Classical	91.20 ± 1.55	528 (37 %)	94.90 ± 1.97	96 (07 %)	97.90 ± 1.10	1248 (87 %)
Country	64.70 ± 7.29	624 (43 %)	88.30 ± 2.45	288 (20 %)	90.50 ± 1.78	288 (20 %)
Disco	87.60 ± 3.92	288 (20 %)	89.60 ± 1.43	48 (03 %)	92.90 ± 1.10	288 (20 %)
Hip hop	87.30 ± 4.32	48 (03 %)	90.20 ± 2.53	96 (07 %)	95.00 ± 1.76	1200 (83 %)
Jazz	74.00 ± 3.77	624 (43 %)	90.30 ± 2.45	864 (60 %)	92.50 ± 2.37	1200 (83 %)
Metal	72.20 ± 4.42	144 (10 %)	90.10 ± 2.85	288 (20 %)	93.10 ± 1.66	432 (30 %)
Pop	87.30 ± 4.00	48 (03 %)	92.40 ± 2.12	1440 (100 %)	94.60 ± 1.35	240 (17 %)
Reggae	84.90 ± 4.63	1440 (100 %)	89.10 ± 2.60	240 (17 %)	93.00 ± 1.83	432 (30 %)
Rock	58.90 ± 3.70	432 (30 %)	86.40 ± 2.59	48 (03 %)	90.00 ± 0.00	48 (03 %)
			ReliefF			
Blues	89.50 ± 3.89	144 (10 %)	90.00 ± 1.49	144 (10 %)	92.60 ± 2.41	1056 (73 %)
Classical	90.30 ± 1.95	1344 (93 %)	94.00 ± 1.56	240 (17 %)	97.90 ± 1.29	1296 (90 %)
Country	72.40 ± 2.80	144 (10 %)	87.80 ± 2.62	1200 (83 %)	90.20 ± 1.93	240 (17 %)
Disco	84.60 ± 3.72	912 (63 %)	88.90 ± 3.14	288 (20 %)	92.60 ± 2.07	432 (30 %)
Hip hop	86.00 ± 4.42	48 (03 %)	90.40 ± 1.78	96 (07 %)	94.80 ± 1.32	1296 (90 %)
Jazz	72.70 ± 3.92	1440 (100 %)	90.80 ± 1.99	432 (30 %)	92.50 ± 2.27	480 (33 %)
Metal	78.30 ± 2.98	48 (03 %)	89.50 ± 3.06	480 (33 %)	91.80 ± 2.35	1152 (80 %)
Pop	87.10 ± 5.28	48 (03 %)	92.40 ± 2.46	1344 (93 %)	94.80 ± 1.55	144 (10 %)
Reggae	84.60 ± 4.09	1392 (97 %)	88.30 ± 1.25	1008 (70 %)	93.30 ± 1.95	1056 (73 %)
Rock	59.50 ± 5.74	48 (03 %)	86.90 ± 4.48	48 (03 %)	90.00 ± 0.00	48 (03 %)

Table 4.2: Evaluation of the feature selection based on the genre discrimination of every feature according to the Rhythm Pattern descriptor on the GTZAN collection. The best classification accuracy with the corresponding standard deviation and the related number of selected features (relative amount of selected features) are listed for every one-vs.-rest genre situation and each of the three calculation models.

dates is recognizable for the majority of the 10 genres. An important fact is that independent from the actual learning algorithm the genres Country and Rock are always under the top three genres which have the strongest variations in terms of the feature selection performance. Even the decline of the classification accuracy according to the SMO can be seen clearly in the case of selection more discriminative features.

Generally, the classification accuracy results of each heuristic calculation model confirm that the feature ranking based on genre discrimination constitutes an effective feature selection approach. Although the employed learning algorithms represent very different learning concepts, a considerable feature set reduction could be achieved in combination with a slight or even insignificant decline of the classification accuracy. Another very interesting general observation is that even a high feature set reduction of more than 50 % had a very limited effect on the classification performance. In the case of the Naive Bayes this limitation is approximately 5 %,

Chi-square						
	Naive Bayes		J48		SMO	
Genre	**Accuracy**	**# Features**	**Accuracy**	**# Features**	**Accuracy**	**# Features**
Classical	88.00 ± 1.62	432 (30 %)	86.97 ± 2.32	48 (03 %)	91.56 ± 1.78	1152 (80 %)
Electronic	87.65 ± 2.15	48 (03 %)	88.20 ± 2.13	432 (30 %)	92.11 ± 1.43	336 (23 %)
Jazz & Blues	85.81 ± 3.61	48 (03 %)	95.40 ± 1.13	192 (13 %)	97.33 ± 1.27	432 (30 %)
Metal & Punk	85.46 ± 4.24	48 (03 %)	92.94 ± 1.82	1344 (93 %)	94.31 ± 1.80	1104 (77 %)
Rock & Pop	81.83 ± 3.27	48 (03 %)	85.67 ± 2.48	144 (10 %)	86.83 ± 2.50	528 (37 %)
World	43.97 ± 3.48	1344 (93 %)	81.41 ± 2.50	48 (03 %)	84.57 ± 1.21	768 (53 %)
Balanced Information Gain						
Classical	88.13 ± 2.18	144 (10 %)	87.38 ± 3.19	48 (03 %)	91.43 ± 2.00	1152 (80 %)
Electronic	86.08 ± 2.65	96 (07 %)	87.79 ± 1.79	720 (50 %)	92.04 ± 1.88	288 (20 %)
Jazz & Blues	77.23 ± 6.06	336 (23 %)	95.89 ± 1.55	48 (03 %)	97.40 ± 1.32	480 (33 %)
Metal & Punk	84.78 ± 3.09	1392 (97 %)	92.80 ± 2.14	1344 (93 %)	94.38 ± 1.57	1248 (87 %)
Rock & Pop	81.00 ± 3.66	144 (10 %)	85.73 ± 1.87	192 (13 %)	86.69 ± 1.90	288 (20 %)
World	43.97 ± 3.48	1344 (93 %)	80.86 ± 2.59	144 (10 %)	84.23 ± 2.42	1056 (73 %)
ReliefF						
Classical	88.20 ± 2.37	432 (30 %)	88.14 ± 2.47	144 (10 %)	91.63 ± 1.61	960 (67 %)
Electronic	83.81 ± 2.97	1056 (73 %)	88.20 ± 3.28	288 (20 %)	91.43 ± 1.56	1392 (97 %)
Jazz & Blues	84.22 ± 2.15	624 (43 %)	96.57 ± 0.73	48 (03 %)	97.26 ± 1.07	384 (27 %)
Metal & Punk	84.30 ± 3.40	1392 (97 %)	92.25 ± 1.11	384 (27 %)	94.03 ± 1.77	1296 (90 %)
Rock & Pop	81.07 ± 3.00	48 (03 %)	86.08 ± 2.23	48 (03 %)	87.93 ± 3.09	480 (33 %)
World	59.53 ± 4.23	48 (03 %)	81.00 ± 3.36	192 (13 %)	85.12 ± 1.86	528 (37 %)

Table 4.3: Evaluation of the feature selection based on the genre discrimination of every feature according to the Rhythm Pattern descriptor on the ISMIR 2004 Genre collection. The best classification accuracy with the corresponding standard deviation and the related number of selected features (relative amount of selected features) are listed for every one-vs.-rest genre situation and each of the three calculation models.

for the learning algorithms J48 and SMO this limitation is even smaller with approximately 2 %.

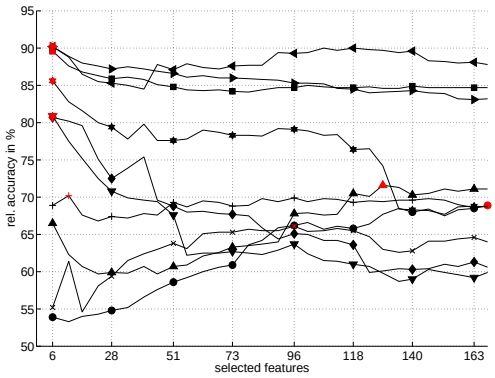
Table 4.2 lists the best classification accuracy with the corresponding standard derivation and the related number of selected features for every one-vs.-rest genre situation. As already concluded, the feature selection performance is very similar in the case of the four calculation models implementing the impurity function. Both the best accuracy and the related number of selected features is very similar among these calculation models. An interesting fact is that independent from the actual learning model and calculation model the genres Classical and Pop are always under the top three genres having the highest accuracy of all genres. As the discriminant analysis of chapter 3 already showed, the highest discrimination values were estimated for these two genres. Thus, these two genres appear to be well distinguishable from the other genres. Figure 4.3 already illustrates a slight divergence of the feature selection performance based on the ReliefF in comparison to the other four calculation models. The corresponding results of table 4.2 also suggest slightly different results. In particular in terms of the Bayes better accuracy results were achieved in combination with a similar amount of feature set reduction. As pointed

out in [48], the ReliefF model also incorporates the contribution of feature dependencies into the estimation of the discriminative power of a specific feature. Since the Naive Bayes assumes all features to be independent, the feature selection based on the ReliefF model appears to compensate the negative effect of this strict assumption the Naive Bayes based on. For the Naive Bayes a high feature set reduction was achieved where only in the case of the genres Blues and Reggae the use of the entire feature set yielded the best accuracy. Naturally, the SMO learning algorithm achieved better accuracy results when more discriminative feature had been employed. The average relative amount of selected features[1] according to the Balanced Information Gain is 38.9 % for the Naive Bayes, 24.7 % for the J48 and remarkable 38 % for the SMO learning algorithm. According to the calculation model ReliefF, the relative amount of selected features is 38.5 % for the Naive Bayes, 36.6 % for the J48 and 49.9 % for the SMO learning algorithm. Thus, a potential feature set reduction can be concluded according to each of the three learning models. The feature selection maintains an acceptable classification performance with respect to using the entire feature set and offers a clear saving of calculation time.

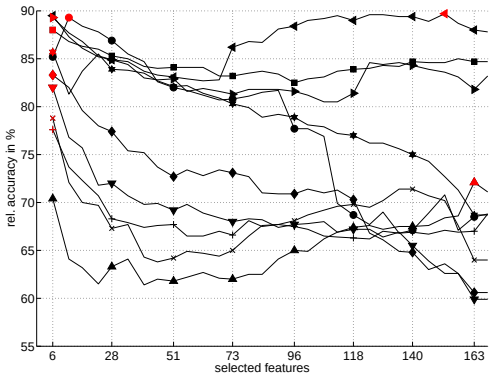
ISMIR 2004 Genre

Also the feature selection evaluation on the ISMIR 2004 Genre collection does basically affirm the conclusions regarding the evaluation on the GTZAN collection. Table 4.3 shows the results of the feature selection evaluation based on this music collection. Again, both the classification accuracy and related the number of selected features are quite similar for the calculation models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain. The average relative amount of selected features according to the Balanced Information Gain is 40 % for the Naive Bayes, 28.7 % for the J48 and 52.2 % for the SMO learning algorithm. According to the ReliefF calculation model, the relative amount of selected features is 41.5 % for the Naive Bayes, remarkable 12.7 % for the J48 and 58.5 % for the SMO learning algorithm. Like in the case of the GTZAN collection a potential feature set reduction can also be concluded for all three learning models.

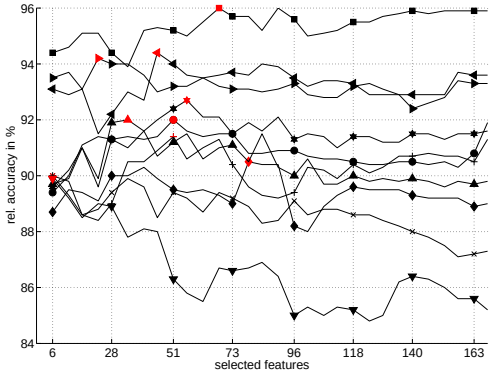
Figure 4.4 depicts the classification accuracy of the feature selection candidates related to the same genre according to the ISMIR 2004 Genre collection. Since the feature selection results based on the calculation models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain are quite similar, only the two calculation models Balanced Information Gain and ReliefF were used. Basically, the usefulness of the feature selection is also approved by these evaluation results, as the classification accuracy of the feature selection candidates related to the same genre does only vary within a narrow margin for the majority of the 6 genres. This margin can be approximately defined with 5 % for the Naive Bayes and J48 and with 2 % for the SMO.



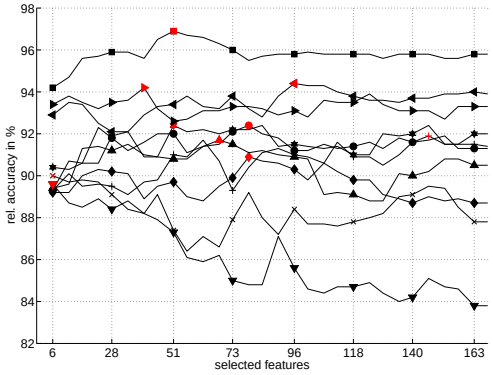
(a) Naive Bayes and Balanced IG



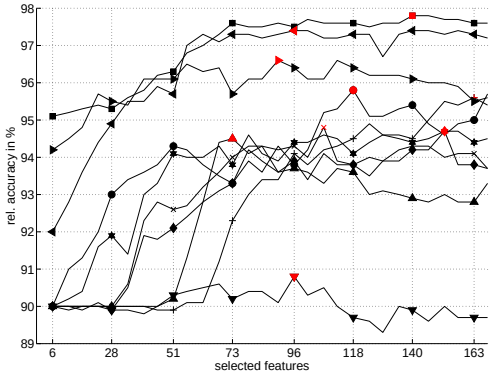
(b) Naive Bayes and ReliefF



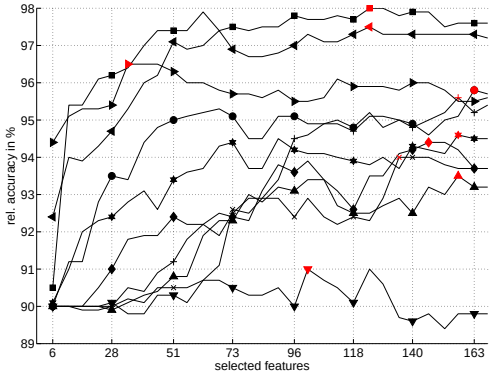
(c) J48 and Balanced IG



(d) J48 and ReliefF

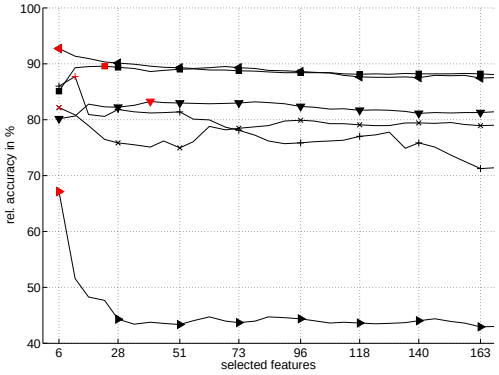


(e) SMO and Balanced IG

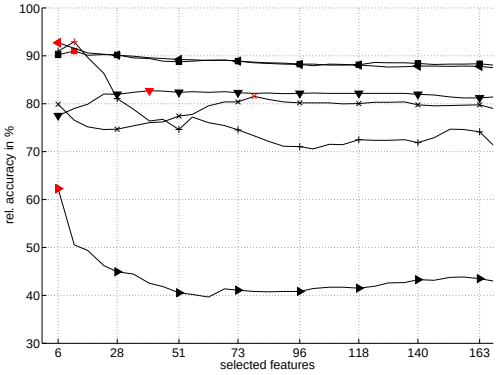


(f) SMO and ReliefF

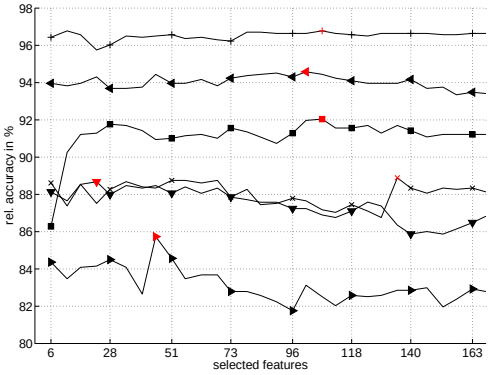
Figure 4.5: Classification accuracy results according to the Statistical Spectrum Descriptor on the GTZAN collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $\lceil i \cdot \frac{168}{30} \rceil$ most discriminative features respectively. Three different learning algorithms were employed with the calculation models Balanced Information Gain and ReliefF each. The following symbols constitute every individual genre: ● for Blues, ■ for Classical, ◆ for Country, × for Disco, ✸ for Hip hop, + for Jazz, ◀ for Metal, ▶ for Pop, ▲ for Reggae and ▼ for Rock. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre.



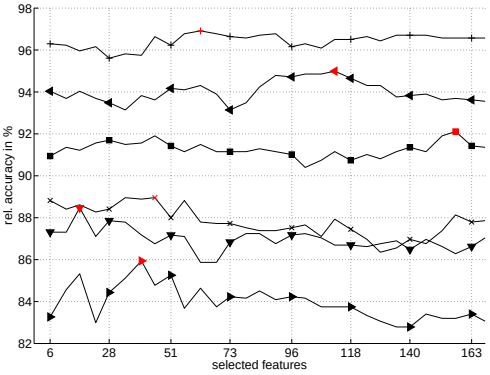
(a) Naive Bayes and Balanced IG



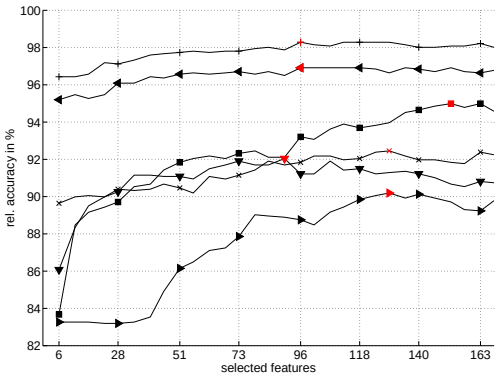
(b) Naive Bayes and ReliefF



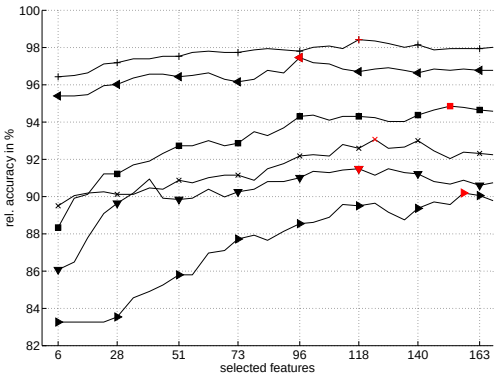
(c) J48 and Balanced IG



(d) J48 and ReliefF



(e) SMO and Balanced IG



(f) SMO and ReliefF

Figure 4.6: Classification accuracy results according to the Statistical Spectrum Descriptor on the ISMIR 2004 Genre collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $\lceil i \cdot \frac{168}{30} \rceil$ most discriminative features respectively. Three different learning algorithms were employed with the calculation models Balanced Information Gain and ReliefF each. The following symbols constitute every individual genre: ■ for Classical, × for Electronic, + for Jazz & Blues, ◀ for Metal & Punk, ▼ for Rock & Pop and ▶ for World. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre.

4.3.2 Statistical Spectrum Descriptor

Focusing first on the GTZAN collection, the classification accuracy of all feature selection candidates related to the same genre is illustrated for every one-vs.-rest genre situation in the figure 4.5. As the results of the Rhythm Pattern descriptor already showed in the previous subsection, the feature selection performances based on either of the four heuristic discrimination models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain are very similar, while the feature selection based on the ReliefF model introduces slight differences in the feature selection performance. Thus, only the Balanced Information Gain model representing the four models implementing the impurity function and the ReliefF model were used to compute the accuracy results depicted in this figure.

Beginning with the results of the Naive Bayes and the Balanced Information Gain presented in figure (a), the classification accuracy of all feature selection candidates related to the same genre varies by a margin of classification for the majority of the 10 genres. This variation of the classification accuracy is higher than the variation according to the accuracy results based on the Rhythm Pattern descriptor. Only the accuracy variations according to the genres Classical, Jazz, Metal and Pop are limited by a margin of 5 % at most. An interesting observation is that when at least 96 of the maximum 168 features were taken into account during the classification the accuracy variations of 9 of 10 genres was limited by a margin of 5 % at most. As the selection of 96 features represents a feature set reduction of approximately 40 % according to the original size of the feature set, this is actually a very promising observation. Another interesting fact is that 5 of 10 genres are related to a decline of the classification accuracy when more discriminative features were selected for the classification, while for the other genres an improvement of the classification accuracy was achieved in the case of selecting many discriminative features.

In the case of the J48 learning algorithm and the Balanced Information Gain in figure (c) the variation of the classification accuracy is surprisingly limited by a margin of 2 % for all genres but Rock. In fact, the accuracy of the feature selection candidates related to rock music considerably decreases when more discriminative features were taken into account. This implies that if very few discriminative features are used to classify rock music the classification performance will actually improve. This assumption is also confirmed in the case of the Naive Bayes learning algorithm but also in terms of all three learning algorithm and the previously examined Rhythm Pattern descriptor. Similar to the results of the Naive Bayes the classification accuracy is always within a margin of 2 % for all genres if only feature selection candidates containing at least 96 of the maximum 168 features will be considered. Thus, a feature set reduction of approximately 40 % does guarantee a quite acceptable classification performance for every genre.

The feature selection performance based on the SMO learning algorithm and the Balanced Information Gain is visualized in figure (e). Contrary to the corresponding results based on the Rhythm Pattern descriptor, an improvement of the classification accuracy can be clearly observed for all genres but rock music when more discriminative features were used during the

[1] The average relative number of selected features is aggregated over all musical genres defined by the GTZAN or by ISMIR 2004 Genre collection.

classification. Considering the variation of the accuracy among all feature selection candidates of the same genre, a stronger variation with at most 6 % is recognizable. Again, this observation is valid for all genres but rock music. Nevertheless, in the case of selecting more than 96 of the maximum 168 features the accuracy is remarkable consistent with a variation of 1 % at most according to all genres. Thus, the feature selection approach also achieves a promising feature set reduction regarding the SMO. It is important to note that the classification accuracy among the feature selection candidates of rock music does not deteriorate that much in comparison to accuracy results based on the Rhythm Pattern descriptor. In fact, the accuracy of feature selection candidates possessing many features is slightly lower but the variation among all candidates is limited by approximately 1 % which might not even be significant.

Each of the figures (b), (d) and (f) illustrates the feature selection performances based on the ReliefF model and the three learning algorithms respectively. Basically, the results of each learning algorithm do not vary that much comparing with the corresponding results based on the Balanced Information Gain. In fact, the feature selection performances according to the J48 and SMO learning algorithms are very similar and therefore the same conclusions can be made. In particular the consistent classification accuracy of feature selection candidates containing more than 96 of 168 most discriminative features can be recognized clearly for all genres. Only the accuracy results according to the Naive Bayes diverge as more genres are related to decreasing classification accuracy when more features were selected.

From a general point of view, both heuristic discrimination models promise a proper feature ranking for feature selection where at least 40 % of the original feature set can be reduced without having a considerable decline in the classification accuracy. Actually, a decline of at most 2 % must be expected which might not even be significant in some situations. Another interesting fact is that almost all evaluation results confirm that the Statistical Spectrum Descriptor outperforms the Rhythm Pattern descriptor on the GTZAN collection in terms of classification accuracy. In particular the achieved accuracy results based on the SMO learning model are equal or better in comparison to the corresponding results based on the Rhythm Pattern descriptor.

In order to compare the contributions of the two calculation models in terms of the feature selection for every calculation and learning algorithms, table 4.4 lists the classification accuracy and the corresponding standard derivation as well as the related number of selected features for every one-vs.-rest genre situation. Basically, according to all genres the assumption holds true that in order to achieve the best classification accuracy the number of selected features is equal or lower in terms of the Naive Bayes and J48 learning algorithms than in the case of the Rhythm Pattern descriptor. This is also reflected by the average relative number of selected features which have already been computed for the Rhythm Pattern descriptor. According to the Balanced Information Gain this relative amount of selected features is 26.9 % for the Naive Bayes, 24.9 % for the J48 and 70.5 % for the SMO learning algorithm. According to the calculation model ReliefF, the relative amount of selected features is 22.2 % for the Naive Bayes, 37 % for the J48 and 77.1 % for the SMO learning algorithm. It can be followed that the feature set reduction is considerably high according to the Naive Bayes and quite similar in the case of

Balanced Information Gain						
	Naive Bayes		**J48**		**SMO**	
Genre	**Accuracy**	**# Features**	**Accuracy**	**# Features**	**Accuracy**	**# Features**
Blues	68.90 ± 4.53	168 (100 %)	92.00 ± 2.49	51 (30 %)	95.80 ± 1.81	118 (70 %)
Classical	89.60 ± 2.17	6 (04 %)	96.00 ± 2.26	68 (40 %)	97.80 ± 1.32	140 (83 %)
Country	80.70 ± 4.92	6 (04 %)	90.50 ± 2.55	79 (47 %)	94.70 ± 2.00	152 (90 %)
Disco	66.10 ± 5.95	96 (57 %)	90.00 ± 0.82	6 (04 %)	94.80 ± 1.23	107 (64 %)
Hip hop	85.60 ± 2.76	6 (04 %)	92.70 ± 3.74	56 (33 %)	94.70 ± 2.50	152 (90 %)
Jazz	70.20 ± 3.65	12 (07 %)	91.40 ± 2.37	51 (30 %)	95.60 ± 2.22	163 (97 %)
Metal	90.20 ± 3.08	6 (04 %)	94.40 ± 3.50	45 (27 %)	97.40 ± 1.43	96 (57 %)
Pop	90.20 ± 3.01	6 (04 %)	94.20 ± 1.40	23 (14 %)	96.60 ± 1.71	90 (54 %)
Reggae	71.60 ± 4.86	129 (77 %)	92.00 ± 1.33	34 (20 %)	94.50 ± 1.35	73 (43 %)
Rock	80.90 ± 3.70	6 (04 %)	89.90 ± 0.57	6 (04 %)	90.80 ± 1.23	96 (57 %)
ReliefF						
Blues	89.30 ± 4.40	12 (07 %)	92.40 ± 1.51	79 (47 %)	95.80 ± 1.55	163 (97 %)
Classical	88.00 ± 2.16	6 (04 %)	96.90 ± 1.91	51 (30 %)	98.00 ± 1.15	124 (74 %)
Country	83.30 ± 3.86	6 (04 %)	90.90 ± 1.52	79 (47 %)	94.40 ± 2.46	146 (87 %)
Disco	78.80 ± 4.52	6 (04 %)	90.00 ± 0.00	6 (04 %)	94.00 ± 2.21	135 (80 %)
Hip hop	85.70 ± 3.20	6 (04 %)	92.40 ± 2.59	51 (30 %)	94.60 ± 2.95	157 (93 %)
Jazz	77.60 ± 2.80	6 (04 %)	91.90 ± 2.77	146 (87 %)	95.60 ± 2.01	157 (93 %)
Metal	89.70 ± 2.16	152 (90 %)	94.40 ± 2.27	96 (57 %)	97.50 ± 1.72	124 (74 %)
Pop	89.30 ± 2.67	6 (04 %)	94.20 ± 1.40	40 (24 %)	96.50 ± 1.27	34 (20 %)
Reggae	72.10 ± 3.51	163 (97 %)	91.70 ± 1.64	68 (40 %)	93.50 ± 1.58	157 (93 %)
Rock	82.00 ± 2.62	6 (04 %)	89.60 ± 0.97	6 (04 %)	91.00 ± 2.98	101 (60 %)

Table 4.4: Evaluation of the feature selection based on the genre discrimination of every feature according to the Statistical Spectrum Descriptor on the GTZAN collection. The best classification accuracy with the corresponding standard deviation and the related number of selected features (relative amount of selected features) are listed for every one-vs.-rest genre situation and each of the two calculation models.

the J48 learning algorithm, while significantly more features were taken into account in the case of the SMO. However, a potential feature set reduction can be concluded according to all three learning models. A closer look at the respective genre evaluation results of table 4.4 reveals an aspect which have been also observed for the Rhythm Pattern descriptor. The genres Classical and Pop are always under the top three genres in terms of the achieved classification accuracy independent from the actual learning algorithm and calculation model. Only in the case of the Naive Bayes and the ReliefF the genre Classical is in fourth place. Consequently, these two genres appear to be better represented by the discriminative features of both the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor.

Balanced Information Gain						
	Naive Bayes		**J48**		**SMO**	
Genre	**Accuracy**	**# Features**	**Accuracy**	**# Features**	**Accuracy**	**# Features**
Classical	89.57 ± 2.36	23 (14 %)	92.04 ± 2.45	107 (64 %)	94.99 ± 2.45	152 (90 %)
Electronic	82.16 ± 4.75	6 (04 %)	88.89 ± 1.46	135 (80 %)	92.45 ± 2.17	129 (77 %)
Jazz & Blues	87.72 ± 4.35	12 (07 %)	96.78 ± 1.33	107 (64 %)	98.29 ± 1.17	96 (57 %)
Metal & Punk	92.73 ± 1.37	6 (04 %)	94.58 ± 1.10	101 (60 %)	96.91 ± 1.56	96 (57 %)
Rock & Pop	83.26 ± 2.92	40 (24 %)	88.68 ± 2.92	23 (14 %)	92.04 ± 2.11	90 (54 %)
World	67.15 ± 5.37	6 (04 %)	85.74 ± 2.34	45 (27 %)	90.19 ± 2.81	129 (77 %)
ReliefF						
Classical	91.01 ± 2.60	12 (07 %)	92.11 ± 1.95	157 (93 %)	94.86 ± 2.90	152 (90 %)
Electronic	81.55 ± 2.82	79 (47 %)	88.96 ± 1.98	45 (27 %)	93.07 ± 2.38	124 (74 %)
Jazz & Blues	92.94 ± 2.72	12 (07 %)	96.91 ± 1.23	62 (37 %)	98.42 ± 1.37	118 (70 %)
Metal & Punk	92.73 ± 1.58	6 (04 %)	94.99 ± 1.22	112 (67 %)	97.46 ± 0.86	96 (57 %)
Rock & Pop	82.71 ± 3.67	40 (24 %)	88.48 ± 2.77	17 (10 %)	91.49 ± 1.19	118 (70 %)
World	62.28 ± 6.34	6 (04 %)	85.94 ± 1.77	40 (24 %)	90.19 ± 2.23	157 (93 %)

Table 4.5: Evaluation of the feature selection based on the genre discrimination of every feature according to the Statistical Spectrum Descriptor on the ISMIR 2004 Genre collection. The best classification accuracy with the corresponding standard deviation and the related number of selected features (relative amount of selected features) are listed for every one-vs.-rest genre situation and each of the two calculation models.

ISMIR 2004 Genre

To compare the feature selection performance regarding the ISMIR 2004 Genre collection, figure 4.6 illustrates the classification accuracy of every feature selection candidate according to every genre. Again, only the calculation models Balanced Information Gain and the ReliefF were employed to generate the accuracy results of this figure. It can be observed clearly that the classification performances based on those two calculation models only differs slightly. In particular the variation of the classification accuracy is quite similar for each genre and also each learning algorithm. Contrary to the corresponding accuracy results based on the GTZAN collection, the observed variation of the classification accuracy is more limited for the Naive Bayes and the J48 learning algorithms. This means that the classification accuracy according to a selection of at least 51 of 168 most discriminative features varies by a margin of 2 % at most for the J48 and the SMO learning algorithms. In terms of the Naive Bayes the classification accuracy is always within a margin of 5 %. This important observation holds true for all one-vs.-rest genre situations but the genre World where a higher variation of the classification accuracy can be seen.

Table 4.5 lists the classification accuracy and the corresponding standard derivation as well as the related number of selected features for every one-vs.-rest genre situations. Similar to the GTZAN collection the average relative numbers of selected features according to each of the two calculation models and the three learning algorithms reveal an interesting conclusion. The average relative amount of selected features according to the Balanced Information Gain is 9.5 %

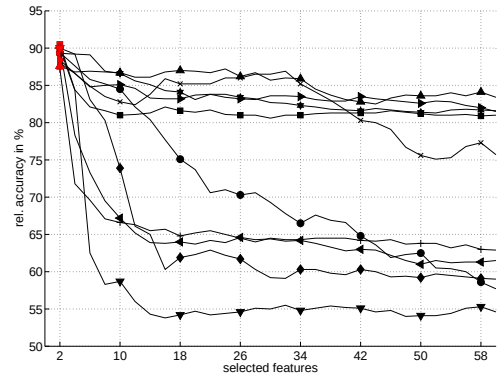
for the Naive Bayes, 51.5 % for the J48 and 68.7 % for the SMO learning algorithm. According to the calculation model ReliefF, the relative amount of selected features is 15.5 % for the Naive Bayes, 43 % for the J48 and 75.7 % for the SMO learning algorithm. This numbers approve the conclusion that in terms of the Statistical Spectrum Descriptor the expected reduction of the original feature set is considerable higher for the Naive Bayes, while the potential feature set reduction according to the J48 and the SMO learning algorithms is lower.

4.3.3 Rhythm Histogram

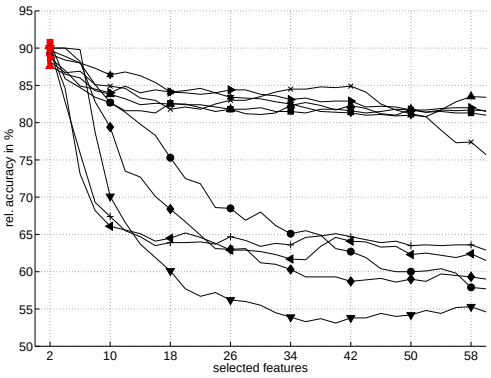
Like in discussion of the descriptors Rhythm Pattern and Statistical Spectrum Descriptor, at first the feature selection performances according to the GTZAN collection will be discussed. Figure 4.7 depicts the classification accuracy of all feature selection candidates according to every of the 10 one-vs.-rest genre situations. Again, the three learning algorithms Naive Bayes, J48 and SMO were used because they represent three quite different concepts of learning. The two heuristic discrimination models Balanced Information Gain and ReliefF were separately employed to generate the required feature ranking. This focus on only two of the five possible calculation models is sufficient, since the four calculation models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain achieve very similar feature rankings and therefore only marginal divergences could be observed in the performances of the corresponding feature selection evaluations.

The classification accuracy of the feature selection candidates based on the Balanced Information Gain are illustrated in the figures (a), (c) and (e) for the three learning algorithms Naive Bayes, J48 and SMO, respectively. The classification accuracy of the selection candidates evaluated with the Naive Bayes introduces an interesting difference to the corresponding accuracy results based on the other two examined descriptors. Those feature selection candidates which possess the highest classification accuracy contain very few features. This observation is valid for all 10 genres and means that a significant feature set reduction is directly related to a higher classification accuracy. Actually, this fact is a strong confirmation for the effectiveness of using the feature ranking based on the Balanced Information Gain to reduce the original feature set defined by the Rhythm Histogram descriptor. Contrary to the application of the Naive Bayes with the other two descriptors, the accuracy rate is close to 90 % for all genres in the case of the Rhythm Histogram descriptor. Another very promising fact is that the variation of the classification accuracy is also limited among the feature selection candidates of the same genre. Considering those feature selection candidates containing at least 18 of the maximum 60 features, the limitation of the accuracy variation is given by a margin of 1 % according to 8 of 10 possible genres. In fact, this is a promising limitation as it might not even be significant at all. A stronger variation of the classification accuracy can only be observed with respect to the genres Blues and Disco. From this observation follows that the classification accuracy will only be effected by an error of 1 % at most if the size of the original feature set is reduced by remarkable 70 %.

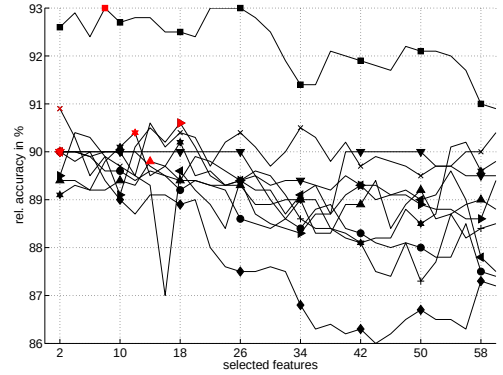
At first the accuracy results according to the Balanced Information Gain and the J48 learning



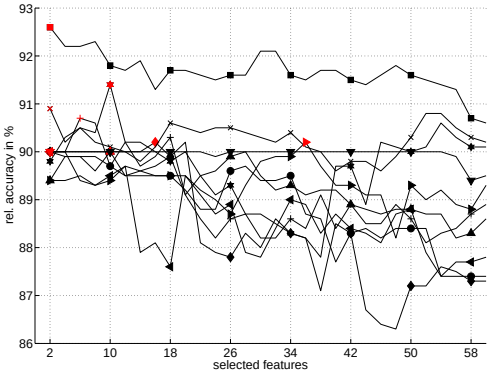
(a) Naive Bayes and Balanced IG



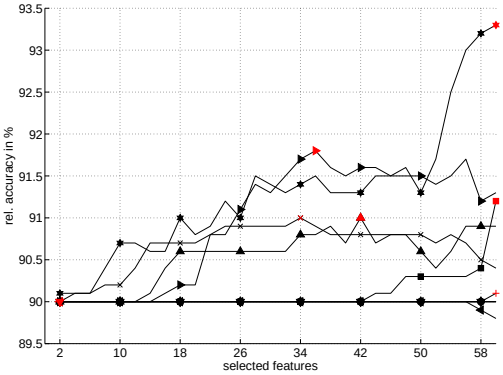
(b) Naive Bayes and ReliefF



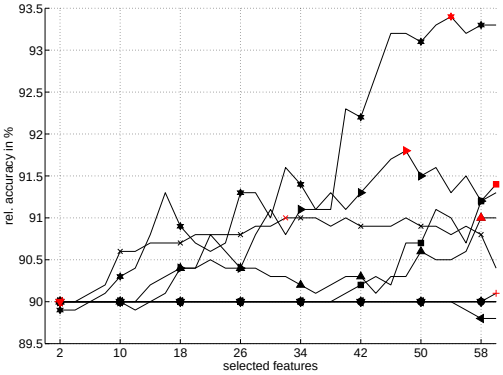
(c) J48 and Balanced IG



(d) J48 and ReliefF

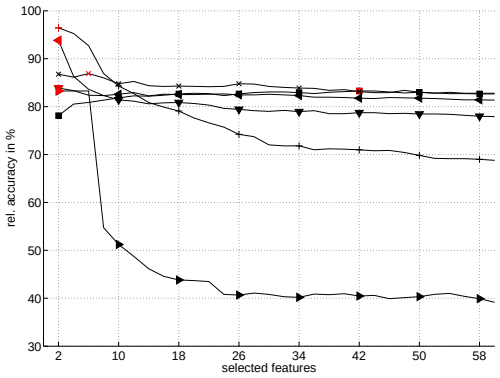


(e) SMO and Balanced IG

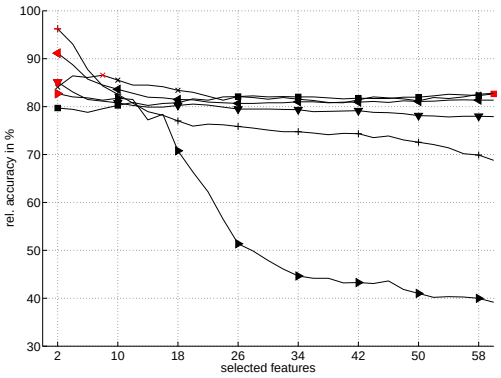


(f) SMO and ReliefF

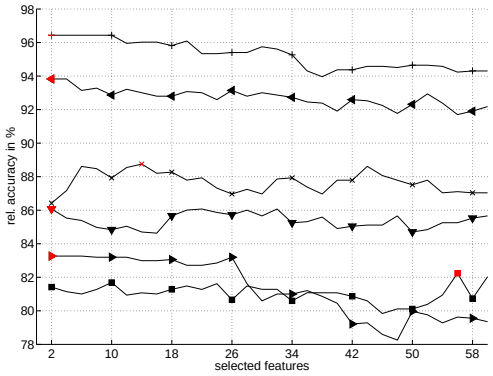
Figure 4.7: Classification accuracy results according to the Rhythm Histogram descriptor on the GTZAN collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $2 \cdot i$ most discriminative features respectively. Three different learning algorithms were employed with the calculation models Balanced Information Gain and ReliefF each. The following symbols constitute every individual genre: ● for Blues, ■ for Classical, ◆ for Country, × for Disco, ★ for Hip hop, + for Jazz, ◀ for Metal, ▶ for Pop, ▲ for Reggae and ▼ for Rock. A red (bright) symbol indicates the best classification accuracy achieved for the respective genre.



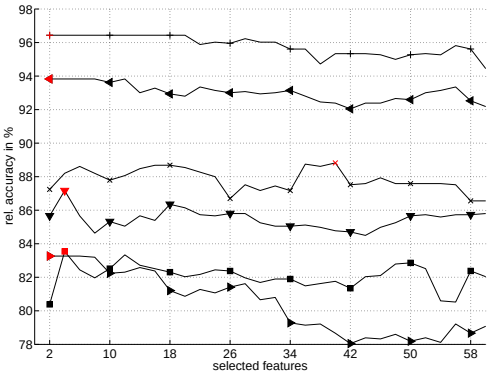
(a) Naive Bayes and Balanced IG



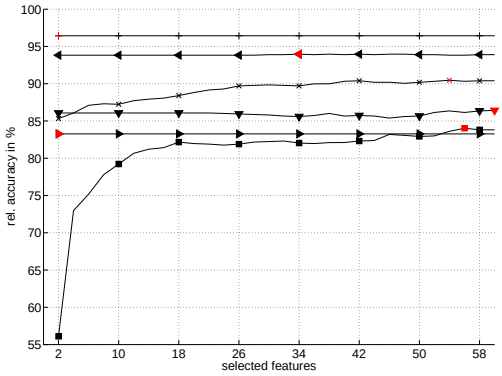
(b) Naive Bayes and ReliefF



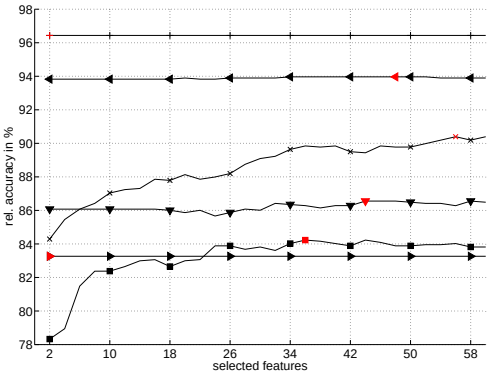
(c) J48 and Balanced IG



(d) J48 and ReliefF

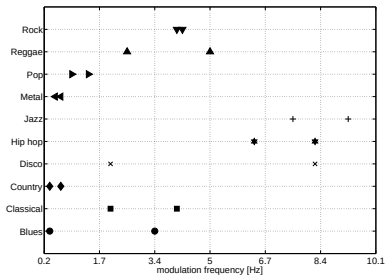


(e) SMO and Balanced IG

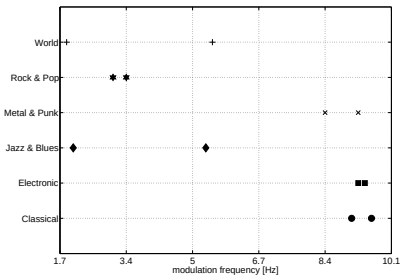


(f) SMO and ReliefF

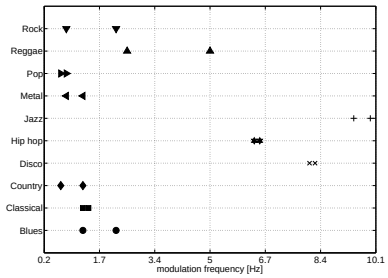
Figure 4.8: Classification accuracy results according to the Rhythm Histogram descriptor on the ISMIR 2004 Genre collection. 30 feature selection candidates have been evaluated where every selection candidate C_i with $1 \leq i \leq 30$ contains the $2 \cdot i$ most discriminative features respectively. Three different learning algorithms were employed with the calculation models Balanced Information Gain and ReliefF each. The following symbols constitute every individual genre: ■ for Classical, × for Electronic, + for Jazz & Blues, ◀ for Metal & Punk, ▼ for Rock & Pop and ▶ for World.



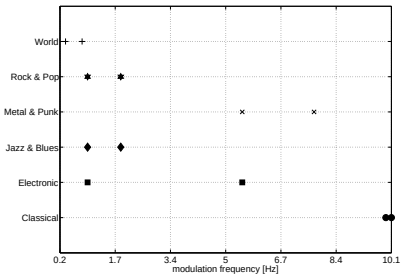
(a) Balanced IG and GTZAN



(b) Balanced IG and ISMIR 2004 Genre



(c) ReliefF and GTZAN



(d) ReliefF and ISMIR 2004 Genre

Figure 4.9: Illustration of the two most discriminative features of the Rhythm Histogram Descriptor according to every genre. The results of the figures(a) and (c) are based on the GTZAN collection, while the figures (b) and (d) illustrate results calculated on the ISMIR Genre 2004 collection.

algorithm illustrated in figure (e) appear to vary more comparing with corresponding results based on the Naive Bayes or the SMO learning algorithms. In fact, the classification accuracy does clearly oscillate between successive feature selection candidates of the same genre. Because of this oscillation various "peaks" exist along the progression of the classification accuracy. Such "peaks" actually occur seldom in the results based on both the Naive Bayes and the SMO learning algorithms. Regarding a Decision tree learning algorithm like the J48 this characteristic is more usual in terms of feature selection, since not only the contribution of the single features but also the contributions exhibited by the feature dependencies influence the performance of a Decision tree learner. Nevertheless, a considerable limitation regarding the variation of the classification accuracy can be recognized. Actually, the classification accuracies of the feature selection candidates having at least 26 of the maximum 60 most discriminative features are consistently within a 1 – 2 % margin for all 10 genres. The strongest variation of 3 % occurs for the selection candidates of the genre Country where the classification accuracy considerably deteriorate when more discriminative features were taken into account.

Considering the classification accuracy of the selection candidates according to the SMO learning algorithm depicted in figure (e), the classification accuracy is quite consistent among the selection candidates of the same genre as the achieved accuracy only varies within margin of 1 % to 2 % for all genres but Hip hop. In the case of Hip hop a variation of 3.5 % can be observed. This limitation is even stronger if only the selection candidates containing at least 18

Balanced Information Gain						
	Naive Bayes		**J48**		**SMO**	
Genre	**Accuracy**	**# Features**	**Accuracy**	**# Features**	**Accuracy**	**# Features**
Blues	89.40 ± 1.26	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)
Classical	90.50 ± 3.14	2 (03 %)	93.00 ± 1.76	8 (13 %)	91.20 ± 1.69	60 (100 %)
Country	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)
Disco	88.00 ± 2.58	2 (03 %)	90.90 ± 0.88	2 (03 %)	91.00 ± 0.94	34 (57 %)
Hip hop	89.30 ± 3.95	2 (03 %)	90.40 ± 2.55	12 (20 %)	93.30 ± 2.83	60 (100 %)
Jazz	87.20 ± 3.01	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.10 ± 0.32	60 (100 %)
Metal	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)
Pop	88.40 ± 2.07	2 (03 %)	90.60 ± 2.99	18 (30 %)	91.80 ± 2.35	36 (60 %)
Reggae	87.60 ± 2.37	2 (03 %)	89.80 ± 1.75	14 (23 %)	91.00 ± 1.33	42 (70 %)
Rock	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)
ReliefF						
Blues	89.20 ± 0.92	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)
Classical	90.80 ± 4.13	2 (03 %)	92.60 ± 1.35	2 (03 %)	91.40 ± 1.71	60 (100 %)
Country	90.00 ± 0.00	2 (03 %)	90.20 ± 1.69	16 (27 %)	90.00 ± 0.00	2 (03 %)
Disco	88.20 ± 2.44	2 (03 %)	90.90 ± 1.10	2 (03 %)	91.00 ± 0.94	32 (53 %)
Hip hop	89.70 ± 3.27	2 (03 %)	91.40 ± 2.22	10 (17 %)	93.40 ± 2.76	54 (90 %)
Jazz	90.00 ± 0.00	2 (03 %)	90.70 ± 0.82	6 (10 %)	90.10 ± 0.32	60 (100 %)
Metal	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)
Pop	88.70 ± 1.64	2 (03 %)	90.20 ± 3.05	36 (60 %)	91.80 ± 2.49	48 (80 %)
Reggae	87.60 ± 2.37	2 (03 %)	90.00 ± 1.15	10 (17 %)	91.00 ± 1.70	58 (97 %)
Rock	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)	90.00 ± 0.00	2 (03 %)

Table 4.6: Evaluation of the feature selection based on the genre discrimination of every feature according to the Rhythm Histogram descriptor on the GTZAN collection. The best classification accuracy with the corresponding standard deviation and the related number of selected features (relative amount of selected features) are listed for every one-vs.-rest genre situation and each of the two calculation models.

of the maximum 60 most discriminative features will be considered. In this case the variation is limited by a margin of 1 % for all genres but Hip hop. Another very important fact is that the variation of the classification accuracy is close to zero for 5 of the 10 genres defined by the GTZAN collection. Among these genres are also Country and Rock whereupon the feature selection results based on either the other two learning algorithms or the SMO together with the other two descriptors Rhythm Pattern and Statistical Spectrum Descriptor show a declining tendency when more discriminative features were used in the classification.

The figures (b), (d) and (f) illustrate the feature selection performances based on the ReliefF model and the three learning algorithms, respectively. Similar to the other two descriptors, the accuracy results based on the ReliefF do not vary much comparing with the corresponding results based on the Balanced Information Gain. Actually, the feature selection results according to the J48 and SMO learning algorithms exhibit such a high degree of similarity that the same

Balanced Information Gain						
	Naive Bayes		**J48**		**SMO**	
Genre	**Accuracy**	**# Features**	**Accuracy**	**# Features**	**Accuracy**	**# Features**
Classical	83.27 ± 2.25	42 (70 %)	82.24 ± 4.92	56 (93 %)	84.02 ± 3.76	56 (93 %)
Electronic	86.90 ± 1.75	6 (10 %)	88.75 ± 1.86	14 (23 %)	90.46 ± 1.74	54 (90 %)
Jazz & Blues	96.43 ± 0.29	2 (03 %)	96.43 ± 0.29	2 (03 %)	96.43 ± 0.29	2 (03 %)
Metal & Punk	93.83 ± 0.02	2 (03 %)	93.83 ± 0.02	2 (03 %)	93.96 ± 0.29	34 (57 %)
Rock & Pop	83.88 ± 2.73	2 (03 %)	86.08 ± 0.32	2 (03 %)	86.42 ± 1.69	60 (100 %)
World	83.26 ± 0.34	2 (03 %)	83.26 ± 0.34	2 (03 %)	83.26 ± 0.34	2 (03 %)
ReliefF						
Classical	82.65 ± 2.51	60 (100 %)	83.54 ± 2.86	4 (07 %)	84.23 ± 3.18	36 (60 %)
Electronic	86.56 ± 2.78	8 (13 %)	88.82 ± 1.86	40 (67 %)	90.40 ± 1.56	56 (93 %)
Jazz & Blues	96.23 ± 0.67	2 (03 %)	96.43 ± 0.29	2 (03 %)	96.43 ± 0.29	2 (03 %)
Metal & Punk	91.15 ± 1.62	2 (03 %)	93.83 ± 0.02	2 (03 %)	93.96 ± 0.43	48 (80 %)
Rock & Pop	85.12 ± 2.57	2 (03 %)	87.17 ± 2.96	4 (07 %)	86.56 ± 2.06	44 (73 %)
World	82.72 ± 0.47	2 (03 %)	83.26 ± 0.34	2 (03 %)	83.26 ± 0.34	2 (03 %)

Table 4.7: Evaluation of the feature selection based on the genre discrimination of every feature according to the Rhythm Histogram descriptor on the ISMIR 2004 Genre collection. The best classification accuracy with the corresponding standard deviation and the related number of selected features (relative amount of selected features) are listed for every one-vs.-rest genre situation and each of the two calculation models.

conclusions can be made.

Only the accuracy results according to the Naive Bayes differ as the variation of the accuracy is higher among the feature selection candidates of the same genre for 6 of the 10 genres in comparison to the corresponding results based on the Balanced Information Gain. This means that a similar variation limit of 1 % is only valid among feature selection candidates having at least 42 features. This refers to a feature set reduction of approximately 30 %. The genres Disco and Reggae are related to a slightly higher variation of the classification accuracy.

Basically, both the Balanced Information Gain and the ReliefF are proper models to compute a feature ranking for feature selection. Actually, at least 30 % of the original feature set can be reduced without having a considerable decline in the classification accuracy. In the case of the Balanced Information Gain a feature set reduction of approximately 55 % is even possible according to all three learning algorithms. A decline of at most 2 % must be expected which might not even be significant in some cases. In order to compare the best classification accuracy and the corresponding standard derivation as well as the related number of selected most discriminative features for every one-vs.-rest genre situation, table 4.6 presents an overview based on both calculation models and the three used learning algorithms. The results based on the Naive Bayes are quite surprising because the best classification accuracy according to every one-vs.-rest genre situation was achieved by taking only the two most discriminative features into account. These two features only represent a relative amount of 3 % according to the original feature set. More surprisingly, this observation is valid for both calculation models. Figure 4.9 visualizes these two

most discriminative features for both calculation models, respectively, in order to emphasize the modulation frequencies constituted by those features which appear to be such decisive regarding the performed one-vs.-rest genre classification. The illustration shows that for the most genres the modulation frequencies according to the two most discriminative features do notably diverge in the case of the Balanced Information Gain. Contrarily, the modulation frequencies of the two most discriminative features based on the ReliefF are quite similar for 7 of 10 genres. Another interesting observation is that the classification accuracy is equal or better according to all one-vs.-rest genre situations than the corresponding accuracy results based on the Naive Bayes with the Rhythm Pattern descriptor or the Statistical Spectrum Descriptor. Also the difference between the best achieved classification accuracy is considerable smaller for every genre.

In order to compare the potential reduction of the feature set in a more convenient way, again the average relative number of selected features has been calculated for each calculation model and learning algorithm. According to the Balanced Information Gain the average relative amount of selected features is 3 % for the Naive Bayes, 10.4 % for the J48 and 49.9 % for the SMO learning algorithm. Regarding the classification accuracy, the SMO basically performed best, but for the genres Jazz & Blues and World all three learning algorithms achieved the same accuracy. The Naive Bayes and the J48 performed quite similar. In terms of the calculation model ReliefF, the relative amount of selected features is 3 % for the Naive Bayes, 14.6 % for the J48 and 53.2 % for the SMO learning algorithm. Considering the achieved classification accuracy results, the SMO learning algorithm performed best for all 6 one-vs.-rest genre situations, while the Naive Bayes achieved a slightly worse performance than the J48. Especially the average numbers of selected features according to the Naive Bayes and the J48 are very remarkable, since these results suggest that a feature ranking based on the Rhythm Histogram descriptor implies the highest reduction of the original feature set in comparison to the other two discussed descriptors. Also the average feature set reduction according to the SMO learning model is better than the reduction achieved in terms of the Statistical Spectrum Descriptor. From this follows that a potential feature set reduction can be concluded for all three learning models.

ISMIR 2004 Genre

Comparing to the GTZAN collection, the feature selection performance based on the ISMIR 2004 Genre collection is presented in figure 4.8 where the two heuristic calculation models Balanced Information Gain and ReliefF were used again.

The classification accuracy of the feature selection candidates according to the two calculation models only differs slightly for the Naive Bayes and the J48 learning algorithms. In fact, the accuracy results based on these two learning algorithms also suggest that both the Balanced Information Gain and the ReliefF are effective heuristics to build a feature ranking for feature selection because the respective variation of the classification accuracy of the corresponding selection candidates is quite limited for almost all of the 6 genres defined by the ISMIR 2004 Genre collection. Actually, the results based on the Naive Bayes reveal a variation margin of approximately 5 %, while the variation margin related to the results according to the J48 is given

with 3 %. In both cases the accuracy results representing the genre World have to a stronger variation. Contrarily, the results of the two calculation models considerably diverge in the case of the SMO learning algorithm as the variation of the accuracy related to 4 of the 6 genres are tightly limited by a margin of less than 1 %. The Balanced Information Gain appears to be more useful as that strong limitation of the classification accuracy holds even true for 5 genres when feature selection candidates having a number of at least 18 of the maximum 60 features were considered only. According to the ReliefF the same scale of limitation can only be observed when feature selection candidates having a number of at least 34 features were taken into account.

Table 4.7 lists the classification accuracy and the corresponding standard derivation as well as the related number of selected features for every one-vs.-rest genre situation. Similar the results based on the GTZAN collection, only the two most discriminative features were taken into account to achieve the best classification accuracy for 4 of maximum 6 genres. Although the data instances of the ISMIR 2004 Genre collection are not equally distributed among the 6 genres, this similar feature selection performance is promising comparing with the GTZAN collection. Again, the average relative numbers of selected features according to each of the two calculation models have been calculated to better compare the feature selection performance among the three discussed descriptors. The average relative amount of selected features according to the Balanced Information Gain is 15.3 % for the Naive Bayes, 21.3 % for the J48 and 57.7 % for the SMO learning algorithm. According to the calculation model ReliefF, the relative amount of selected features is 20.8 % for the Naive Bayes, 15 % for the J48 and 52 % for the SMO learning algorithm. Comparing to the corresponding results of the Rhythm Pattern descriptor and the Statistical Spectrum Descriptor, the average numbers of selected features imply that the expected reduction of the original feature set is considerable higher for the J48 and the SMO learning models and even the potential feature set reduction according to the Naive Bayes is only marginal lower. In the case of ISMIR 2004 Genre collection the ReliefF model outperforms the Balanced Information Gain, while for the GTZAN collection the Balanced Information Gain is the better heuristic discrimination model.

4.4 Conclusion

This chapter presented an evaluation of a feature selection approach based on a feature ranking according to the contribution of every feature for genre discrimination. The key advantage of this feature selection approach is that the feature evaluation can be performed without the need of splitting the original training and test set into specific sets which are individually used for the feature evaluation and the actual learning. Contrary to the Wrapper feature selection approach discussed in [31], this means that both the feature selection algorithm and the successive learning algorithm can use the full training and test set.

The feature selection evaluation compared the contributions of the five heuristic discrimination models Chi-square, Information Gain, Gain Ratio, Balanced Information Gain and ReliefF to estimate an effective feature ranking for feature selection. The three descriptors Rhythm Pat-

tern, Statistical Spectrum Descriptor and Rhythm Histogram represented the original feature sets used in the feature selection evaluation where all three descriptors were available for both the GTZAN and the ISMIR 2004 Genre collection. Also the three learning algorithms Naive Bayes, J48 and SMO were separately employed during the evaluation because they represent three quite different concepts of learning. Moreover, the feature selection evaluation was done based on one-vs.-rest genre situations only. Two key questions were examined during the feature selection analysis. The first question concerned the degree of consistency regarding the classification accuracy which was observed among the feature selection candidates of the same genre. In particular the actual limitation regarding the variation of the classification accuracy was examined as it represents a decisive aspect of the quality of the feature selection approach. The second question was related to the potential scale of feature set reduction which can be expected for all genres according to the respective learning algorithm and descriptor.

The evaluation clearly showed that the feature selection performances according to the four calculation models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain marginally vary. Since the accuracy results based on the Gain Ratio and the Balanced Information Gain suggested a slightly higher consistency according to the classification candidates and also slightly better classification accuracy for many genres, these two calculation models should be preferred against the Chi-square. The feature selection performance based on the ReliefF diverged more where in particular the limitation of the classification accuracy was different.

The results of the feature selection evaluation suggested that the classification accuracy of the selection candidates related to the same genre was actually strongly limited although the actual scale of limitation particularly depended on the learning algorithm used and the descriptor. In many cases the accuracy was limited by a margin of approximately 5 % where some stronger variation were observed for some specific one-vs.-rest situations. Yet, already this limitation is remarkable as it basically held true for all three descriptors, the two examined music collections and the three learning algorithms. More specifically, in terms of the J48 and the SMO in particular, the limitation of the variation given a by margin of 1 – 2 % was stronger for almost all genres. Only the results based on the genres Country and Rock of the GTZAN collection varied stronger, by at most 4 %. In the case of the Rhythm Histogram descriptor, the variation of the classification accuracy was always within a margin of 1 % for all learning models and at least 8 of 10 genres according to the GTZAN collection and 4 of 6 genres according to the ISMIR 2004 Genre collection. An important conclusion is that the limitation of the accuracy variation was always within a margin of 1 – 2 % for all music collections, learning algorithms and descriptors if the feature selection candidates containing at least 50 % of the most discriminative features were considered only. From this observation can be followed that the original feature set size can actually be reduced by 50 % without having a considerable, in some situation even a potentially insignificant, decline of the classification accuracy. Thus, the effectiveness of the feature selection approach based on genre discrimination can definitely be concluded.

The results according to the average relative number of selected features over all genres implied a strong dependence on the music collection used. The average relative number of se-

lected features was calculated by only considering those selection candidates over all one-vs.-rest genre situations for which the best classification accuracy was achieved. Nevertheless, important conclusions could be made holding true for all three music collection. Considering the GTZAN collection, the potential average feature set reduction according to the Rhythm Histogram descriptor was remarkably good for the two learning algorithms Naive Bayes and the J48. More specifically, for the Naive Bayes the achieved feature set reduction was 97 % according to the Balanced Information Gain and the ReliefF. In the case of the J48 the achieved feature set reduction was 89.6 % and 85.4 % according to the Balanced Information Gain and the ReliefF, respectively. The largest feature set reduction according to the SMO learning algorithm was observed in terms of the descriptors Rhythm Pattern and Rhythm Histogram. In the case of the Rhythm Pattern descriptor an average reduction of 62 % and 50.1 % was achieved according to the Balanced Information Gain and the ReliefF, respectively, while an average reduction of 50.1 % and 46.8 % was observed with the Rhythm Histogram descriptor. The feature set reduction regarding the Statistical Spectrum Descriptor is 30 % according to the SMO and approximately 70 % for the other two learning algorithms. According to the GTZAN collection, it could definitely be followed that the feature ranking based on the Balanced Information Gain achieved a larger the feature set reduction. The classification accuracy was considerably better in the case of the SMO learning algorithm compared with the Naive Bayes and the J48. Very interesting is that the performance of the SMO is only slightly better according to the Rhythm Histogram descriptor, while the margin between the accuracy achieved by the SMO and the other learning algorithms is considerably larger in the case of the other two descriptors.

In the case of the ISMIR 2004 Genre collection, the dependency on the applied learning algorithm was more important in order to conclude which calculation model achieves a larger feature set reduction together with a limited decline of the classification accuracy. According to the Naive Bayes the Balanced Information Gain performed better, while the ReliefF model should be preferred for the Decision tree *J*48. According to the SMO learning algorithm, the Rhythm Histogram descriptor with the ReliefF outperformed the Balanced Information Gain with a feature set reductions of 48 % and 42.3 %, respectively, while the Balanced Information Gain was the better model for feature ranking in terms of the other two descriptors. Nevertheless, the average feature set reduction among the best classification accuracy of every genre was also remarkable in terms of the ISMIR 2004 Genre collection. For both the Naive Bayes and the J48 an average feature set reduction of approximately 50 % was achieved according to all three descriptors where the average reduction was enormously high with 75 % or even more in the case of the Rhythm Histogram descriptor. For the SMO learning algorithm a feature set reduction of at least 40 % was achieved in with the descriptors Rhythm Pattern and Rhythm Histogram, while a reduction of at least 25 % was observed with the Statistical Spectrum Descriptor. It could be concluded that the Balanced Information Gain performed better with the Naive Bayes and the SMO learning algorithms, while the ReliefF appeared to be preferable with the J48 learning algorithm. Regarding the achieved classification accuracy quite the same conclusions were valid as in the case of the GTZAN collection.

Chapter 5

Applications

This chapter focuses on two specific applications which have been designed and implemented to obtain significant empirical data to examine all questions defined in this thesis. The first application represents the discriminant analysis tool *DiscriminationAnalyzer* which has been partly developed in the popular scientific programming environment MATLAB. Section 5.1 describes the intended purpose as well as all main components of the DiscriminationAnalyzer tool. Moreover, this section also gives a compact overview of all included user interface windows and its most relevant user controls and actions.

The second application which is described in section 5.2 is not actually meant to be a standalone system with an own user interface but rather extends the popular machine learning workbench WEKA. The purpose of this extension is to provide new learning algorithms which on the one hand can simultaneously handle multiple different feature sets regarding a unique classification problem, and on the other hand use a hierarchical taxonomy for classifying.

5.1 DiscriminationAnalyzer

Although the abstract idea of using a potential discriminative power analysis of variables concerning the determination of underlying classes has been originally established for problems basically based on all types of information, results of this approach also imply meaningful conclusions

to Music Information Retrieval. In comparison to some other feature evaluation techniques the key advantage of using a discriminant analysis to measure the "quality" of features is that the actual evaluation can be performed on the very same training set which is also used to train the classification system. Thus, a further partitioning of the available training set is not required and both the evaluation as well as learning will certainly benefit by yielding more robust estimation results. Another advantage of this approach to for qualify musical descriptors is that the easy application, since discriminant analysis provides a generic solution for analyzing musical descriptors.

The key aim of the DiscriminationAnalyzer is to provide specific tools to perform an interactive discriminant analysis based on arbitrary heuristic discrimination models and feature sets in a convenient manner. This means that although the rhythmic descriptors Rhythm Patterns, the Statistical Spectrum Descriptor and the Rhythm Histogram are used within the thesis only, the DiscriminationAnalyzer tool works an arbitrary feature sets. Additionally, new heuristic discrimination models can also be added. Basically, the DiscriminationAnalyzer tool consists of two separate parts. The first part realizes the graphical and control components of the user interface and the second part includes the underlying computational components. As the original intention had been to design an analyzing tool which runs in MATLAB, user interface and all main control components were developed in the MATLAB environment. Thus, an active session of MATLAB is necessary to use the DiscriminationAnalyzer tool. However, most components concerning the calculation of the discrimination values, the evaluation of feature selections or other data processing are implemented with the WEKA workbench. WEKA is a Java-based open-source framework containing various supervised and unsupervised learning methods and has been first introduced in [57]. It should be noted that in addition of using original classes of the WEKA workbench also specific class extension and self-designed classes have been implemented. All classes are integrated into the WEKA class hierarchy to guarantee full compatibly.

The use of the DiscriminationAnalyzer tool enriches the deployment of discriminant analysis concerning arbitrary feature sets by providing the following core functionalities:

- Arbitrary datasets which are defined either in the popular ARFF format of WEKA or in the specific SOMLib dataset format are allowed for input. Detailed description of ARFF and SOMLib can be reviewed in [57] and, respectively, in [49]. Both dataset formats can also be chosen to save already processed datasets and selected feature subsets. Additionally, the discrimination values computed so far can be explicitly stored by using a self-created data format in order to allow convenient reuse.

- Basically, seven heuristic discrimination models can be chosen to calculate the discrimination values of features according to the currently loaded feature sets. Five of these seven calculation models are discussed in section 3.2. Moreover, the DiscriminationAnalyzer tool offers an interface to import arbitrary heuristic discrimination models implemented in Java into the currently running application[1].

[1] To guarantee successful inclusion of a new calculation model, it must be designed as a single Java class and

- Arbitrary feature sets are processable within the DiscriminationAnalyzer. Since quite many feature sets like rhythmic descriptors can also be represented in matrix form, a dynamic matrix visualization model has been implemented: Both matrix dimensions can be manually redefined to adapt the actual feature set representation at any time.

- Both an independent analysis of single feature sets and a simultaneous analysis of multiple feature sets are provided by the DiscriminationAnalyzer tool, whereas those feature sets can even refer to different music collections. As a consequence, the DiscriminationAnalyzer tool can be run either in a *single feature set mode* or in a *multiple feature set mode* and both modes can be manually set by the user.

- Visualization of the relation of the discrimination values against the ranking order based on a specific genre and the selected calculation model.

- Intuitive selection of k *most* discriminative features based on the visualization of the relation of the discrimination values against the ranking order.

- Examination of all computed discrimination values can be done either in graphical or numerical manner.

- Evaluation of feature subset selections by choosing arbitrary WEKA-based learning algorithms in connection with cross validation or simple validation by splitting into separate training and test sets. Additionally, an interface to integrate miscellaneous Java-based learning algorithms is also provided[2].

The main functionalities are grouped into six core components in order to increase the development efficiency and usability of the DiscriminationAnalyzer tool as a whole. Moreover, a specific *analyzing pipeline* defines the underlying concept of the DiscriminationAnalyzer tool and each component actually represents a specific step of this pipeline. The analyzing pipeline aggregates all main steps to establish meaningful discriminant analysis and successive feature subset selection. The main steps of the analyzing pipeline are: dataset input/output, data normalization, computation of the discrimination values, visualization, feature subset evaluation as well as evaluation set up. Table 5.1 gives an overview of the analyzing pipeline and its components.

The following contributions are responsible for the realization of the DiscriminationAnalyzer tool and its included components. My own contributions are the design as well as the complete implementation of all control components, the basic program architecture and the user interface. Furthermore, the heuristic discrimination model *Balanced Information Gain* which is introduced in [58] has been adjusted and implemented by myself. Also, an implementation of the *attribute-discrimination* model [11, 18] has been carried out by myself. All other heuristic discrimination

must also implement the abstract class *weka.attributeSelection.AttributeEvaluator* contained in the WEKA framework.

[2] All learning algorithms which are intended to be added into the DiscriminationAnalyzer tool must be designed in a single Java class and must also extend the abstract class *weka.classifiers.Classifier* of the WEKA framework.

Dataset input/output	ARFF data format (defined in WEKA), VEC data format (defined in SOMLib), MAT data format (self-defined)
Data normalization	Zero-Mean standardization, Max-Min length normalization, Squared-Sum length normalization *or without normalization*
Computation of discrimination values	Chi-square, Balanced Information Gain [58], Attribute Discrimination [11,18], Information Gain, Gain Ratio, ReliefF [33, 48], Symmetrical Uncertainty [58] *and an interface for adding additional calculation models*
Visualization	distribution of the discrimination values, color matrix representation, feature set correlation, feature set value frequency *and various visualizations to examine already calculated discrimination values*
Selection evaluation	Decision tree J48, Nearest Neighbor (lBk), Naive Bayes, OneR, SMO (i. e. SVM), Random forest, ZeroR *and an interface for adding additional learners*
Evaluation set ups	Cross validation or train/test set split

Table 5.1: Core components of the DiscriminationAnalyzer

models which are ad hoc provided in the DiscriminationAnalyzer tool are already included in the standard WEKA workbench and, therefore, have been integrated without any modification.

5.1.1 Controlling the analyzing pipeline

The main window of DiscriminationAnalyzer tool which is shown in figure 5.1 includes all control elements and options to modify the analyzing pipeline accordingly. This compact interface design has been chosen in order to provide simple and efficient usage. Basically, the interface is divided into three separate control layers which represent the data/feature set input selection, adjustable options concerning the computation of the discrimination values and the actual visualization of the respective computation.

The data/feature set input control is realized in the upper section of the main window. Because of including two separate list boxes within the data input selection the feature subset selection which will be actually incorporated into successive discriminant analysis is absolutely independent from the collection of currently loaded feature sets. This separation of loaded and selected feature set sets guarantees a convenient way to manage large sets simultaneously. Thus, only those feature sets which are actually listed in the right list box will be considered for successive computations of discrimination values. Additionally, on the right hand side of this upper section general information about the selected feature set set as well as the related music collection is displayed.

The middle section of the main window provides both adjustable options concerning the analyzing pipeline and controls to actually initiate computations. The first row of this section

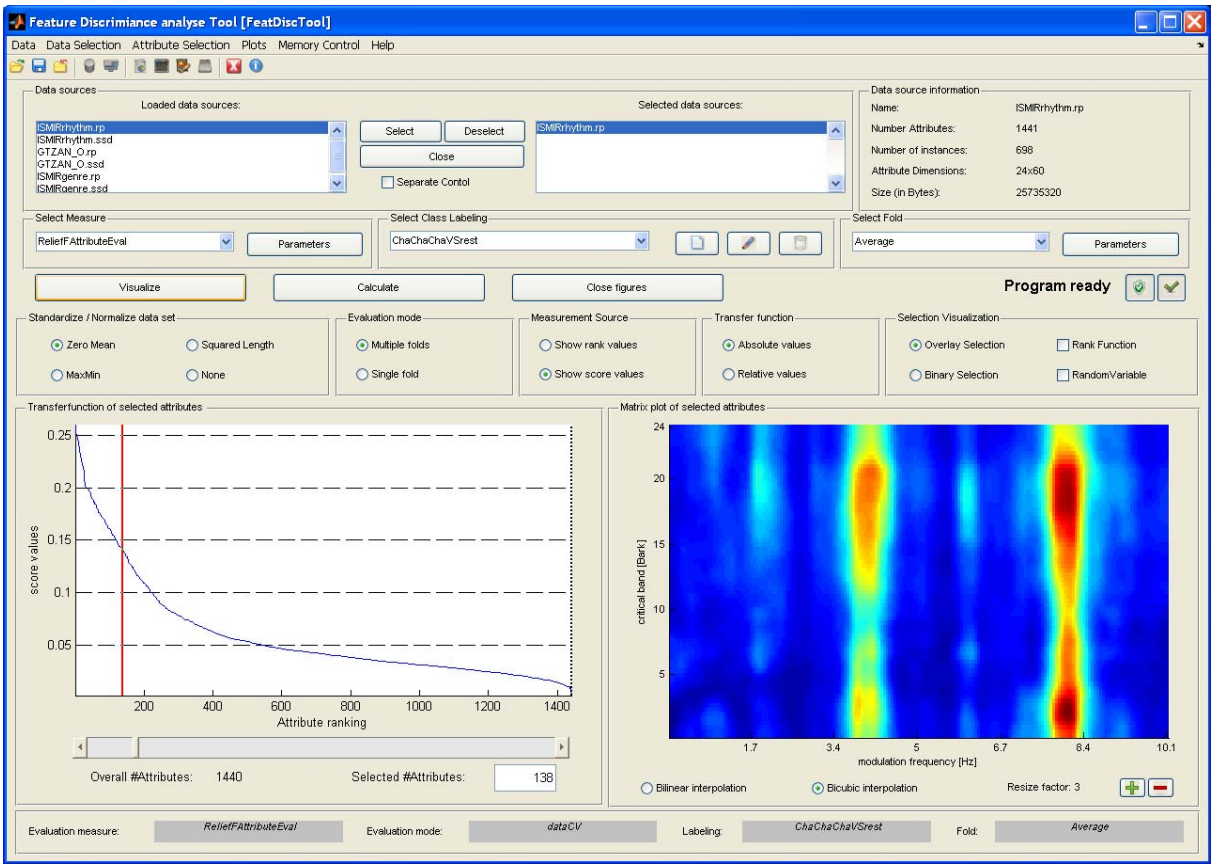


Figure 5.1: Main window of the DiscriminationAnalyzer

includes options to choose a particular calculation model which estimates the discriminative power of all features according to a given feature set. Furthermore, options to define the labeling situation which should be used during the computation and, eventually, to set the current calculation fold. The labeling situation refers to the underlying class (e. g. genre) range which will be used in all successive calculations. In case of using the multiple calculation fold mode, the current fold selection decides which results will be finally visualized by provided result plots. The successive row contains such controls to initiate calculations, to display obtained results and to show the current program status. Finally, various options concerning the computation of discrimination values as well as the visualization of the results are placed within the third row of the middle section. Following options can be set by the user – from left to right considering the order within the main window:

- Normalization/standardization of the feature set prior to the computation of the discrimination values.
- Two different computational modes to estimate the discrimination values. The first computational mode combines an independent calculation of multiple folds with an eventual rank correlation test to obtain final estimations for every single feature of the feature set. The rank correlation test is based on Kendall's rank correlation coefficient described in [1]

and estimates a statistically robust ranking of all included features according to a specific feature set. Eventually, this ranking represents the discriminative ranking among all features.

Contrary to the first, the second computational mode only applies a single calculation fold to estimate the discriminative ranking. Comparing to the multiple calculation fold approach, the single calculation fold mode often causes a distorted and, therefore, unreliable ranking. On the other hand the computation time significantly decreases.

- Either the rank value or the actual discriminative measurement value is used for successive steps of the analysis, e. g. the visualization of the computed results.
- Display of the relation of the discrimination values against the ranking order with respect to absolute or relative measurement values. The application of relative measurement values imposes a normalization of the relation which can be desirable when the actual range of measurement values is too large.
- Visualization of user-defined feature selection either by using an overlay plot based on the original color matrix representation or by generating a binary matrix representation in which white matrix cells emphasize selected features.

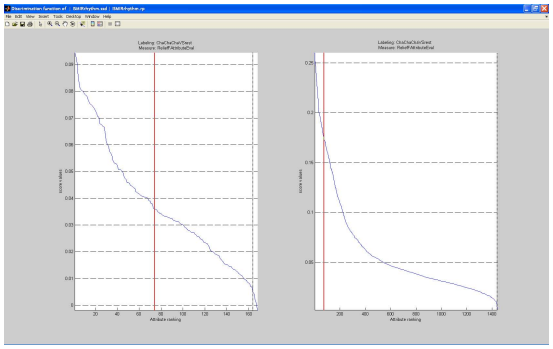
The bottom section of the main window displays the two main visualizations, namely a specific relation of the discrimination values against the ranking order (left) and the color matrix representation which conveniently illustrates the assignment of discrimination values to the corresponding features (right). In the case of the multiple feature set mode a combined output of all computational results would yield an unclear visual representation. Thus, only the respective visualization of a single feature set[3] is included in the main window and a combined visualization of all results will be shown in a separate window as figure 5.2 illustrates.

The menu bar at the top of the main window additionally enriches the possibilities of interaction. Various menus are included to provide fast access to data loading and saving, data selection and removal, to generate several different plots and to start the numerical examination as well as the feature subset evaluation. All plot types which are described in the visualization component of table 5.1 can be chosen via the corresponding menu. The numerical examination of the calculation of the discrimination values and the feature selection evaluation are explicitly reviewed in subsections 5.1.2 and 5.1.3. Eventually, a tool bar is also available directly below the menu bar for convenient access of some frequently used actions.

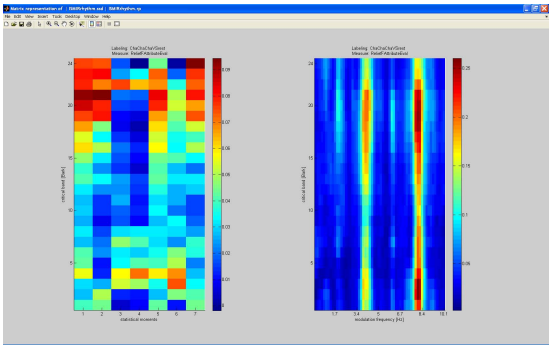
5.1.2 Evaluation of discrimination results

All previously computed discrimination values according to selected feature sets can be examined either in a graphical or in a numerical manner. Among others the graphical representation is mainly established by two related plots which on the one hand visualize the relation of the

[3] This feature set can be chosen by selecting the corresponding feature set in the selection list box.

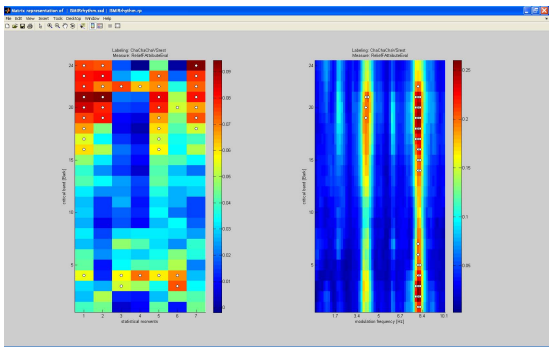


(a) Relation of the discrimination values against the ranking order

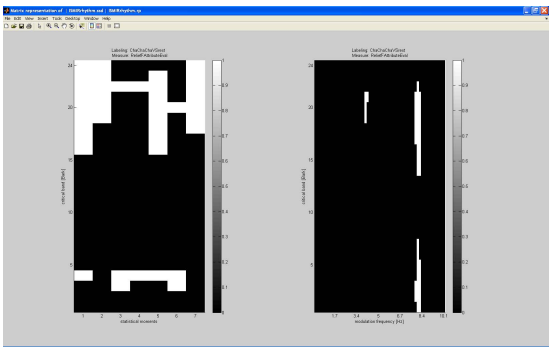


(b) Matrix representation

Figure 5.2: Simultaneous visualization of a Rhythm Patterns and Statistical Spectrum Descriptor related to the music collection ISMIR 2004 Rhythm.



(a) Overlay matrix plot



(b) Binary matrix plot

Figure 5.3: Feature subset selection of the 40 most discriminative features concerning Rhythm Patterns and Statistical Spectrum Descriptor which have been extracted from the music collection ISMIR 2004 Rhythm. Subfigure (a) illustrates the feature selection by an overlay plot. In (b), a binary representation is shown.

discrimination values against the ranking order, and on the other hand generate a color matrix to display the discriminative power of every feature conveniently. In the previous subsection both plots have been already declared as the main visualization of the computational results. If the single feature set mode is currently activated, both plots will be shown at the bottom section of the main DiscriminationAnalyzer window. This case is illustrated in figure 5.1. Otherwise, if the multiple feature set mode is activated, both the relation of the discrimination values against the ranking order and the color matrix representations of the corresponding feature sets will be displayed in separate windows. Figure 5.2 shows two separately generated plot windows.

In addition to the result visualization, the feature selection based on the discrimination values is a further core examination tool and can be directly combined with the original color matrix representation. The underlying selection mechanism aggregates the *best* k features, i. e. the k *most discriminative* features, whereas the upper bound k can be set by the user. If multiple feature sets are used simultaneously, the k most discriminative features of each set will be chosen

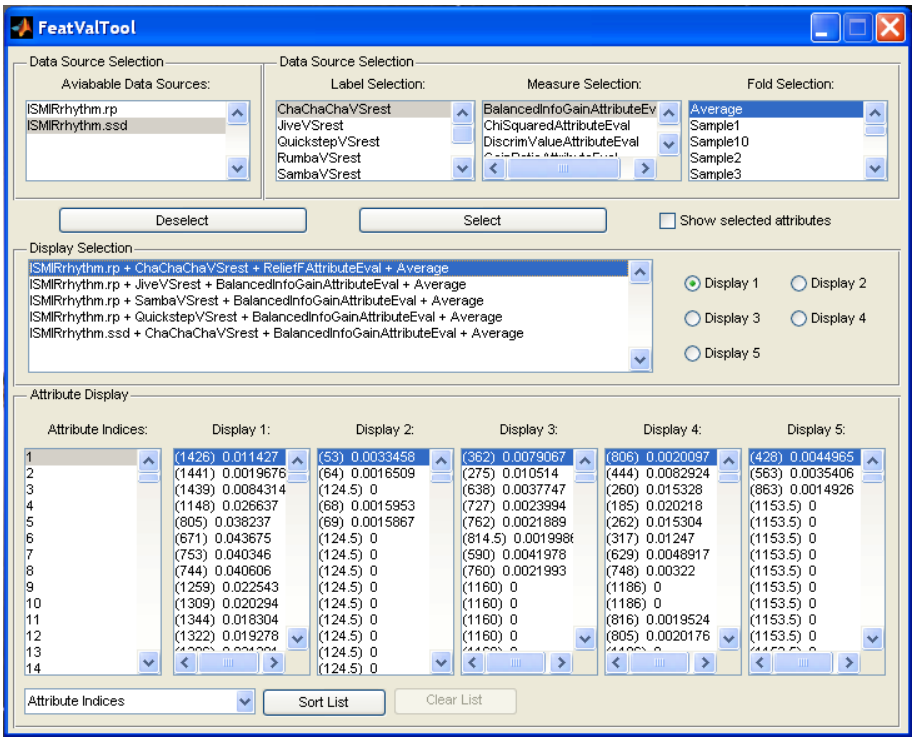


Figure 5.4: Numeric examination window of the DiscriminationAnalyzer

to build feature subsets, respectively. Yet, a particular reference feature set must be set by the user in order to define a common k. Regarding the selection illustration within the color matrix representation, either an overlay plot or a binary plot can be employed. The overlay plot depicts selected features with white dots on the corresponding matrix cells. In terms of the binary plot white matrix cells indicate selected features. An exemplary selection visualization with the upper bound $k = 40$ of both the overlay plot and the binary plot is shown in figure 5.3.

Besides the graphical representation of computational results a tabular summarization of currently available calculation results is provided in the DiscriminationAnalyzer tool. This tabular representation also supports arbitrary comparisons of those results based on different heuristic discrimination models or calculation modes as well. In figure 5.4, the separate numerical examination window is introduced which can be opened via the menu bar or tool bar of the main window. In the upper section of this window particular calculation result settings can be defined by choosing the feature set, the heuristic discrimination model, the labeling situation and a calculation fold[4]. The middle section contains controls to select predefined result settings for displaying and, eventually, the bottom selection actually shows the chosen result sets in tabular form.

[4] Results of the single calculation mode can also be chosen.

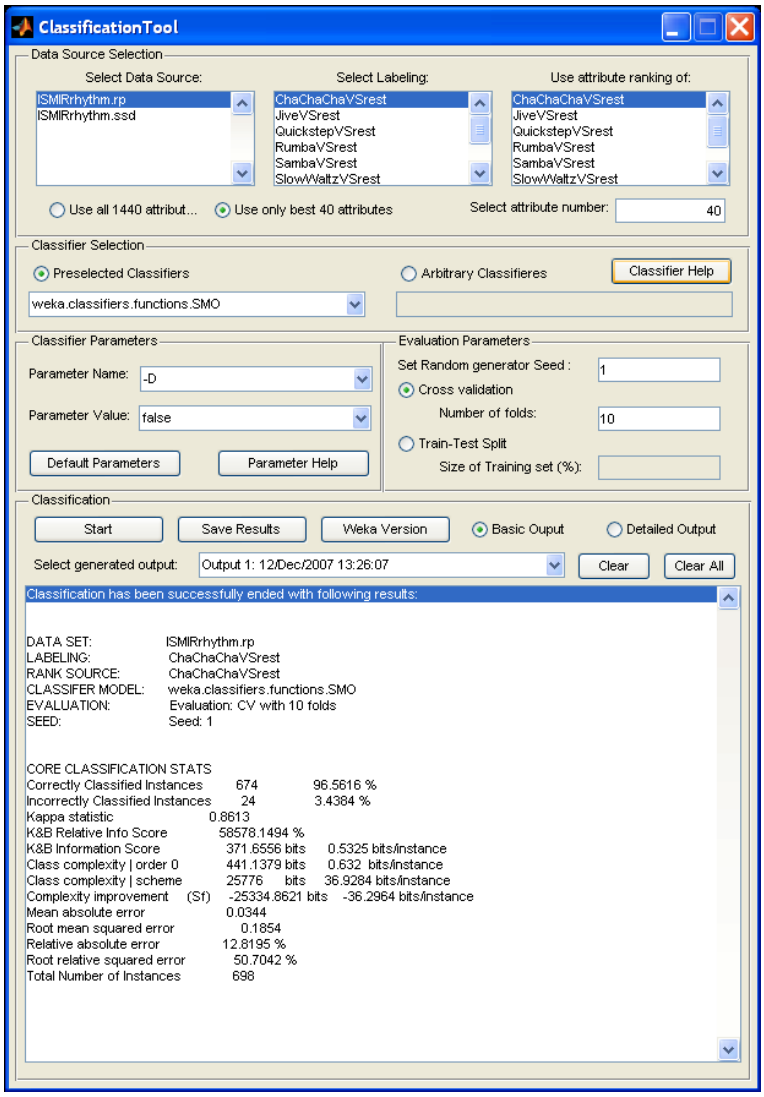


Figure 5.5: Evaluation window of the DiscriminationAnalyzer.

5.1.3 Feature selection evaluation

After the desired discrimination values have been calculated and the k most discriminative features have been successively selected, the next step of the analyzing pipeline is the specific evaluation of those selected features with respect to the classification performance. A separate evaluation window offers a basic evaluation environment by using WEKA-based learning algorithms and two different evaluation modes. Figure 5.5 visualizes this evaluation window which can be directly opened via the menu bar or the tool bar of the main DiscriminationAnalyzer window.

Similar to the numerical evaluation window introduced in the previous subsection, the upper section of the evaluation window contains controls to choose a particular feature set, the corresponding labeling situation, the calculation model for estimating the discriminative ranking and, eventually, a particular calculation fold. The middle section of the evaluation window provides

all available options for the evaluation environment consisting of both the selection of a specific learning algorithm and an evaluation mode. Several popular learning algorithms are integrated in the DiscriminationAnalyzer tool and can be chosen by the user. For instance Support vector machines, Decision trees and probabilistic learners are available. But arbitrary learning algorithms can also be imported if such models fulfill two requirements. First, the learning algorithm must be implemented within a single Java class. Second, the respective Java class must extend the abstract class *weka.classifiers.Classifier* which is included in the original WEKA workbench. Additionally, the user has the possibility to perform the evaluation either by a cross validation or by an usually faster train/test set split. Fundamental options to adequately adjust those basic evaluation procedures are also available. Eventually, the evaluation output and some basic output control are placed in the bottom section of this selection evaluation window.

5.1.4 Data input and output

The DiscriminationAnalyzer tool supports both the popular ARFF dataset format and the more specific SOMLib dataset format for both input and output. Only datasets which are described in either of these two formats can be successfully loaded. These two dataset formats can also be chosen to save given feature selections after corresponding calculations. Both loading and saving of data sets as well as feature selections can be initiated either by using the menu bar or the tool bar of the main window.

The ARFF dataset format was created to provide a compact definition of entire dataset including all features with their names, the value type and value range of every feature attribute and the actual data instances. The entire dataset description and all data instances are stored in a single ASCII-coded file which file name must append the extension `.arff`. A detailed overview of the ARFF dataset format can be reviewed in [57] and on the website of the WEKA machine learning workbench [55].

The SOMLib dataset format has been introduced in the SOMLib Digital Library Project [49] and is also used to extract the descriptors Rhythm Pattern, Statistical Spectrum Descriptor and Rhythm Histograms described in [38] or in related applications such as PlaySom [44]. It defines a *vector file* which contains the dataset description and all included data instances as well. The vector file is also coded in ASCII and the underlying file name must have the extension `.vec`. Contrary to ARFF, the SOMLib format does not explicitly include class assignments inside the vector file. Thus, a second file which is called the *ground truth file* (is defined as a tab-separated ASCII file) must be delivered as well. This ground truth file actually defines a particular class assignment for every included data instance. In order to successfully load or save datasets described by SOMLib, both the main vector file and the related ground truth file must be declared.

Additionally, to avoid the loss of calculation results by closing the current session, the DiscriminationAnalyzer tool supports an self-defined dataset format based on the usual MATLAB file format having the extension `.mat`. All currently computed discrimination results will be fully described by this dataset format. Contrary to the other supported dataset formats, only a

physical link, i. e. the absolute path to the dataset, will be stored instead of saving all data instances. When the link becomes invalid or the corresponding file has been created on another computer, the correct link referring to the dataset can be manually set via the menu bar of the main window.

5.2 Extension of the WEKA workbench

Basically, the Java-based open source machine learning workbench WEKA [55, 57] is used to evaluate feature subset selections based on the discriminative feature ranking pointed out in chapter 4. Since some more complex learning algorithms and evaluation procedures were employed during the evaluation phase of this thesis, the original WEKA framework had to be enhanced accordingly by designing new classes or extending already existing classes. In particular, original WEKA does not contain a hierarchical learning algorithm or can not simultaneously process several feature sets within a single learning algorithm. Although these WEKA extensions do not constitute an explicit stand-alone system in relation with a graphical user interface, I have aggregated all those WEKA extensions in a second application. Because all self-designed interfaces and classes are directly embedded into the original WEKA workbench, this application has been called *Extended WEKA*.

A detailed UML class diagram of relevant interfaces and classes which are part of Extended WEKA is illustrated in figure 5.6. It is worth noting that in parallel to the original WEKA classifier structure, which always defines `weka.classifiers.Classifier` as the top class, an additional classifier structure has been created to handle the simultaneous use of multiple feature sets within a single learning algorithm. Every learning algorithm which includes this new class structure must also implement the primary class structures in order to guarantee full compatibility with WEKA-based applications. The following list gives a compact description of the most relevant members of Extended WEKA:

MultipleSetClassifier introduces the design of *multiple set learning algorithms* into the WEKA framework which do actually accept multiple independent feature sets extracted from the same underlying data source. Basically, the idea of combing classification results based on different feature sets was already pointed out in [30], but in particular Flexer et al. [21] emphasize the use of this approach in context of music classification. As standard WEKA framework does only support multiple classifier combination on a single feature set[5], the design of a new class structure which has this interface as the top is ineluctable. However, the interface `MultipleSetClassifier` should always be implemented parallel to the standard WEKA classifier structure[6] in order to guarantee full compatibility within the WEKA framework.

[5] Class `weka.classifiers.meta.Vote` provides classifier combination on a single feature set by different combination rules (e. g. average, max, min) which can be surveyed in [30].

[6] A respective classifier class should implement both the standard class `weka.classifiers.Classifier` and the interface `MultipleSetClassifier`.

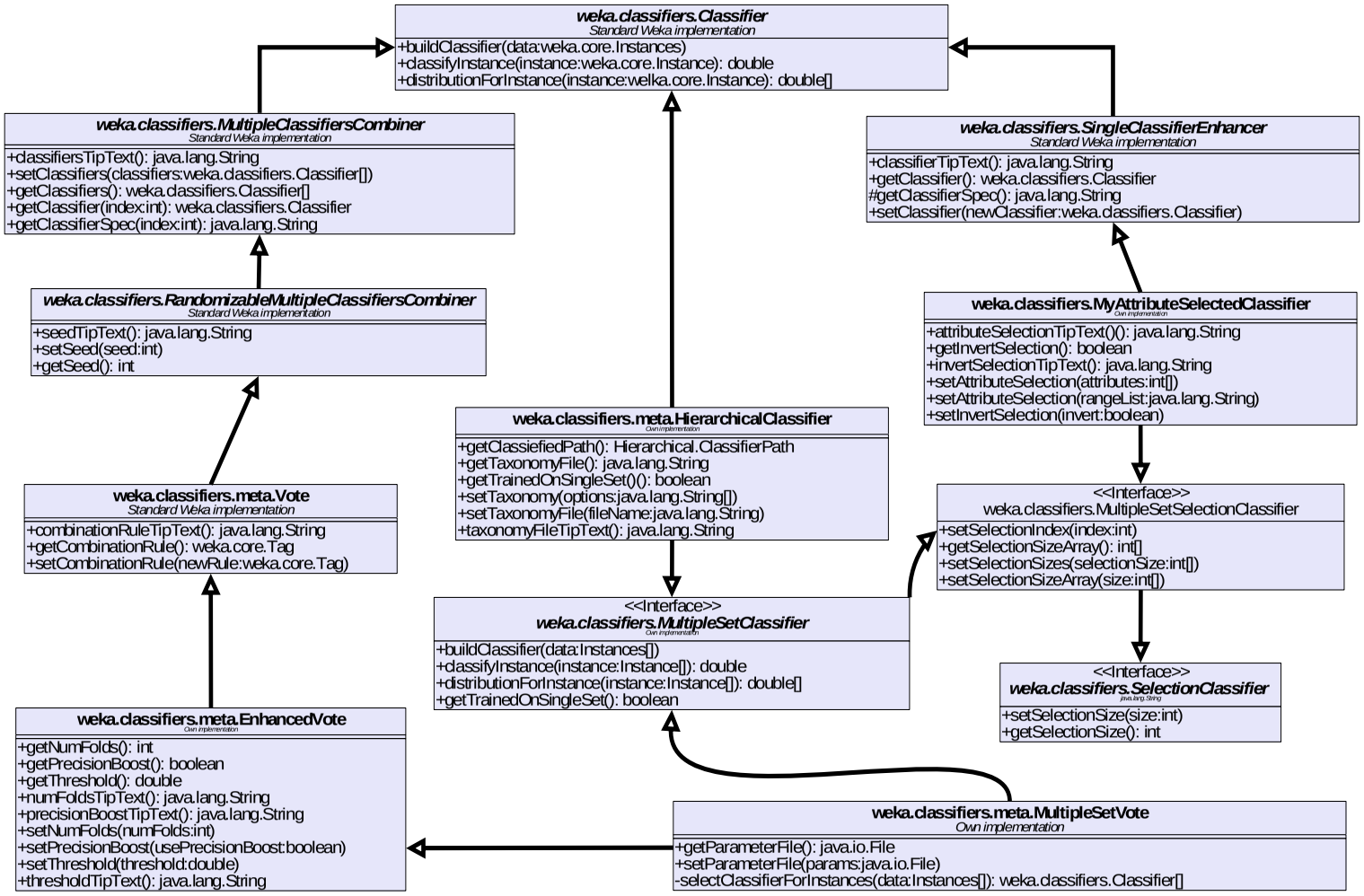


Figure 5.6: UML class diagram of all relevant WEKA extensions. Public members are denoted with the prefix "+", private members have the prefix "−" and protected members are emphasized with the prefix "# ".

SelectionClassifier provides additional control to restrict the number of features to the *first k* features which will be used during the successive learning and classification of the underlying learning algorithm. This selection is always performed on a single feature set.

MultipleSetSelectionClassifier establishes *first k* feature selection on multiple feature sets and therefore is directly related to both interfaces `MultipleSetClassifier` as well as `SelectionClassifier`. The implementation of this interface should guarantees the definition of selection sizes for each feature set independently.

EnhancedVote extends class `weka.classifiers.meta.Vote` of the original WEKA framework by additionally including the combination rule *Sum of probabilities* to aggregate probability results of multiple learning algorithms. Because this class should be considered as an extension of the primary WEKA class, it does not support the simultaneous application of multiple feature sets.

MyAttributeSelectedClassifier employs a user-defined feature selection prior to successive learning and classification procedures. Contrary to the interfaces `SelectionClassifier` and `MultipleSetSelectionClassifier`, this meta learning algorithm performs an arbitrary feature selection by assigning a list of feature indices corresponding to the underlying data instance. The actual learning algorithm must also be defined during initialization. It is worth noting that this meta learning algorithm is required to sufficiently realize the `HierarchicalClassifier` in order to offer particular feature selections at every included inner node of the class taxonomy.

MultipleSetVote combines the concept of simultaneous use of multiple feature sets with multiple independent learning algorithms. Thus, a `MultipleSetVote` learner is a very generic meta learning algorithm which on the one hand can process a single feature set or multiple feature sets simultaneously, and on the other hand employs a single or a combination of several learning algorithms. Additionally, the assignment of given learning algorithms to corresponding feature sets can be done either by the user or automatically due to prior accuracy estimations. To obtain accuracy estimation for a particular learning algorithm, separate cross-validation procedures for every available feature set will be employed. Subsection 5.2.1 refers to implementation details as well as relevant adjustable options.

HierarchicalClassifier performs hierarchical classification based on a user-defined taxonomy. Although this learning algorithm is used within this thesis only in context of musical genre classification, it has been deliberately implemented to support arbitrary classification problems which can be effectively solved by a hierarchical classification approach. Taxonomy definitions have to be formulated by using the standard WEKA XML scheme which is originally introduced to describe classifiers or entire experiment environments. The XML-based taxonomy description can be easily imported into a new `HierarchicalClassifier` instance by directly defining the absolute path of the file source. Furthermore, classification

Processing mode	Description
Single-Multiple	A single learning algorithm is employed to different and independent feature sets.
Multiple-Single	Multiple independent learning algorithms use the same feature set respectively.
Multiple-Multiple	Multiple independent learning algorithms use different and independent feature sets respectively.

Table 5.2: Available processing modes of `MultipleSetVote`.

results are represented by a *classification path* which contains the entire path from root to the respective leaf node which actually depicts the eventual class assignment. Thus, all assigned sub classes[7] as well as the actual result class itself are easily accessible. Subsection 5.2.2 focuses on implementation details and summarizes relevant adjustable options of the `HierarchicalClassifier` as well as some examples of valid XML taxonomy definitions.

In addition to above described classes and interfaces, Extended WEKA also includes classes to perform particular experiments based on the choices of learning algorithms, experiment modes and feature sets. As original WEKA again only supports single feature set learning, existing experiment classes[8] have been adapted to handle the simultaneous use of multiple feature sets. Consequently, classification experiments will only support the full scale of possibilities regarding `MultipleSetVote` and `HierarchicalClassifier` if these experiment class extensions are used.

Eventually, some notes concerning the programming environment and the used original WEKA framework should be given. Extended WEKA is based on the WEKA version 3-5-6 and has been developed under the programming environment Eclipse 3.2 and Windows XP SP2. Unfortunately, a downward compatibility can not be guaranteed. In order to avoid compile or runtime errors, the Java SE 5.0 should be used at least.

5.2.1 MultipleSetVote

The `MultipleSetVote` meta learning algorithm combines the concept of multiple learning algorithms with the simultaneous application of multiple independent feature sets. Obviously, those feature sets contain different types of features but share the same data origin. Since the works of Kittler et al. [30] and also Flexer et al. [21] impressively show that the combination of multiple learning algorithms with a single or several independent feature sets actually has the potential to improve the classification accuracy, the aim of `MultipleSetVote` is to encapsulate both meta learning approaches. As a consequence of this generic implementation, three possible processing modes which are presented in table 5.2 can be selected.

In order to set the desired processing mode of a `MultipleSetVote` instance, following methods must be used:

[7] Sub classes refers to inner nodes of the taxonomy description.

[8] The original WEKA framework contains the class `weka.experiment.Experiment` for defining arbitrary classification experiments.

- **public void** setClassifiers(weka. classifiers . Classifier [] classifiers) must be used to select those learning algorithms which will be successively used for classification. If an array with only one instance of type `weka.classifiers.Classifier` is defined, the current processing mode will be automatically set to *single*-<*any*>. Otherwise mode *multiple*-<*any*> will be activated.

- **public void** buildClassifier(weka.core.Instances data) initiates the learning procedure with the given feature set `data`. Because of using a single feature set during learning, the current processing mode is set to <*any*>-*single*.

- **public void** buildClassifier(weka.core.Instances[] data) starts the learning procedure based on the given feature set array `data`. Thus, the current processing mode is set to <*any*>-*multiple*. The array `data` includes all independent feature sets which will be successively used during learning.

Table 5.3 lists all relevant options for modifying learning and classification procedures of `MultipleSetVote`. Actually, three different ways are available in order to adjust options for a certain instance. First, every option can be manipulated separately by calling corresponding *setter*-methods. Second, a whole selection of desired options can be constituted as an array of `java.lang.String` where every array element represents a single option. In other words, every array element contains both the option name and its value (if required). The invocation of the method

public void setOptions(java.lang.String[] options) **throws** Exception

processes such an option array. Figure 5.7 shows an example set up in array representation. Third and last, an option representation with XML can be imported by employing the method

public void setParameterFile(java.util.File parameterFile) **throws** Exception

Basically, this XML representation is very similar to the original XML support for defining arbitrary WEKA learning algorithms. The WekaDoc website [56] introduces the use and creation of such XML option files. The basic XML scheme for a valid selection of options is sketched in listing 5.1. A comparison of the three different kinds of option representation is presented in figure 5.7.

Since the accuracy and performance of almost all learning algorithms, e. g. rule-based or probabilistic learners, are strongly affected by certain properties of applied feature sets, it definitely makes sense to specifically choose learning algorithms depending on the actually used feature set. Such feature set properties are for instance the dimensionality or value ranges of included features. Additionally, the complexity of class regions has a strong impact on the performance of learning algorithms. In other words, only those models should be selected among all primarily chosen learning algorithms which achieve a classification accuracy higher than some

[9] One of the following combination rules can be chosen: AVG, PROD, MAJ, MIN, MAX, MED, SUM.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE options[
 <!ELEMENT options (option)*>
<!ATTLIST options type CDATA "classifier">
<!ATTLIST options value CDATA "">
<!ELEMENT option (#PCDATA | options)*>
<!ATTLIST option name CDATA #REQUIRED>
<!ATTLIST option type (flag | single | hyphens | quotes) "single">
]>
<options type="class" value="weka.classifiers.meta.MultipleSetVote">
    <option name="R">SUM</option>
    <option name="Pfolds">5</option>
    <option name="Sfolds">2</option>
    <option name="B" type="quotes">
        <options type="classifier" value="weka.classifiers.functions.SMO"/>
    </option>
</options>
```

Listing 5.1: Outline of a XML scheme to describe an option setting for the class `MultipleSetVote`.

Option name	Default	Option description
-E	inactive	Apply classifier selection before actual learning
-Sfolds < *num* >	3	Cross validation folds for estimating "best" classifier for every feature set
-file < *path* >	—	Absolute path to XML source which specifies classifier options.
-R < *rule*[9] >	AVG	The combination rule to use
-P	inactive	Active Precision-Boosted Combination for chosen combination rule (Class attribute must be nominal)
-threshold < *num* >	0.5	Only those classifiers will be boosted which actually deliver a classification accuracy greater than the threshold
-Pfolds < *num* >	3	Cross validation folds for estimating classifier weights
-D	inactive	If set, classifier is run in debug mode and may output additional info to the console.

Table 5.3: Available options of MulitpleSetVote

user-defined threshold respectively. In context of `MultipleSetVote` this model selection is called *primary selection* and independently performs cross-validation for estimating model accuracy. An important implication of primary selection is that if it is performed it can not be guaranteed that all learning algorithms which have been originally assigned will actually be used. Obviously, the meaning of primary selection is directly related to the processing mode *multiple-multiple*. In all other modes primary selection will be neglected.

Another additional option of `MultipleSetVote` regarding the meta learning approach of model combination is *precision boosting* introduced in [13]. The idea of precision boosting is to weight the estimations of every learning algorithm during the combination. This means that instead of considering all learning algorithm equally, some certain model bias is included during the combination. The estimation of usable model weights will be done in an independent cross validation procedure.

```
Combination rule:                          Sum of probabilities
Threshold for precision boosting:          0.5
CV folds for precision boosting:           5
CV folds for best classifier estimation:   2
Status of best classifier estimation:      active
Status of precision boosting:              active
Embedded Learning model:                   weka.classifiers.functions.SMO
```

transformed into String array representation

```
String[] options = {"-R SUM", "-threshold 0.5", "-Pfolds 5",
      "-Sfolds 2", "-E", "-P", "-B \"weka.classifiers.functions.SMO\""};
```

transformed into XML option scheme

```
<options type="class" value="weka.classifiers.meta.MultipleSetVote">
    <option name="R">SUM</option>
    <option name="threshold">0.5</option>

    <option name="Pfolds">5</option>
    <option name="Sfolds">2</option>

    <option name="E" type="flag"/>
    <option name="P" type="flag"/>

    <option name="B" type="quotes">
        <options type="classifier" value="weka.classifiers.functions.SMO"/>
    </option>
</options>
```

Figure 5.7: Exemplary option setting of `MultipleSetVote` in different representations.

5.2.2 HierarchicalClassifier

The class `HierarchicalClassifier` implements a specific realization of a hierarchical classifier which also incorporates the classification results of several *sub* learning algorithms which are based on a given hierarchical class structure. But contrary to `MultipleSetVote` in which the results of included models are aggregated simultaneously, the classification result of each sub learning algorithm will be independently evaluated against a particular hierarchical class structure, also called taxonomy. Every learning algorithm is embedded into a unique node of this taxonomy and the final classification result is obtained by a successive aggregation of all partial estimations. In other words, this aggregation can also be defined as a certain path of the taxonomy tree which starts at the root node and always ends at some particular leaf node which describes the final classification result.

Hierarchical classification is already known in machine learning theory for some time and has

Option name	Default	Option description
-file $< path >$	—	Absolute path to XML source defining the class taxonomy. This option is absolutely required.
-C	inactive	Ignore unused classes at every classifier node.
-S $< num_1 \ldots num_N >$	all	Actual indices of features which will be used during learning. The corresponding feature selection only considers attributes at given positions.
-D	inactive	If set, classifier is run in debug mode and may output additional info to the console.

Table 5.4: Available options of `HierarchicalClassifier`

been employed in various applications. One of those applications is musical genre classification as section 2.4 emphasizes. In the following, only relevant aspects concerning the design of the Java class `HierarchicalClassifier` will be pointed out.

Similar to the class `MultipleSetVote` several options are provided to modify learning and classification of the `HierarchicalClassifier`. Most importantly, a user-defined taxonomy description must be given via a corresponding option before any learning procedure can actually start. All available options are listed in table 5.4 in which a short description of every option is included as well. Since `HierarchicalClassifier` and `MultipleSetVote` share the same option assignment interface, again three ways are offered to adjust available options. Thus, options can be set either by calling the corresponding *setter*-method directly, by assigning a `String` array containing desired options or by using an option representation by XML. Review subsection 5.2.1 for further details.

In order to define a particular taxonomy, all nodes of the taxonomy tree must be described by using a specific XML representation. Again, this XML representation is derived from the original XML support of WEKA for formulating learning algorithms in a single XML file. The fundamental outline of a taxonomy description is illustrated in listing 5.2. Since meaningful taxonomy descriptions are very long, this code example only includes the root node, one inner node and a leaf node. The related taxonomy tree is visualized in figure 5.8. A valid XML taxonomy representation can be assigned to an initialized instance by two different ways. First, the *setter*-method

public void setTaxonomy(java.lang.String taxonomyFile)**throws** Exception

is called with the parameter `taxonomyFile` which contains the absolute path to the XML representation source. Second and last, instead of defining an explicit XML specification of the taxonomy within a file, all desired taxonomy attributes can be formulated by an array representation which has to be constructed in similar manner as the array representation of previously mentioned learning algorithm options. Every element of such an array includes the full description of a particular node. It should be noted that the hierarchical relation of a single node to other nodes is defined by the respective options *parent* and *child*. The *setter*-method

```
1 <?xml version="1.0" encoding="utf-8"?>
2 <!DOCTYPE options[
3  <!ELEMENT options (option)*>
4 <!ATTLIST options type CDATA "classifier">
5 <!ATTLIST options value CDATA "">
6 <!ELEMENT option (#PCDATA | options)*>
7 <!ATTLIST option name CDATA #REQUIRED>
8 <!ATTLIST option type (flag | single | hyphens | quotes) "single">
9 ]>

11 <options type="class" value="weka.classifiers.meta.HierarchicalClassifier!">
12     <!-- root node -> nodeID = 1 -->
13     <option name="node" type="quotes">
14         <options type="node"
                value="weka.classifiers.meta.HierarchicalClassifier.ClassificationNode">
15             <option name="id">1</option>
16             <option name="parent">0</option>

18             <option name="child">2</option>
19             <option name="child">3</option>
20             <option name="child">4</option>
21             <option name="child">5</option>

23             <option name="W" type="hyphens">
24                 <options type="classifier" value="weka.classifiers.functions.SMO"/>
25             </option>

27             <!-- ... and more options! -->
28         </options>
29     </option>

31     .                                   .                                   .
32     .                                   .                                   .
33     .                                   .                                   .

35     <!-- last node -> nodeID = 15 -->
36     <option name="node" type="quotes">
37         <options type="node"
                value="weka.classifiers.meta.HierarchicalClassifier.ClassificationNode">
38             <option name="id">15</option>
39             <option name="parent">5</option>

41             <option name="label">Country</option>
42             <option name="classes">country</option>
43         </options>
44     </option>
45 </options>
```

Listing 5.2: The outline of a taxonomy description for the `HierarchicalClassifier`.

public void setTaxonomy(String[] options)**throws** Exception

handles the necessary processing of the given taxonomy array representation.

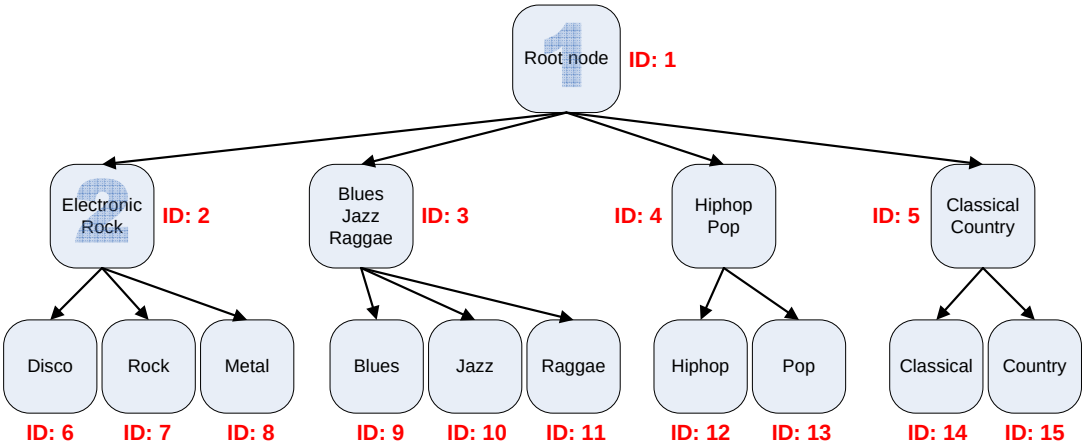
Finally, it should be noted that the `HierarchicalClassifier` learning algorithm also implements the interface `MultipleSetSelectionClassifier`. As a consequence, every feature set which will be used during learning can be additionally reduced by only taking into account the first k features only. In case of using multiple feature sets during learning, this auxiliary feature selection is applicable independently for every feature set. Thus, in order to define respective values for the upper bound k, use the *setter*-methods

public void setSelectionSize(**int** selectionSize)

in case of applying a single set or

```
public void setSelectionSizeArray(int[] selectionSize )
```

when multiple feature sets are applied.




```
<option name="node" type="quotes">
	<options type="node">
		<option name="id">1</option>
		<option name="parent">0</option>
		<option name="notes">root node of GTZAN music genre database</option>
		<option name="label">GTZAN_Root</option>
		<option name="child">2</option>
		<option name="child">3</option>
		<option name="child">4</option>
		<option name="child">5</option>
		<option name="W" type="hyphens">
			<options value="weka.classifiers.meta.MultipleSetVote">
				<option name="file">MultipleSetVoteOptions.xml</option>
			</options>
		</option>
	</options>
</option>
```

```
<option name="node" type="quotes">
	<options type="node">
		<option name="id">2</option>
		<option name="parent">1</option>
		<option name="label">Electronic/rock</option>
		<option name="classes">disco,rock,metal</option>
		<option name="notes">Electronic/Rock</option>
		<option name="child">6</option>
		<option name="child">7</option>
		<option name="child">8</option>
		<option name="W" type="hyphens">
			<options value="weka.classifiers.meta.MultipleSetVote">
				<option name="file">MultipleSetVoteOptions.xml</option>
			</options>
		</option>
	</options>
</option>
```

Figure 5.8: An exemplary musical genre taxonomy based on the GTZAN music collection. This taxonomy was first introduced in [34]. Additionally, the corresponding XML description of the root node and one successive inner node is shown below.

Chapter 6

Conclusions

This chapter summarizes the main conclusions of the discriminant analysis as well as of a feature selection approach using a feature ranking based on the genre discrimination of every feature. As being a key contribution of this thesis, the DiscriminationAnalyzer tool is also reviewed and its main functionalities are summarized. During the analysis and evaluation phase of this thesis further ideas and enhancements have arisen regarding the discriminant analysis and the evaluation of the feature selection which could not be realized in this thesis. Those ideas are also outlined as well as possible applications in terms of genre classifications for which the proposed feature selection approach could be used.

6.1 Discriminant analysis

The discrimination analysis showed clearly according to the three music collections used in this thesis that the heuristic discrimination models implementing the impurity function estimated very consistent feature patterns. In particular, this means that the same features were recognized to be discriminative, while the actual discrimination values slightly varied among the calculation models. It has been pointed out in section 3.2 that entropy-based calculation models tend to overestimate features having a large range of values. Different approaches of normalizing the estimates exist in order to reduce the distortion of such multi-valued features. Thus, it is not surprising that the discrimination values varied. Nevertheless, it could be concluded that the different approaches of normalizing have a limited influence in the calculation of discriminative features according to all three discussed music collections.

Considering the performances of the Gain Ratio, the Balanced Information Gain and the ReliefF, the computed discriminative feature patterns according to the Statistical Spectrum Descriptor were considerable more similar among the three calculation models as in the case of the other two descriptors. This conclusion was valid for both the GTZAN and the ISMIR 2004 Genre collections but not for the ISMIR 2004 Rhythm collection. Regarding the genre-to-genre comparisons according to the partially related music collections GTZAN and ISMIR 2004 Genre, it could also be concluded in the case of the Statistical Spectrum Descriptor that

the three calculation models estimated quite similar discriminative feature patterns for each of the four examined genre comparisons. In that sense the ReliefF calculation model performed better compared with the Gain Ratio and the Balanced Information Gain. According to the Rhythm Pattern descriptor and the Rhythm Histogram descriptor the discrimination results based on the three calculation models diverged more for both the GTZAN and the ISMIR 2004 Genre collection. Only for few genres the corresponding discriminative feature patterns revealed a notable degree of similarity. Also the genre-to-genre comparisons of these two partially related collections revealed diverging discriminative feature patterns. Also in the case of the Classical genre in both collections, which should correlate at most, a certain degree of similarity could not be concluded for both descriptors as only in the case of the Rhythm Pattern descriptor a considerable similarity was shown. Thus, a similarity regarding the performances of three heuristic discrimination models could not be concluded for all three descriptors. The highest degree of similarity was observed in the case of the Statistical Spectrum Descriptor but only for two of three music collections.

Another very important fact was concluded in terms of the Statistical Spectrum Descriptor. The majority of features related to the statistical measures variance and skewness appeared to be irrelevant over all three music collections. In fact, for all three music collections it was shown that only few features were estimated to be discriminative and even only for a small number of genres. A large number of features corresponding to these two measures consistently exhibited zero or very low genre discrimination.

6.2 Feature selection

The evaluation clearly showed that the feature selection performances according to the four calculation models Chi-square, Information Gain, Gain Ratio and Balanced Information Gain vary marginally only. The feature selection performance based on the ReliefF diverged more where in particular the limitation of the classification accuracy was different. Nevertheless, the results of the feature selection evaluation suggested that the classification accuracy of the selection candidates related to the same genre was strongly limited by a margin of approximately 5 % although the actual scale of limitation depended on the learning algorithm used and the respective descriptor. The limitation of the accuracy variation was always within a margin of 1 – 2 % for all music collections, learning algorithms and descriptors if the feature selection candidates containing 50 % or more of the most discriminative features were considered only. It can be followed that a reduction of 50 % according to the original feature set only slightly decreases the classification accuracy by a margin of 1 – 2 %, in some situation this decline of the accuracy may even be insignificant.

Considering the GTZAN music collection, the potential average feature set reduction according to the Rhythm Histogram descriptor was remarkably good for the two learning algorithms Naive Bayes and the J48. The average relative number of selected features was calculated by only considering those selection candidates over all one-vs.-rest genre situations for which the

best classification accuracy was achieved. More specifically, for the Naive Bayes the achieved feature set reduction was 97 % according to the Balanced Information Gain and the ReliefF. In the case of the J48 learning algorithm the achieved feature set reduction was 89.6 % and 85.4 % according to the Balanced Information Gain and the ReliefF, respectively. The largest feature set reduction according to the SMO learning algorithm was observed in terms of the descriptors Rhythm Pattern and Rhythm Histogram. In the case of the Rhythm Pattern descriptor an average reduction of 62 % and 50.1 % was achieved according to the Balanced Information Gain and the ReliefF, respectively, while an average reduction of 50.1 % and 46.8 % was observed with the Rhythm Histogram descriptor. The feature set reduction regarding the Statistical Spectrum Descriptor is 30 % according to the SMO and approximately 70 % for the other two learning algorithms. According to the GTZAN musiccollection, it could definetly be followed that the feature ranking based on the Balanced Information Gain achieved a larger the feature set reduction. It could be concluded that the Balanced Information Gain performed better with the Naive Bayes and the SMO learning algorithms, while the ReliefF appeared to be preferable with the J48 learning algorithm. Very interesting is that the performance of the SMO is only slightly better according to the Rhythm Histogram descriptor, while the margin between the accuracy achieved by the SMO and the other learning algorithms is considerably larger in the case of the other two descriptors.

In the case of the ISMIR 2004 Genre music collection, the dependency on the applied learning algorithm was more crucial in order to conclude which calculation model achieves a larger feature set reduction together with a limited decline of the classification accuracy. According to the Naive Bayes the Balanced Information Gain performed better, while the ReliefF model should be preferred for the J48 learning algorithm. According to the SMO learning algorithm, the Rhythm Histogram descriptor with the ReliefF outperformed the Balanced Information Gain with a feature set reductions of 48 % and 42.3 %, respectively, while the Balanced Information Gain was the better model for feature ranking in terms of the other two descriptors. Nevertheless, the average feature set reduction among the best classification accuracy of every genre was also remarkable in terms of the ISMIR 2004 Genre music collection. For both the Naive Bayes and the J48 Decision tree an average feature set reduction of approximately 50 % was achieved with the best classification accuracy according to all three descriptors where the average reduction was enormously high with 75 % or even more in the case of the Rhythm Histogram descriptor. For the SMO learning algorithm a feature set reduction of at least 40 % was achieved in with the descriptors Rhythm Pattern and Rhythm Histogram, while a reduction of at least 25 % was observed with the Statistical Spectrum Descriptor. It could be concluded that the Balanced Information Gain performed better with the Naive Bayes and the SMO learning algorithms, while the ReliefF appeared to be preferable with the J48 learning algorithm. Regarding the achieved classification accuracy quite the same conclusions were valid as in the case of the GTZAN music collection.

6.3 DiscriminationAnalyzer

The DiscriminationAnalyzer tool was developed within this thesis to provide specific tools to perform an interactive discriminant analysis based on arbitrary heuristic discrimination models and also arbitrary feature sets in a convenient manner. The user can choose between various calculation parameters regarding the normalization of the feature set, the class labeling and the calculation of discriminative features. Additionally, specific tools are provided in order to interactively select features and to evaluate given feature subsets by arbitrary learning algorithms. The tool already includes various heuristic discrimination models like those calculation models which were used in this thesis but new heuristic discrimination models can also be integrated by using a specific interface. The graphical user interface of the DiscriminationAnalyzer tool as well as the choice of various graphical representations regarding the discrimination results guarantee both an effective analysis and a convenient way to compare discriminative feature patterns. The DiscriminationAnalyzer tool basically runs in MATLAB, while the main part of the computations is performed by specific Java programs including the WEKA machine learning workbench.

6.4 Future Work

Throughout the phases of analyzing and evaluating the key questions of this thesis some ideas and enhancements arose which could not be realized because respective implementations would have been too time-consuming. The most relevant ideas regarding the discriminant analysis and the feature selection approach as well as two possible applications of the proposed feature selection approach are described in this final section.

The discriminant analysis introduced in chapter 3 was performed on three music collections and five heuristic discrimination models. In fact, four of the five calculation models implement the impurity function and therefore the discriminative feature patterns computed by those calculation models were very similar. During the analysis the idea came up to use another calculation model which utilizes a different approach of estimating the dependency of a specific feature to genres than the five models used in the thesis. This specific heuristic discrimination model is called Attribute Discrimination and has already been introduced in chapter 2. The key concept of the Attribute Discrimination model is described in [11, 18]. It would be an interesting questing whether this calculation model also estimates diverging discriminative feature patterns and how it affects the feature selection if the feature ranking differs compared with those based on the other calculation models. Another idea was to use binary genre situations during the discriminant analysis in addition to the employed one-vs.-rest genre situations. With using binary genre situations features related to a specific genre might be more emphasized as the discriminant analysis is focused on specific genre pairs only. This can introduce a better insight to the question which feature patterns are particularly discriminative for a specific genre. Eventually, a problem of the discriminant analysis according to all three descriptors was that the genres

of the two collections GTZAN and ISMIR 2004 Genre were only partially correlated, while the genres of the ISMIR 2004 Rhythm collection were completely different. Thus, the use of music collections containing more correlated or even similar genres would guarantee clearer comparisons regarding the performances of the heuristic discrimination models as possible influences based on only partially correlated genres would be eliminated.

During the evaluation of the feature selection approach in chapter 4 also some interesting enhancements were not integrated. One of these enhancements is an extensive correlation analysis of the feature selection candidates representing feature subsets containing the most k discriminative features. In fact, the introduced evaluation results of the feature selection suggested that a considerably large number of selected features might be correlated to each other. This assumption was supported by the fact that the variation of the classification accuracy was very limited among the feature selection candidates. From the view point of feature selection it would certainly be worth knowing whether such a correlation between the selected features exist and which features are correlated to each other. Another interesting enhancement would be to additionally perform a feature subset evaluation instead of using feature ranking. In that sense it would be worth knowing how the feature selection performance differs if the feature selection candidates would be assembled by evaluating proper feature subsets of size k which do not necessarily need to contain the k most discriminative features. Such feature subsets could be found by a greedy approach searching through all discriminative features to assemble a proper feature subset. Additionally, a certain threshold could be used to further limit the feature space being browsed. Eventually, the last enhancement concerns the learning algorithms used during the evaluation. Chapter 2 pointed out that the Gaussian Mixture Models and k-Nearest Neighbor learning algorithms are often used in genre classification. Thus, it would also be interesting to know how the feature ranking approach based on heuristic discrimination models works with these learning algorithms because the choice of the learning algorithm certainly affects the overall classification performance according to machine learning theory.

Finally, two specific applications regarding genre classification should be mentioned which can directly be combined with the introduced feature ranking approach. The first application concerns ensemble learners which use multiple learning algorithms to solve the classification problem by splitting it into several smaller subproblems. Thus, every learning algorithm solves a subproblem and the partial results of every learning algorithm must be aggregated by a certain method. Such ensemble learners were discussed in chapter 2. Since the computation of discriminative features can easily be adapted to specific genre situations, a feature ranking based on heuristic discrimination models can be used to select features according to specific subproblems. Thus, every learning algorithm could be combined with an individually adapted feature selection. This idea can also work within a hierarchical classification approach representing the second application. In fact, ensemble and hierarchical learning have in common that the original multi-class problem is partitioned into smaller subproblems related to a reduced number of classes. However, the main difference is that hierarchical learning utilizes a specific taxonomy to combine the classification results based on the subproblems. Nevertheless, an individually

adapted feature selection may also improve hierarchical learning algorithms which have been discussed in terms of genre classification in chapter 2.

Appendix A

Mathematical notation

The following table lists the mathematical notation which was used throughout this thesis. Since the key goal of this notation is to guarantee consistent formulations of the heuristic discrimination models in particular, it was sometimes necessary to deviate from some of the conventions used in the corresponding research literature.

Vectors are denoted by lower case bold Roman letters such as $\mathbf{x}$, and all vectors are assumed to be row vectors. For the convenient use of vectors, the notation $\mathbf{x}^i$ additionally indicates the i^{th} element of the vector $\mathbf{x}$. Uppercase bold Roman letters, such as $\mathbf{G}$, denote matrices. The notation $(g_1, \cdots, g_m)$ represents a row vector with m elements, while $(\mathbf{g}_1, \cdots, \mathbf{g}_m)$ denotes a matrix with m columns. A closed interval between the boundaries a and b is defined by $[a, b]$, and $\{a_1, \cdots, a_m\}$ denotes a set of m elements. Particular sets are usually named with Roman capitals, but in some specific cases calligraphic capitals such as $\mathcal{A}$ or $\mathcal{C}$ are used to emphasize the meaning of those sets. Functions are constituted by Roman or Greek letters such as $f(\cdot)$ or $\gamma(\cdot)$. The density function of a random variable X is denoted by $P(X)$, while the elementary probability of a particular observation x of X is given by $p(x)$. In terms of the joint probability according to for random variables X and Y, $P(X, Y)$ and $p(x, y)$ are used respectively.

Eventually, the following table summarizes important functions and some specific sets and vectors which are frequently used in sections 3.2.1 – 3.2.5 to define the five heuristic discrimination models employed in this thesis. The operation $A \setminus B$ denotes the usual set difference of the arbitrary sets A and B defined by $A \setminus B = \{x : x \in A \text{ and } x \notin B\}$.

$\mathcal{D}$	The data model of an arbitrary classification problem which is described by the set $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$ with the size m. $\mathbf{X}$ and $\mathbf{Y}$ are the sets of data instances and class targets respectively.

$\mathcal{A}$	The set of all attributes $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n\}$ according to $\mathcal{D}$.
$\mathcal{C}$	The set of all unique class labels according to the underlying classification problem.
$\mathbf{X}$	The matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ of m data instances with n feature values is defined by $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m)$.
$\mathbf{x}$	Denotes vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ representing a single data instances of $\mathbf{X}$.
$\mathbf{Y}$	This vector constitutes the target vector describing the class assignments of every data instance by a specific target value and is defined as $\mathbf{Y} = (y_1, y_2, \cdots, y_m)$ for m instances according to $\mathcal{D}$.
$\mathbf{a}$	Constitutes a specific attribute $\mathbf{a_j} \in \mathcal{A}$ and denotes the vector $\mathbf{a_j} = (\mathbf{x}_1^{j)}, \mathbf{x}_2^j, \cdots, \mathbf{x}_m^j)$ where $1 \leq j \leq n$ denotes the unique index of the attribute in $\mathcal{A}$. Thus, this vector contains the values of a specific attribute for all instances in $\mathcal{D}$.
$\bar{\mathbf{a}}$	Denotes the set of all attributes defined in $\mathcal{A}$ except the specific attribute $\mathbf{a}$ by $\bar{\mathbf{a}} = \mathcal{A} \setminus \mathbf{a}$.
c	Denotes a specific class label of the set $\mathcal{C}$.
$\bar{c}$	Denotes the set of all class labels contained in $\mathcal{C}$ but the specific class label c $in\mathcal{C}$.
$\eta(\cdot)$	The function $\eta : \mathbb{R}^m \mapsto \{1, \cdots, \lvert\mathcal{A}\rvert\}$ returns the unique index of the feature $\mathbf{a} \in \mathbb{R}^m$ with respect to the feature set.
$\gamma(\cdot)$	The function $\gamma : \mathbb{R} \mapsto \mathcal{C}$ returns a unique class label which is directly related to the given target value $y \in \mathbf{Y}$.
$H(\mathbf{a})$	The entropy of a specific feature $\mathbf{a} \in \mathcal{A}$.
$H(c)$	The joint entropy of a specific class $c \in \mathcal{C}$.
$H(\mathbf{a}, c)$	The joint entropy of a specific feature $\mathbf{a} \in \mathcal{A}$ and class $c \in \mathcal{C}$.
$f(\mathbf{a})$	Returns the discrimination value of a given feature $\mathbf{a} \in \mathcal{A}$ based on a particular heuristic discrimination model.

Bibliography

[1] Hervé Abdi. Kendall rank correlation. In Neil J. Salkind, editor, *Encyclopedia of Measurement and Statistics*, pages 508–510, Thousand Oaks (CA), US, 2006. Sage Publications, Inc.

[2] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2001.

[3] Jean-Julien Aucouturier and François Pachet. Representing Musical Genre: A State of the Art. *Journal of New Music Research*, 32(1):83–93, 2003.

[4] BallroomDancers.com. http://www.ballroomdancers.com/Music/style.asp, February 2008.

[5] James Bergstra, Norman Casagrande, Dumitru Erhan, Douglas Eck, and Balázs Kégl. Aggregate Features and AdaBoost for Music Classification. Machine Learning, 2006.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[7] Stefan Brecheisen, Hans-Peter Kriegel, Peter Kunath, and Alexey Pryakhin. Hierarchical Genre Classification for Large Music Collections. In *Proceedings of IEEE International Conference on Multimedia and Expo*, volume 5, pages 1385–1388. IEEE, July 9–12 2006.

[8] Leo Breiman, Jerome Friedman, R.A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth Inc, Belmont, California, US, 1984.

[9] Juan José Burred and Alexander Lerch. A Hierarchical Approach To Automatic Musical Genre Classification. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 8–11 2003.

[10] Donald Byrd and Tim Crawford. Problems of music information retrieval in the real world. *Information Processing and Management*, 38(2):249–272, 2002.

[11] Fazli Can and Esen A. Ozkarahan. Computation of term/document discrimination values by use of the cover coefficient. *Journal of the American Society for Information Science*, 38(3):171–183, 1987.

[12] Christopher DeCoro, Zafer Barutcuoglu, and Rebecca Fiebrink. Bayesian Aggregation for Hierarchical Genre Classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 07)*, Vienna, Austria, September 2007.

[13] Da Deng and Jianhua Zhang. Combining Multiple Precision-Boosted Classifiers for Indoor-Outdoor Scene Classification. In *International Conference on Information Technology and Applications*, pages 720–725. IEEE Computer Society, 2005.

[14] Thomas G. Dietterich. Ensemble Methods in Machine Learning. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2000.

[15] Simon Dixon, Elias Pampalk, and Gerhard Widmer. Classification of dance music by periodicity patterns. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 03)*, 2003.

[16] J. Stephen Downie. Music Information Retrieval. *Annual Review of Information Science and Technology (ARIST)*, 37:295–340, 2003.

[17] J. Stephen Downie et al. MIREX 2007. http://www.music-ir.org/evaluation/, 2007.

[18] David Dubin. Further Cautions for the Calculation of Discrimination Values. Technical Report UIUCLIS--1999/3+IRG, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL, 1998.

[19] Rebecca Fiebrink and Ichiro Fujinaga. Feature selection pitfalls and music classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 06)*, Victoria, Canada, October 8–12 2006.

[20] Rebecca Fiebrink, Cory McKay, and Ichiro Fujinaga. Combining D2K and JGAP for Efficient Feature Weighting For Classification Tasks in Music Information Retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 05)*, London, UK, September 11–15 2005.

[21] Arthur Flexer, Fabien Gouyon, Simon Dixon, and Gerhard Widmer. Probabilistic combination of features for music classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 06)*, Victoria, Canada, October 8–12 2006.

[22] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, pages 196–204, London, UK, June 17–29 2004.

[23] Marco Grimaldi, Padraig Cunningham, and Anil Kokaram. An Evaluation of Alternative Feature Selection Strategies and Ensemble Techniques for Classifying Music. Technical report, The University of Dublin, Trinity College, Department of Computer Science, May 30 2003.

[24] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[25] Mark A. Hall and Geoffrey Holmes. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447, 2003.

[26] Earl. B. Hunt and Philip J. Stone. *Experiments in Induction*. Academic Press, New York, US, 1966.

[27] The International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) Project. http://www.music-ir.org/mirex2007/index.php/Main_Page, February 2008.

[28] ISMIR 2004 Audio Description Contest. http://ismir2004.ismir.net/ISMIR_Contest.html, February 2008.

[29] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *ML92: Proceedings of the ninth international workshop on Machine learning*, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[30] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(3):226–239, 1998.

[31] Ron Kohavi and George H. John. The Wrapper Approach. In Liu Huan and Hiroshi Motoda, editors, *Feature Selection for Knowledge Discovery and Data Mining*, volume 454 of *The Springer International Series in Engineering and Computer Science*, pages 33–50, 1998.

[32] Igor Kononenko. On Biases in Estimating Multi-Valued Attributes. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI 95)*, pages 1034–1040, 1995.

[33] Igor Kononenko and Edvard Simec. Induction of decision trees using RELIEFF. In G. Della Riccia, R. Kruse, and R. Viertl, editors, *Mathematical and Statistical Methods in Artificial Intelligence: Proceedings of the Issek94 Workshop*, pages 199–220, Vienna, Austria, July 1995. Springer-Verlag.

[34] Tao Li and Mitsunori Ogihara. Music genre classification with taxonomy. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 05)*, volume 5, pages 197–200. IEEE, 2005.

[35] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 03)*, pages 282–289, New York, NY, USA, 2003. ACM.

[36] Tao Li and George Tzanetakis. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, chapter Factors in automatic musical genre classification of audio signals, pages 143–146. IEEE, October 2003.

[37] Thomas Lidy, Georg Pölzlbauer, and Andreas Rauber. Sound re-synthesis from rhythm pattern features - audible insight into a music feature extraction process. In *Proceedings of the International Computer Music Conference (ICMC 05)*, Barcelona, Spain, September 5–9 2005.

[38] Thomas Lidy and Andreas Rauber. Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 05)*, pages 34–41, 2005.

[39] Cory McKay and Ichiro Fujinaga. Automatic Genre Classification Using Large High-Level Musical Feature Sets. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 04)*, 2004.

[40] Cory McKay and Ichiro Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved? In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 06)*, pages 101–106, 2006.

[41] François Pachet and Daniel Cazaly. A taxonomy of musical genres. In *Proceedings of Content-Based Multimedia Information Access Conference (RIAO 00)*, Paris, France, April 2000.

[42] Elias Pampalk. Islands of Music: Analysis, Organization, and Visualization of Music Archives. Master's thesis, Vienna University of Technology, Vienna, Austria, December 2001.

[43] John C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Alex J Smola, Peter L. Bartlett, Bernhard Schöelkopf, and Dale Schuurmans, editors, *Advances in large margin classifiers*, pages 61–74. The MIT Press, 2000.

[44] PlaySOM - Intuitive access to Music Archives. http://www.ifs.tuwien.ac.at/mir/playsom.html, February 2008.

[45] J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986.

[46] Andreas Rauber and Markus Frühwirth. Automatically analyzing and organizing music archives. In *In Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries (ECDL 01)*, Darmstadt, Germany, September 4–8 2001.

[47] Andreas Rauber, Elias Pampalk, and Dieter Merkl. Using psychoacoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles. In *In Pro-*

ceedings of the International Conference on Music Information Retrieval (ISMIR 02), pages 71–80, Paris, France, October 13–17 2002.

[48] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, 53(1–2):23–69, 2003.

[49] Specification of SOMLib Data Files. http://www.ifs.tuwien.ac.at/~andi/somlib/download/SOMLib_Datafiles.html, February 2008.

[50] David M. J. Tax, Martijn van Breukelen, Robert P. W. Duin, and Josef Kittler. Combining multiple classifiers by averaging or by multiplying. *Pattern Recognition*, 33(9):1475–1485, 2000.

[51] Rainer Typke. A survey of Music Information Retrieval Systems. http://mirsystems.info/index.php?id=mirsystems, November 28 2007.

[52] Rainer Typke, Frans Wiering, and Remco C. Veltkamp. A survey of Music Information Retrieval Systems. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 05)*, pages 153–160, 2005.

[53] George Tzanetakis. *Manipulation, analysis and retrieval systems for audio signals.* PhD thesis, Princeton University, Princeton, NJ, USA, 2002.

[54] George Tzanetakis and Perry R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

[55] University of Waikato. Website of the machine learning workbench WEKA. http://www.cs.waikato.ac.nz/ml/weka/, December 18 2007.

[56] WekaDoc the Documentation project for WEKA. Xml support in weka. http://weka.sourceforge.net/wekadoc/index.php/en:XML_%283.4.6%29, December 18 2007.

[57] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Information Science and Statistics. Morgan Kaufmann, San Francisco, US, 2^{nd} edition, 2005.

[58] Yimin Wu and Aidong Zhang. Feature selection for classifying high-dimensional numerical data. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 2, pages 251–258, 27 June–2 July 2004.

[59] Changsheng Xu, Namunu C. Maddage, Xi Shao, Fang Cao, and Qi Tian. Musical genre classification using support vector machines. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 03)*, volume 5, pages 429–32. IEEE, April 6–10 2004.