Feature Mining and Predictive Model Construction from Severe Trauma Patient's Data

Janez Demšar [1], Blaž Zupan [1,2,3], Noriaki Aoki [3,4],
Matthew. J. Wall [5], Thomas H. Granchi [5] and J. Robert Beck [3]

[1] Faculty of Computer and Information Sciences, University of Ljubljana, Slovenia
[2] J. Stefan Institute, Ljubljana, Slovenia
[3] Office of Information Technology, Baylor College of Medicine, Houston, TX
[4] Department of General Medicine and Clinical Epidemiology, Kyoto University, Japan
[5] Ben Taub General Hospital, Houston, TX

Abstract

In management of severe trauma patients, trauma surgeons need to decide which patients are eligible for damage control. Such decision may be supported by utilizing models that predict the patient's outcome. This paper investigates the possibility to construct patient outcome prediction models from retrospective data. As the retrospective data included in this study comprises rather excessive number of features, special attention was paid to the problem of selecting only the most relevant features. We show that a small subset of features may carry enough information to construct reasonably accurate prediction models.

1 Introduction

Trauma surgeons face complex management and decision making problems when treating patients with severe traumatic injury. In the initial period of treatment, the patient's continued hemodynamic instability may increase the risk of difficulty of definitive repair of all injuries. Bold attempts to completely correct acute surgical problems, especially trauma, were explored in 1970s and 1980s. The application of extracorporeal support, extensive resections, aggressive fluid resuscitation (including blood products), and primary extensive reconstruction was reported during this era. Often, with these early excursions into complex trauma reconstruction and resection, the surgical goals were achieved at the initial operation, but the patients went on to die of respiratory failure, multiple organ distress, and coagulopathy. Even with such aggressive attempts, undesirable outcomes in terms of cost, lengthy stay in intensive care unit (ICU) or hospital, cerebral insufficiency, and death were encountered [7].

Within the last ten years, the damage control approach emerged from a need to meet the challenge of the changing scope and severity of injury. The basic concept of damage control for trauma patients is to avoid extensive procedures on unstable patients, stabilizing fatal problems at initial operation, and applying staged surgery after successful initial resuscitation. As the success of damage control has grown, so has its acceptance in the traditional surgical community.

Several problems regarding damage control remain. One of the most important but difficult issues is determining which patients are eligible for damage control. As there is no widely accepted standard, the surgeons are still looking for the systematized approach to define the eligibility of patients.

Damage control requires a massive investment of personnel, efforts, and resources in a small group of critical injured patients who carry a mortality rate in excess of 50%, even under the best circumstances [3]. From the viewpoint of resource allocation, it would be desired to be able to accurately predict patient's outcome to prevent futile use of limited resources. This paper investigates the possibility of constructing an outcome prediction model for severe trauma patients after the first surgery. The model is derived from the data that includes patient's features and the observed outcome. As the original dataset included rather excessive number of features, special attention was paid to the problem of selecting only the most relevant features. We show that a small subset of features may carry enough information to construct reasonably accurate prediction models.

2 Data

We retrospectively examined 68 patients who required damage control surgery at Trauma and Critical Care Center, Ben-Taub General Hospital in the period from 1994 to 1997. A set of 174 features including patient characteristics, features of prehospital care, and physical and laboratory findings in emergency room, operating room and intensive care unit was used in the analysis. The dataset included many missing values: preliminary dataset inspection showed that for 78 features data was missing for at least 50% of patients – these features were not included in the further analysis. The resulting dataset (68 patients, 96 features) had 20.7% of missing values.

3 Methods

A number of preprocessing, modeling and quality estimation methods were used in this work. We first describe feature mining techniques that were used to narrow a list of features used. From the resulting dataset, prediction models were derived by classification trees and naive Bayesian techniques. The quality of the models was assessed through using various criteria and statistical tests.

3.1 Feature mining

Feature (or attribute) mining is a data mining preprocessing stage where, for a classification tasks, a subset of most relevant features is identified and potentially reformulated [12]. The identification of most relevant features often refers to their ranking, subset selection and to their categorization.

In the first step of the preprocessing, we categorized (discretized) the continuous features. This was required for naive Bayesian modeling technique which does not directly handle continuous features. Besides, the mere information on how the features were categorized can be interesting for the domain expert to verify the relevance of the data base (if categorization is as expected) or to point out for new and interesting categories and cut-off points. We have used two approaches for categorization: quartiles and entropy-MDL based discretization. The quartile discretization splits the range of feature's values into four intervals, so that the number of patients within each interval is approximately equal. The more sophisticated entropy discretization [2] uses a top-down approach, similar to clustering methods. It start with an interval covering all the feature's values and finds a cut-off point which maximizes the informativity. If the gained information is greater than the increase of the minimal description length for the feature values, the interval is cut into two subintervals and the procedure is repeated on both of them. However, it often happens that the process stops at the first step already. If this is the case, there was no cut-off point and such features can be regarded as irrelevant. In this way the entropy-based discretization can also be used as a feature selection tool.

As the quartile discretization considers only the values of the feature that is being discretized independently of other features or outcomes, it tends to be more noise-proof on one side but potentially less interesting for the domain expert on the other side. Besides, the number of intervals for the quartile discretization is fixed, so it cannot be used for the feature subset selection.

After categorization, features were ranked using RELIEFF [4, 5]. RELIEFF measures usefulness of a feature by observing the relation between its value and patient's outcome. Intuitively, if there is a group of patients with the same or similar feature values, the observed feature is "valuable" as a predictor if it has different values on pairs of patients with different outcomes (thus distinguishing between them), but the same value on pairs with the same outcome. The features with negative RELIEFF estimate may be considered irrelevant. The features with the highest score are presumed to be the most useful for predicting the outcome. In our pilot study, features were ordered according to their scores and presented to the expert who performed the final selection.

3.2 Data modeling

After we have reformulated the trauma patients' descriptions by categorizing and selecting the features, we used two well-known machine learning techniques to induce the predictive models. The first one was our own implementation of *classification trees* derived from a commonly-known ID3 recursive partitioning algorithm [10]. The basic idea of ID3 is to partition the patients into ever smaller groups until creating the groups with all patients corresponding to the same class (e.g., survives, does not survive). The partition criteria is a function computed from predictor variables. To avoid overfitting, we have used a simple pruning criterion that stops the induction when the sample size for a node falls under the prescribed number of examples or when a sufficient proportion of a subgroup has the same output.

The second machine learning method used was a naive Bayesian classifier. Assuming the independence of predictive variables, the probability that a patient described with values of predictor variables $V = (v_1 ... v_n)$ survives can be estimated by *naive Bayesian formula* [6]

$$P(R|V) = P(R) \prod_{i=1}^{n} \frac{P(R|v_i)}{P(R)}$$

where $P(R)$ is the apriori probability of survival and $P(R|v_i)$ is the conditional probability of survival if i-th predictor variable has the value v_i; both are estimated from the training set of patients.

Naive Bayes and classification trees were chosen because they represent two essentially different approaches for induction of predictive models. Naive Bayesian models include all of predictive variables used in the data, while classification trees in general only use a subset of most informative features. Naive Bayesian models are in essence linear, while classification trees may represent more complex models. For modeling from medical data, however, it was observed that naive Bayes most often performs best, outscoring classification trees, rules, and even artificial neural networks [1, 6].

A baseline for comparison with above two methods was a *majority classifier* that uses a training set to determine the most frequent class and then classifies all cases from the test set to that class.

3.3 Model evaluation methods, metrics and comparison statistics

Having categorized and selected the features and induced outcome prediction models, different statistical measures can be used to evaluate the quality of derived models. From those which we used in this study, the first two (classification accuracy, sensitivity and specificity) consider the class prediction while for the other two (average probability assigned to correct class, area under ROC curve) use the model to predict the probabilities of classes.

- **Classification accuracy (CA)** measures the proportion of correctly classified test examples, therefore estimating the probability of the correct classification.

- **Sensitivity and specificity (Sens/Spec)** measure the model's ability to "recognize" the patients of a certain group. If we decide to observe the surviving patients, *sensitivity* is a probability that a patient who has survived is also classified as surviving, and *specificity* is a probability that a not-surviving patient is classified as not-surviving (or, more generally, *specificity* is the number of patients *not wrongly classified as surviving* divided by the number of patients which *could be wrongly classified as surviving*).

- **Average probability assigned to the correct class (AP)** is related to classification accuracy, but it gives an additional information on the reliability of the classifier's decisions. If this measure is low, the classifier can still have a good classification accuracy but its decisions are, on the average, marginal.

- **Area under ROC curve (aROC)** is based on a non-parametric statistical sign test and estimates a probability that for a pair of patients of which one has survived and the other has not, the surviving patient is given a greater probability of survival. This probability was estimated from the test data using relative frequency.

The above metrics and statistics were assessed through stratified *ten-fold cross validation* [8]. This divides the patient's dataset to 10 sets of approximately equal size and equal distribution of outcomes. In each experiment, a single set is used for testing the classifier that has been developed from the remaining nine sets. The statistics for each method are then assessed as an average of 10 experiments. The same training and testing datasets were used for all classification methods.

The described statistics measure the quality of a single classifier. Although they can be used to compare them, a better and more statistically correct test is available for this purpose. *McNemar's test* compares two classifiers by counting the examples which were classified correctly by the first but not by the second classifier (n_{10}) and vice-versa (n_{01}). As the same training and test sets are used for both induction methods, counts can be summed for all ten cross validation experiments. Under the null hypothesis, the classifiers are equal and so are the counts, $n_{10} = n_{01}$. The statistics D, computed as

$$D = \frac{(|n_{01} - n_{10}| - 1)}{n_{01} + n_{10}}$$

is distributed approximately by the χ^2 distribution with one degree of freedom. The difference is significant at $\alpha = 0.05$ level if D is larger than 3.84.

Another evaluation of at least equal importance as the statistical measures is the evaluation done by the domain expert, who ultimately decides whether the models make sense and can be of practical value for solving the problem for which they were induced in the first place.

#	Feature	Categories	Description	Reference
1	APPT_WORST	$< 78.7, \ \geq 78.7$	The worst partial active thromboplastin time	$25 - 33\ s$
2	BE_ICU	$< -12.6, \ \geq -12.6$	Bicarbonate Excess at ICU	$-2\ -\ 2$
3	BLEEDING_T	Yes, No	Physician's impression regarding coagulopathy during operation	No
4	CATECHOLAM	Yes, No	Cathecholamine administration	No
5	EBL	$< 2.5, \ \geq 2.5$	Estimated Blood Loss	
6	MBP_WORST	$< 36.3, \ \geq 36.3$	The worst mean blood pressure	$\geq 60\ mm$Hg
7	PACO2_OR	$< 44.0, \ \geq 44.0$	The worst arterial carbon dioxide tension	$35 - 45\ torr$
8	PH_WORST	$< 7.0, \ \geq 7.0$	The worst pH	$7.35 - 7.45$
9	PT_ICU	$< 22.3, \ \geq 22.3$	Prothrombin time at ICU	$10.7 - 13.0\ s$
10	TYPE_OF_CL	Skin, Bag	The type of closing	
12	PH_ICU	$< 7.20, 7.20 - 7.33,$ > 7.33	The worst pH value at ICU	$7.35 - 7.45$
11	SBP_WORST	$< 57.0, \ \geq 57.0$	The worst systolic blood pressure	$\geq 90\ mm$Hg

Table 1: Selected features and their description (in alphabetical order).

4 Feature mining and model construction

From the set of 96 features, the entropy based discretization found 56 features as irrelevant. RELIEFF assigned negative score to additional four features, thus resulting in a dataset with only 36 features. From these, the expert (a board certified emergency physician) selected 10 predictive features (features 1 to 10 in Table 1) considering also their potential clinical significance. The expert additionally verified and confirmed that among features not included in the set of 36 there are none that should be additionally selected for modeling. This confirms the usefulness of feature subset selection in our setting.

The expert also inspected the categorization found by the entropy-based algorithm by using previous reports, his own pathophysiological knowledge and the analysis of odds-ratio significance. His findings mostly confirm those of the discretization algorithm. For instance, for APPT_WORST he proposed 80 as a simpler boundary than 78.7. For BE_ICU he proposed a higher range of 22.5. Using odds-ratio visualization he also commented that these are the only two attributes that should always be treated as categorical, while other continuous attributes can also be modeled as continuous, if the modeling technique allows it. The remaining features should be, by his opinion, categorized on three rather than two intervals. We can expect that the method used would indeed devise a finer categorization if more patients were available. Apart from this, he found the proposed categorization to be reasonable.

The selected ten features, together with the two-valued outcome (Death, Well) constituted our first dataset. Additionally, the expert proposed another feature subset in which MBP_WORST and PH_WORST were replaced by SBP_WORST and PH_ICU (features 11 and 12 in Table 1), respectively.

Prognostic model	CA	AP	Sens/Spec	aROC
Majority	0.662	0.552	1.000/0.000	0.500
Decision tree (quartiles)	0.824	0.663	0.800/0.870	0.834
Decision tree (entropy)	0.824	0.686	0.822/0.696	0.849
Naive Bayes (quartiles)	0.809	0.777	0.800/0.826	0.891
Naive Bayes (entropy)	0.794	0.777	0.800/0.826	0.882

Table 2: Classification accuracy (CA), average probability assigned to the correct class (AP), sensitivity and specificity (Sens/Spec) and area under ROC curve (aROC).

Modeling algorithms were successful on both datasets. The classification accuracy was especially high for the second one, reaching accuracy of 93% correct classifications. The conditional probabilities in the naive Bayesian classifier and the graphical presentation of the decision tree revealed though the models' main strategy: for decision trees, all the patients which were given Cathecolam were classified to "Death" and similarly, the conditional probability of survival after being given Cathecolam was 0.00. The inspection of the data indeed proved that from 68 patients, all the 16 patients who were given Cathecolam died. The expert confirmed the found relation is sensible but useless. As this drug is usually the last resort used for the most severe patients, it is highly correlated to the patient's outcome but the surgeon cannot use it for making predictions – its use does *not cause* the death but is (in a sense) *caused by* the surgeon's prediction of the probable death of the patient. The expert proposed to remove this feature from the dataset for the further experiments.

We therefore formed a third dataset, with the same features as the second one but with CATECOLAM removed. The results on this dataset are presented in Table 2. Decision trees and naive Bayesian classifier are better than the baseline majority classifier, though the statistical significance of the differences is (at best) marginal, probably due to the low number of patients. Using McNemar's test, the decision tree model with entropy-based discretization was found to be significantly better than majority classifier ($p = 0.04$), while the tree models with quartiles discretization and naive Bayesian models with quartiles and with entropy-based discretization have significance levels of 0.06, 0.09 and 0.14, respectively.

Figure 1 shows a decision tree build from the dataset with all 68 patients. The left tree was obtained using a simple prepruning (minimal number of examples in each leaf is 2, maximal proportion of majority class in each internal node is 90%). For the tree on the right, all decisions with all leaves having the same majority class are removed. Although it is much simpler, the results of the right tree are comparable to the results of the left one for all of the measured statistics.

From the expert's perspective, the right tree is a reasonable model for outcome prediction. It is based on the two important representatives from two of the most important groups of factors which affect the outcome, coagulopathy and acidosis. It is also interesting that the particular importance of this two features to the patient's outcome was theoretically stressed

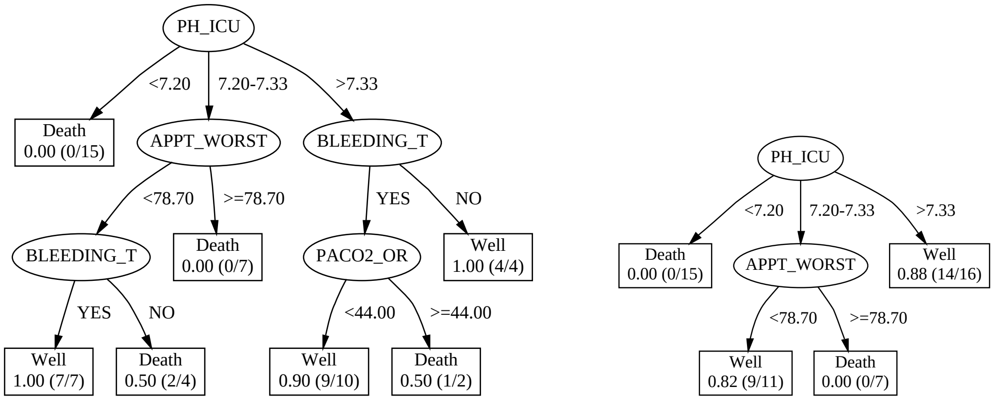

Figure 1: A decision tree model, derived with entropy discretization with simple prepruning (left) and with removing the internal nodes leading to the leaves with the same majority class (right).

in the work of Rotondo et al [11].

The topmost decision in the trees from Figure 1 is based on the blood's pH value (PH_ICU), which reflects many of important aspects of the injury (respiratory and cardiovascular distress, blood loss and cellular damage). As lower value indicates severe damages to patient's vital systems, patients with pH level below 7.20 are not expected to survive. A normal value of pH (> 7.33) predicts probable survival of the patient. The outcome for the patients with the pH values between 7.20 and 7.33 is predicted from the worst partial active thromboplastin time value (APPT_WORST), which assigns a greater probability of survival to the patients with normal blood coagulation.

The larger tree from Figure 1 further divides two subgroups of the patients. The use of the features BLEEDING_T and PACO2OR in the rightmost tree is appropriate, since it predicts the non-bleeding patients and the patients with normal carbon dioxide tension a better chances for survival. Contrary to this, the expert found the left-most node with the BLEEDING_T feature unexpected and in conflict with the domain knowledge and the common sense; at this point the model derivation was probably mislead because of a small sample size. Several nodes of this tree contain very small number of patients. In two cases, exactly half of the patients in the leaf survived and a half did not, so the outcome for patients classified to this two nodes cannot be predicted.

Generally, the domain expert preferred the simpler tree where their leaves represent a higher number of patients. We can, however, speculate that retrospective data that would include a higher number of patients would enable us to induce a larger, yet reasonable decision trees.

5 Conclusion

This paper documents a study to construct outcome prediction models from retrospective data of severe trauma patients. The study should be regarded as pilot since it only includes 68 patients. Despite having such small dataset, the following conclusions can be drawn:

- A rather small subset of features from trauma patient's database seems sufficient for modeling.
- Given a proper selection of features, prognostic models for the outcome for severe trauma patients are plausible.

From methodological point of view, this study has found feature categorization and feature subset selection algorithms useful preprocessing tools. Categorization of most relevant features was inspected and confirmed by the expert. Expert also found the feature rating by RELIEFF to be meaningful. This rating helped him to decide which set of features should be used in the modeling dataset. Both naive Bayes and derivation of classification trees resulted in models of reasonable performance, with two techniques not being significantly different.

The main outcome of this pilot study is the observation that prognostic models can be build for prediction of outcomes for severe trauma patients. This is only a preliminary finding, which needs to be verified in a study that would include a larger number of patients. The present dataset includes many features, and if such comprehensive data collection poses problems — as suggests our present dataset with many missing data — an outcome of this study may help trauma personnel to focus mostly on features that we have observed to be most relevant for prediction models.

References

[1] R. Bellazzi and B. Zupan. Intelligent data analysis in medicine and pharmacology: A position statement. In *IDAMAP-98*, pages 1–4, Brighton, UK, 1988.

[2] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029, Chambery, France, 1993. Morgan-Kaufmann.

[3] T. S. Granchi and K. R. Liscum. The logistics of damage control. *Surg Clin North Am*, 77:921–8, 1997.

[4] K. Kira and L. Rendell. A practical approach to feature selection. In D. Sleeman and P. Edwards, editors, *Proc. 9th Int'l Conf. on Machine Learning*, pages 249–256, Aberdeen, 1992. Morgan Kaufmann Publishers.

[5] I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In F. Bergadano and L. De Raedt, editors, *Proc. European Conf. on Machine Learning (ECML-94)*, pages 171–182. Springer-Verlag, 1994.

[6] I. Kononenko, I. Bratko, and M. Kukar. Application of machine learning to medical diagnosis. In *Machine Learning and Data Mining: Methods and Applications*, pages 389–408. John Wiley & Sons, Chichester, 1998.

[7] K. L. Mattox. Introduction, background, and future projections of damage control surgery. *Surg Clin North Am*, 77:753–759, 1997.

[8] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine learning, neural and statistical classification.* Ellis Horwood, 1994.

[9] T. Niblett and I. Bratko. Learning decision rules in noisy domains. In *Expert Systems 86*, pages 15–18. Cambridge University Press, 1986. (Proc. EWSL 1986, Brighton).

[10] R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[11] M. F. Rotondo and D. H. Zonies. The damage control sequence and underlying logic. *Surg Clin North Am*, 77:761–777, 1997.

[12] I. Kononenko, B. Zupan. Attribute mining: Evaluation, discretization, subset selection and constructive induction In *"From Machine Learning to Knowledge Discovery in Databases" Workshop Notes*, ICML-99, Bled, Slovenia.