GENERATING SUMMARIES FROM RETRIEVED BASE CASES

A. Schuster, K. Adamson, D.A. Bell

University of Ulster at Jordanstown, Faculty of Informatics, School of Information and Software Engineering, Shore Road, Newtownabbey, Co. Antrim, BT37 0QB, Northern Ireland

{a.schuster, k.adamson, da.bell}@ulst.ac.uk

Abstract

Despite ongoing efforts a lack of efficient case evaluation remains a central problem within the relatively young research area of case-based reasoning (CBR). In simple terms case evaluation is the task of identifying the problem-solving potential retrieved cases provide for a given situation. Very frequently this process includes the identification and generation of summaries for all, or some retrieved candidate cases. In many applications case evaluation is done by the system user. This paper addresses the problem by presenting a general method for case evaluation. The method generates expressive summaries for retrieved cases based on the theory of fuzziness. Results from an application in the domain of Coronary Heart Disease Risk Assessment (CHDRA) indicate the value of the method for case evaluation. The paper also identifies the potential of the method for other CBR issues like similarity assessment and case indexing.

1. Introduction

A very common problem-solving strategy for humans is to remember the knowledge and the experience they have gathered in similar, past situations, and to apply that knowledge to solve the current problem. In some situations a solution to the problem at hand might be derived from only a single past situation, in others it might be extracted from more than one past event. CBR is an artificial intelligence problem-solving technique that follows the same route. [Aamodt and Plaza, 1994], [Kolodner, 1993]. In CBR, past experience—i.e., knowledge about situations that have been solved in the past—is represented by entities called *cases*. These cases are stored and organised in a memory-like construct called a *case knowledge base* or simply *case base*. Reasoning in CBR systems is accomplished by retrieving the base case(s) most relevant to a new situation or problem at hand, called the *query case*. A case-based reasoner relies on the assumption that a solution to the query case might be derived from the base cases retrieved. The problem-solving contribution of these cases then has to be determined. In CBR terminology this process is known as *case evaluation*. However, case evaluation is not a trivial affair at all. Actually, in most CBR applications the system user plays the dominant part in this process, and thus case evaluation remains one of the central CBR issues [Petersen, 1997], [Leake and Barletta, 1997]. Figure 1 for example, illustrates a CBR scenario where a system user is confronted with the task of finding a solution for a query case.

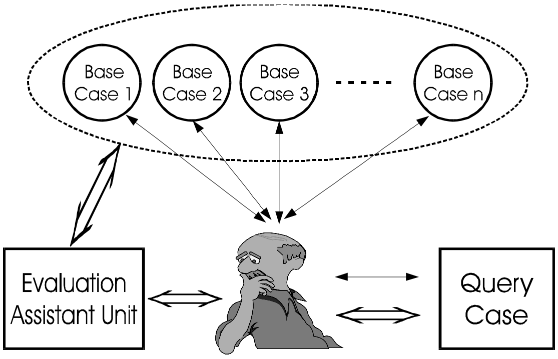

Figure 1: System user evaluating retrieved cases.

Suppose that for the Query Case in Figure 1 a certain number of base cases (Base Case 1 to Base Case n) has been retrieved from the case base. The system users job then is to evaluate the problem-solving potential of these cases and to apply (if necessary adapt) the solutions provided by these cases to the query case. Here, it is important to mention that a solution might be derived from: (a) a single retrieved case, (b) all retrieved cases, or (c) a sub-group of the cases recalled. Since cases are described by so-called *features*, the evaluation process is based on an feature to feature comparison between the query case and **each** retrieved base case. The single-lined arrows in Figure 1 indicate this strategy. Consequently, the larger the number of retrieved cases and the larger the number of features per case, the more complex and hence difficult the case evaluation process for the system user.

Alternatively, it would be very useful to have a so-called *Evaluation Assistant Unit* (EAU) available (Figure 1). The purpose of the EAU would be to identify prominent characteristics of retrieved base cases and to make this information accessible for the system user (double-lined arrows in Figure 1). Given that an EAU of this kind exists the system user's first action would be to check the output of the EAU, and then analyse the problem-solving potential of the retrieved cases according to the suggestions given by the EAU. So, EAU related requirements might be as follows:

- At root the EAU should make the process of case evaluation more efficient. Results generated by the unit should be intuitively appealing to an expert's understanding of the problem in question.
- Case characteristics can very frequently be expressed by vague or imprecise, but nevertheless highly expressive summaries. For example, in the CHDRA domain a system user may recognise that all retrieved cases show *very high blood cholesterol levels*. The EAU would therefore then have to address uncertainty issues.

The purpose of this paper is to propose a method that allows the identification of main characteristics (e.g. *very high* blood pressure) shared by retrieved base cases. The method aims to assist a CBR system user within the process of case evaluation. Specifically, in many situations the information that might be useful for problem-solving can be articulated by a system user via expressive summaries. These summaries often show a certain degree of vagueness or imprecision, and so the proposed method should be able to manage uncertainty [Zadeh, 1973], [Klir and Folger, 1988]. We have therefore used the Fuzzy-C-Means clustering technique to generate (initial) summaries for 'individual' base cases. We have also developed an algorithm that uses the results of the Fuzzy-C-Means to establish meaningful summaries for a 'varying number' of retrieved base cases. The applicability and usefulness of the approach was tested in a generic CBR application in the CHDRA domain to see if our approach has the potential to substantially support case evaluation.

The remainder of this paper is organised as follows: Section 2 identifies the value of abstract case summaries in CBR. The generation of such summaries for a single case, as well as for multiple cases is the content of Section 3. In Section 4 we utilise our approach in a CBR application in the CHDRA domain. Section 5 reviews related work. Finally, Section 6 ends with a discussion, conclusions, and future work.

2. Generation of a Case Summary

Many decision-making situations require the ability to aggregate information into expressive summaries. Even if the meaning of a summary can be represented in terms of simpler pieces of information the need for such higher level entities exists [Wilensky, 1986]. Further, although expressive summaries usually show a degree of uncertainty (e.g., in the form of imprecision or vagueness) in many situations they are sufficient for further information processing [Schuster *et al.*, 1997].

For example, in the CHDRA domain cholesterol has been identified (among other factors) to be a main risk factor for myocardial infarction and subsequent sudden death [Levy, 1993]. Cholesterol travels in the blood in distinct particles called lipoprotein. The two major types of lipoproteins are low-density lipoproteins (LDL) and high-density lipoproteins (HDL). LDL, often called 'bad' cholesterol, delivers the cholesterol to the arterial walls with the ultimate consequence of narrowing the arteries. HDL, often called 'good' cholesterol, protects against heart disease by removing excess cholesterol from the blood. In a fasting blood test a clinician first finds out what a subject's TOTAL cholesterol level is. If the TOTAL cholesterol level is too high then further measurements of LDL and HDL are required. The two ratios TOTAL/HDL and LDL/HDL are also important because they provide more meaningful indicators of coronary heart disease (CHD) risk than TOTAL cholesterol per se. However, having the five values: TOTAL = 5.30 mmoll^{-1}, LDL = 3.82 mmoll^{-1}, HDL = 0.63 mmoll^{-1}, TOTAL/HDL = 8.41, and LDL/HDL = 6.06 in front of him for example, a clinician might say simply that a subject's CHOLESTEROL in terms of CHD risk is *normal*. (Note: in this paper we use 'cholesterol' in general discussions, and CHOLESTEROL when we talk about an abstract summary, aggregated or composed of different cholesterol types and cholesterol type ratios.) Subsequent decision-making can then be undertaken without any reference to the underlying numerical data by using only such abstract summaries. For example, a clinician might say: "Because your CHOLESTEROL is *rather high* we suggest the following diet. …". From the point of view of CBR such abstract summaries can be used as highly expressive case descriptors allowing fast and efficient case evaluation. The strategy in the following sections is therefore as follows:

(1) To address the uncertainty issue mentioned above a clustering technique is first selected that allows the generation of categories with relaxed boundaries. Such relaxed categories are meaningful, because there is no exact boundary between, for example, *normal* and *abnormal* CHOLESTEROL, and therefore the transition between two categories should be gradual or fuzzy rather than abrupt or crisp.

(2) Once the clustering technique is selected, its value is tested on single cases.

(3) The approach is then extended to generate meaningful summaries for more than one retrieved case.

3. Generation of Case Summaries via the Fuzzy-C-Means

The aim of any clustering technique is to find structures contained within data. These structures are usually classes or categories. Once the classes are established they are used as a container for objects described by the data. Classical clustering techniques assign an object to exactly one class. In many situations this is an oversimplification, because very often objects can be partially assigned into two or more classes. CHOLESTEROL assignment is a typical example. The Fuzzy-C-Means clustering algorithm is based on this idea. A detailed explanation of the theoretical foundations of the algorithm is beyond the scope of this paper. It is however important to understand the main characteristics of the algorithm.

At core the Fuzzy-C-Means is a result of an attempt to come to grips with the problem of pattern recognition in the context of imprecisely defined categories [Bezdek, 1981]. A main feature of the Fuzzy-C-Means is that for a single object (x) the algorithm assigns a membership degree ($\mu(x)$) to every single class (C_i). Membership degrees are drawn from the interval [0, 1], thus $\mu_{Ci}(x) \in [0, 1]$. Further, for a single object the membership degrees assigned to all classes sum to one, $\Sigma\mu_{Ci}(x) = 1$. At the beginning the membership degrees are seeded randomly. The algorithm then iteratively determines new cluster centers, and usually terminates when a predefined threshold is reached.

3.1 Summary for a Single Case

We present the process first for one case, using our CHOLESTEROL study for illustration. The base for the study is a data set consisting of 166 records. The data was collected in a wider study in the CHDRA domain [Lopes *et al*, 1994]. Very basically a record holds the personal and medical data of a subject, including the values for TOTAL, LDL, and HDL cholesterol, as well as the two ratios TOTAL/HDL and LDL/HDL. Since our study focuses on determining the (summary) CHOLESTEROL of a subject, in a first step a domain expert was asked to provide expertise on the CHOLESTEROL of each subject. The expert was asked to indicate one of the fields (*normal*, *borderline*, or *abnormal*) for each data record (Table 1).

Table 1: Domain expert's assignment on 166 data records.

CHOLESTEROL / Expert's Decision		
normal / 79	*borderline* / 61	*abnormal* / 26

For example, according to the human expert there are 79 data records whose CHOLESTEROL is *normal* with respect to CHD. The numbers for *borderline* and *abnormal* CHOLESTEROL records are 61 and 26, respectively. It is important to emphasise that the domain expert was asked to classify each record. For quite a few records it was difficult for the domain expert to come up with a confident assignment, because of his opinion that a record belongs to the boundary region between two categories. In these situations the expert was more or less forced to choose one of the categories in question. The same 166 records were classified by the Fuzzy-C-Means technique. Table 2 illustrates three typical Fuzzy-C-Means classification outcomes.

Table 2: Fuzzy-C-Means classification for three data records.

	Fuzzy-C-Means			
Record	***normal***	***borderline***	***abnormal***	Σ
1	**0.88**	0.10	0.02	1.00
2	0.01	**0.97**	0.02	1.00
3	0.04	0.16	**0.80**	1.00

According to the maximum degrees computed for a class the Fuzzy-C-Means categorises the first record in Table 2 to be *normal* (0.88), the second record to be *borderline* (0.97), and the last record to be *abnormal* (0.80) with respect to CHD risk. (Note: the degrees in any row sum up to one.) Provided with the results of the algorithm and the assignment of the domain expert it is possible to determine the total number of correct classifications established by the Fuzzy-C-Means (Table 3).

Table 3: Comparing the Fuzzy-C-Means with the domain expert's opinion on the CHOLESTEROL of 166 subjects.

CHOLESTEROL	**Expert**	**Fuzzy-C-Means**
normal	79	60 = 75.9%
borderline	61	44 = 72.1%
abnormal	26	21 = 80.0%
Σ	166	125 = 75.3%

For example, for the 79 *normal* CHOLESTEROL records identified by the human expert the Fuzzy-C-Means provides 60 matching outcomes, which is equivalent to 75.9%. The results for the classes *borderline* and *abnormal* are 44/72.1% and 21/80.0%, respectively. The last row in Table 3 holds the combined results for all three categories, and indicates a total of 125/75.3% correct classifications by the algorithm. We suggest that these 75.3% should be regarded as a lower bound. It was pointed out earlier that for some records a confident assignment by the domain expert was difficult, because the available data indicated these records as being located in the boundary region between two categories. Table 4 illustrates two such records.

Table 4: Two data records allocated to boundary regions.

Record	**Expert**	**Fuzzy-C-Means**		
		normal	***borderline***	***abnormal***
1	*normal*	0.42	**0.45**	0.13
2	*borderline*	0.37	**0.39**	0.24

For example, the human expert and the Fuzzy-C-Means disagree on the assignment of the first data record in Table 4. The expert's assignment is *normal*, whereas the Fuzzy-C-Means categorises the record to be *borderline* (0.45). A closer look however reveals that the Fuzzy-C-Means computes a similarly high degree for the class *normal* (0.42), and therefore identifies this record as a boundary record rather than as a prototypical member of one of these two classes. On the other hand, the expert and the algorithm agree in their assignment on the second record (expert = *borderline*, Fuzzy-C-Means = *borderline*/0.39). Nevertheless, the membership degree determined for the 'second best' class (*normal*/0.37) is very close to the membership degree of the 'winning' class (*borderline*/0.39), which identifies this record like the record before to be a boundary record falling into the region between these two classes. An investigation on all 166 records, that is correctly as well as incorrectly classified records, identified the described phenomenon of a relatively small class membership difference for quite a few records. This strengthens the view that the 125/75.3% correct classifications in Table 3 should be regarded as a lower bound. There is however a restriction to the interpretation of a record. A record can belong to only one boundary region. We therefore consider only the two categories with the highest membership degrees for subsequent processing. For example, although the second record in Table 4 shows some degree for the class *abnormal* (0.24), only the degrees *borderline*/0.39 and *normal*/0.37 are used for an interpretation. For the current record an interpretation would be: "The CHOLESTEROL of the subject is located in the boundary region between *normal* and *borderline*". Note that the mere assignment *borderline* by the domain expert in Table 4 fails to provide such important information. The Fuzzy-C-Means is thus able to identify and to manage the task of uncertain or difficult class assignment. It can be considered as an alternative for information systems showing a need for such support. From the point of view of CBR the advantages can be summarised as follows:

- It is possible to aggregate individual case features (e.g., TOTAL, LDL, etc.) via the Fuzzy-C-Means to derive expressive summaries (e.g., CHOLESTEROL).
- The Fuzzy-C-Means addresses uncertainty issues. Membership degrees identify cases where a clear category assignment is difficult.
- Case evaluation is commonly undertaken by the system user via a sequential feature to feature comparison between the query case and the returned base cases. Summaries reduce the number of features included within the comparison, leading to a faster and more efficient case evaluation.

3.2 Summary for Multiple Cases

The above arguments indicate the value of the method for a single retrieved base case. The following section extends the approach by introducing an algorithm that allows the generation of summaries for multiple retrieved base cases. For example, consider a retrieval scenario where six base cases have been retrieved for a query case. Table 5 illustrates the CHOLESTEROL values of these cases.

Table 5: CHOLESTEROL values of six retrieved base cases.

Retrieved base cases	CHOLESTEROL *normal*	*borderline*	*abnormal*
1	0.399	0.491	~~0.110~~
2	~~0.092~~	0.217	0.691
3	~~0.129~~	0.632	0.239
4	~~0.019~~	0.048	0.933
5	0.090	0.836	~~0.074~~
6	~~0.073~~	0.289	0.638
$M_{Category}$	0.489	2.513	2.501
normalisation	0.195	1.000	0.995

Table 5 indicates that only the two highest membership degrees are considered for further processing by striking through the lowest class assignment within a row. For example, for base case No.1 in Table 5 only the category assignments *normal*/0.399 and *borderline*/0.491 are selected. A summary for the CHOLESTEROL of all six cases in Table 5 is generated in two steps. Step 1 establishes the total membership degree ($M_{Category}$) for a category via Eq. 1:

$$M_{Category} = \sum_{i=1}^{n} \mu_{\,Category}(Case_I), \qquad \text{Eq. 1}$$

where $\mu_{Category}(Case_i)$ is the CHOLESTEROL membership degree of a retrieved base case within one of the categories *normal*, *borderline* or *abnormal*. For example, for the six cases in Table 5 Eq. 1 establishes a total membership degree for the category *normal* of: $M_{normal} = 0.399 + 0.090 = 0.489$. $M_{borderline}$ and $M_{abnormal}$ sum up to 2.513 and 2.501, respectively. Step 2 produces the final outcome by simply normalising the Step 1 results (last row in Table 5). So, a fast interpretation of the CHOLESTEROL of the six retrieved cases based on the normalised Table 5 results would be: "The CHOLESTEROL of most of the retrieved cases is *borderline* or *abnormal*". This can be easily verified by looking at the retrieved cases individually. Note that such an interpretation is faster and more efficient than an interpretation that is based on a comparison of the individual TOTAL, LDL, HDL, TOTAL/HDL, and LDL/HDL values of the six cases in Table 5.

The EVALUATER System

To test the value of the proposed method we have developed the EVALUATER system (Figure 2).

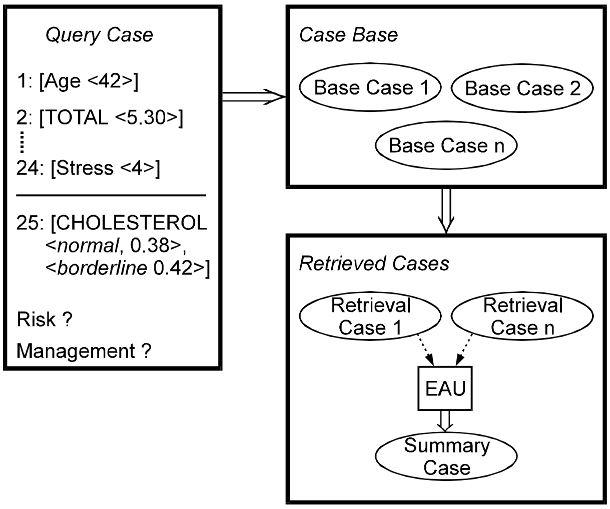

Figure 2: The simplified EVALUATOR system.

EVALUATER uses data that is derived from a study of 83 middle-aged men between 30 and 65 years of age who undertook standard screening tests in 1993 and again in 1996 in order to identify selected CHD risk factors [Lopes *et al.*, 1994]. A series of qualitative and quantitative information were collected including: age, height, weight, body fat percentage, personal and family medical history, smoking and nutrition habits, blood pressure, cholesterol, stress and physical activity levels. To evaluate the individual CHD risk of a subject, EVALUATER relies on a point scoring system proposed by Anderson *et al*., [Anderson *et al.*, 1991]. Individual risk values correspond to a subject's 10-year CHD risk, and range in the interval [1 ≤ risk value ≤ 32]. For example, a risk score of 29 corresponds to a predicted 10-year CHD risk of 36%.

The aim of the application is to gain information about CHD risk and management of a new patient, illustrated as a query case in Figure 2. According to the recorded data EVALUATER retrieves the most similar case(s) from the case base (Base Case 1, Base Case 2, …, Base Case n), assuming that the information available through the retrieved case(s) can be used for the query case. Case retrieval is based on 24 features, including TOTAL, LDL, and HDL cholesterol. Note that feature No.25 in Figure 2 is not included in the case retrieval process. Feature No.25 is generated by the Fuzzy-C-Means and depicts the (summary) CHOLESTEROL of a single case.

According to Figure 2 the system retrieves a certain number of base cases (Retrieval Case 1 to Retrieval Case n) for the query case. Figure 2 further illustrates that an EAU is integrated into the system. The EAU generates a *Summary Case* from the retrieved cases. The *Summary Case* includes summaries that are derived from **all** retrieved base cases within a query. For the feature CHOLESTEROL this process was explained in the previous sections. In a real scenario a system user would first check the output of the EAU, the *Summary Case*, and then analyse the problem-solving potential of the retrieved cases according to the suggestions given by the EAU. To further improve the case evaluation process the EAU also provides a graphical output (Figure 3).

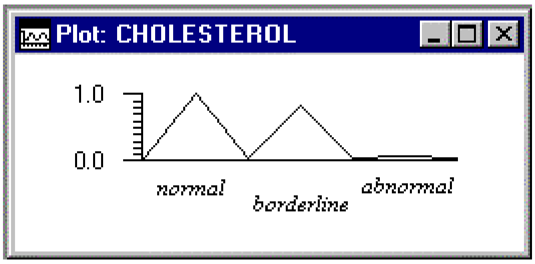

Figure 3: Graphical output of the EAU.

The graphical output presents relevant categories in triangular shape. For example, Figure 3 reveals that the CHOLESTEROL of most of the retrieved base cases is *normal* or *borderline*.

4.1 Results

Testing of the EVALUATER prototype system involves the identification of the value of the system in the domain of CHDRA. Currently we focus on the value and efficiency of the EAU. In its present state the system only provides results for the CHOLESTEROL of retrieved cases. Nevertheless, in a number of test runs it became clear that the EAU allows fast, and sufficiently accurate interpretations of the CHOLESTEROL of multiple retrieved cases, which was the aim of this research. The testing was conducted by a domain expert who was available during the course of this project. It also emerged that the method works better for a somewhat smaller number (n) of retrieved cases (n = 4 to 8). For a larger number of retrieved cases (n > 10) it was frequently the case that all categories showed similarly high total membership degrees per category. This does not necessarily indicate a non-meaningful result. It might be due to the limited present state of the system, because the EVALUATER case base so far, exists of only 83 cases.

5 Related Work

In a similar context Schuster *et al*, use a rule based fuzzy expert system to generate summaries for complex case features. The value of these complex features was then tested in different case retrieval scenarios [Schuster *et al.*, 1997]. The generation of summaries via a fuzzy expert system is a knowledge intensive approach, because it involves the formulation and representation of detailed domain knowledge in the form of rules and fuzzy sets. In contrast, the knowledge used in the presented study is poorer, but it is nevertheless sufficiently expressive for the given application.

Work by Jeng and Liang, and Petersen provides further evidence for the applicability of the theory of fuzziness in CBR [Jeng and Liang, 1995], [Petersen, 1997]. Jeng and Liang's paper addresses the case indexing issue, whereas Petersen's focus is on similarity assessment between fuzzily defined data. In both these papers the use of linguistic expressions, which in our understanding are equivalent to expressive summaries, play a central role. It seems to be possible to apply the presented method to both areas. The CHOLESTEROL summary for example, can be taken as an expressive and meaningful case index. It also appears feasible to develop a similarity assessment strategy that employs the (summary) feature CHOLESTEROL and its associated membership degrees within the process of case retrieval.

6 Conclusions and Future Work

A general method to support the process of case evaluation for an arbitrary number of retrieved base cases has been presented. Based on a fuzzy clustering technique, and expressed in the form of an algorithm the proposed method was integrated in a generic

CBR application specifically developed for this research. First results achieved in a number of tests indicated the value of the method. We belief that many information systems with similar problems can benefit from the presented method.

Present and future work entails the utilisation of the approach to generate summaries for other features included in the EVALUATER system. For example, the available data contains 4 features related to stress, 4 features associated with physical exercise, as well as 2 blood pressure features. The aim would be to generate summaries for stress, physical exercise, and blood pressure, and to test the value of these summaries for case evaluation. Another route further research might take has been mentioned earlier, as it is also intended to investigate the value of generated summaries from the point of view of case indexing and similarity assessment.

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AICOM*, 7(1):39--59, March 1994.

[Anderson et al., 1991] Keaven M. Anderson, Peter W.F. Wilson, Patricia M. Odell, and William B. Kannel. An updated coronary risk profile. *Circulation*, 83(1):356--362, January 1991.

[Bezdek, 1981] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, London, 1981.

[Jeng and Liang, 1995] Bing C. Jeng and Ting-Peng Liang. Fuzzy indexing and retrieval in case-based reasoning. *Expert Systems with Applications*, 8(1):135--142, 1995.

[Klir and Folger, 1988] George J. Klir and Tina A. Folger. *Fuzzy Sets, Uncertainty and Information*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[Kolodner, 1993] Janet L. Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, California, 1993.

[Leake and Barletta, 1997] David Leake and Ralph Barletta. Case-based Reasoning Tutorial, Fifteenth International Joint Conference on Artificial Intelligence IJCAI'97, Nagoya, Japan, 1997.

[Levy, 1993] Daniel Levy. A multi-factorial approach to coronary heart disease risk assessment. *Clin. And Exper. Hypertension*, 15(6)1077—1086, 1993.

[Lopes *et al.*, 1994] Philippe L. Lopes, R.H. Mitchell and John A. White. The relationships between respiratory sinus arrhythmia and coronary heart disease risk factors in middle-aged males. *Automedica*, (16):71--76, 1994.

[Petersen, 1997] Jörg Petersen. Similarity of fuzzy data in case-based fuzzy system anaesthesia. *Fuzzy Sets and Systems*, 85:247—262, 1995.

[Schuster *et al.*, 1997] Alfons Schuster, Philippe Lopes, Kenneth Adamson, David A. Bell, John A. White. Aggregating features and matching cases on vague linguistic expressions. Proceedings of the 15th International Joint Conference on Artificial Intelligence, pages 252--257, Nagoya, Japan, 1997.

[Wilensky, 1986] Robert Wilensky. Knowledge representation-A critique and a proposal. Janet Colodner and K. Riesbeck, editors. *Experience, Memory, and Reasoning*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986

[Zadeh, 1973]. Lotfi A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics, SMC*, 3(1):28--45, January 1973.