THE NATIONAL UNIVERSITY
of SINGAPORE

School of Computing
Lower Kent Ridge Road, Singapore 119260

TRD9/06

Experimental Analysis on Severe Head Injury Outcome Prediction – A Preliminary Study□□□□

Hongli YIN, Guoliang LI, Tze-Yun LEONG, Vellaisamy KURALMANI, Boon Chuan PANG, Beng Ti ANG, Kah Keow LEE and Ivan NG□

September 2006

Technical Report

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

JAFFAR, Joxan
Dean of School

Experimental Analysis on Severe Head Injury Outcome Prediction – A Preliminary Study

Hongli Yin[1], Guoliang Li[1], Tze-Yun Leong[1], Vellaisamy Kuralmani [2],
Boon Chuan Pang[3], Beng Ti Ang[3], Kah Keow Lee [3], Ivan Ng[3]
{yinhl, ligl, leongtz}@comp.nus.edu.sg, {vkmani}@i2r.a-star.edu.sg
{Boon_Chuan_Pang, Beng_Ti_Ang, kah_keow_lee, Ivan_ng}@nni.com.sg
[1] Medical computing Lab, School of Computing, National University of Singapore
[2]Institute for Infocomm Research, Singapore
[3] Acute Brain Injury Research Laboratory, Department of Neurosurgery, National Neuroscience Institute, Singapore

Abstract

Severe head injury management is a very costly and labor-intensive process. There has been growing interest in building outcome analysis models using existing patient records to facilitate decision making and resource planning. However, traditional methods and results in the literature are often inconsistent in variable discretization, accuracy evaluation and class label assignment. In this paper, we examined the effectiveness of applying different outcome analysis methods in head injury management in a uniform manner, based on a set of actual patient records. We have conducted a set of experiments using sound statistical techniques to derive the results. Besides the comparative analysis that highlight the strengths and limitations of different outcome analysis methods, the experiments also show that Minimal-Description-Length (MDL)-based discretization method can help improve prediction accuracy substantially, and that class label assignments in the classification techniques play a very important role on prediction accuracy.

1 Introduction

Severe head injury is one of the major causes of death and disability worldwide. The process to manage head injury patients is very costly and labor-intensive. To optimize head injury management process and resource utilization in hospitals, many efforts have been done in head injury outcome analysis [3,4,6,7]. For example, Choi *et al*. [3] achieved an overall prediction rate of 77.7% using a prediction tree for outcome after severe head injury. Nissen *et al*. [7] used Bayesian Network to get a 84.3% accuracy to predict live and mild disability, 83.6% accuracy to predict death or vegetative survive, and an overall accuracy of 75.8% on a group of 324 patients. Dora *et al.* [4] designed a decision support system to improve severe head injury treatment procedure.

However, we found that some inconsistencies in literature make the comparisons among different results difficult. Specifically, the inconsistencies are as follows. First, the data set for head injury outcome analysis always contains both numeric and categorical variables. Sometimes, the numeric variables are discretized manually for further analysis. Different ways of discretization will affect the significant factor analysis differently. Second, the performance of built models was not well reported in the literature. Usually, a prediction model should be trained on one data set and tested on another different one; the prediction accuracy on both training and testing data should be reported. However, some papers only reported results from training data[3]. Such evaluation results from the training data only are often over optimistic and therefore may not be instructive for prediction model selection. Third, the definitions of class labels for performance evaluation in different papers are inconsistent. Usually, the outcome of a severe head injury patient can be defined as one of the five Glasgow Outcome Scores (GOS 1-5): death, vegetative state, severely disability, moderate disability or good recovery. In head injury outcome analysis, such five categories can be combined in different ways to build a classification model, such as a) death (GOS 1) and live (GOS 2-5) [9], b) {death or vegetative state} (GOS 1-2), {severe disability} (GOS 3), and {moderate disability or good recovery} (GOS 4-5) [7], c) (GOS 1-3) and (GOS 4-5) [1]. Different combinations of GOS scores will affect prediction accuracy significantly (As shown in Section 4), and make results from different papers incomparable.

In this paper, we tried to show the above inconsistencies from experiments. In our experiments, we compared models built with discretization (either supervised or unsupervised) and models built without discretization. The experiment results show that models built with discretization generally perform better than models built without discretization. Particularly, Minimum-Description-Length-based discretization method performs more stably in improving prediction accuracy. Next, we compared evaluation results from both training data and cross validation. The comparisons show that the results from training data are over optimistic, which means that cross validation results should always be

reported for a practical model comparison or selection. We have applied different methods to a data set collected from a local hospital and tried different ways to combine GOS scores as class labels. The results confirmed that different combinations of GOS scores affect prediction results significantly. It suggests that a consistent model has to be able to deal with various GOS combinations, and any fair model comparison should be performed using the same way of GOS combination.

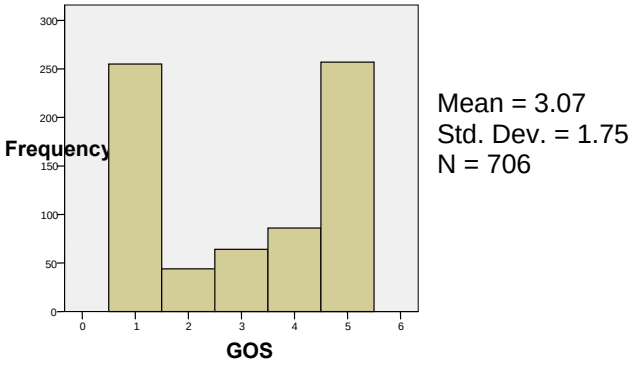
2 Data summary

Our data set contains 706 severe head injury (with Glasgow Outcome Score of 5 or less) patient records, which were collected in a local hospital from January 1999 to March 2005. The follow-up results of the patients at six-month time from the date of admission were collected and measured with Glasgow Outcome Score (GOS): death, vegetative state, serious disability, moderate disability, and good recovery. In the database, there are more than one hundred variables in each patient record. Based on domain knowledge and feature selection, sixteen variables measured at admission time were chosen for experiments. The descriptions of the variables are summarized in Table 1.

The distribution of GOS scores in our data set is shown in Figure 1, from which we know that the data is not equally distributed on different GOS scores: most of the patients are either well recovered or dead.

In the data set, there are some missing values. For numeric variables, we filled missing values with the means of the known values, and for categorical variables, the missing values are filled in with the modes of the known values.

Figure 1 Data distribution with GOS score



3 Our Methodologies

In our experiment, we chose five different data mining methods for outcome analysis: Bayesian Network, Decision Tree, Logistic Regression, Support Vector Machine, and Neural Network.

Table 1 Description of head injury dataset with list of prognostic factors

	Cases	Min	Max	Mean
1. AGE	706	10	97	45.64
2. Gender	706	1	2	1.22
3. Ethnic Group	706	1	4	1.56
4. Mechanism of injury	706	0	6	2.15
5. Types of motor vehicle accident	706	0	7	1.58
6. Alcohol use	706	0	3	.15
7. Presence of traumatic SAH	706	0	2	1.50
8. Presence of cervical injury	706	1	2	1.92
9. Presence of multiple injuries	706	1	2	1.76
10. Pre-resuscitation GCS	703	3	15	9.00
11. Pre-resuscitation papillary light response	703	0	2	1.67
12. Presence of coagulopathy	689	0	2	1.61
13. Presence of hypoxia	706	1	2	1.89
14. Presence of hypotension	706	1	2	1.88
15. Post-resuscitation GCS	698	3	15	7.79
16. Post-resuscitation papillary light response	691	0	2	1.59
Outcome Glasgow Outcome Scale	706	1	5	3.07

Bayesian Network Bayesian Network is a method to model correlations among a group of variables using directed acyclic graph. Bayesian network can inference the states of the unknown variables with prior probabilities and known evidence, and it has an advantage in handling missing data. Besides giving promising performance, Bayesian Network also can reveal underneath relationships among variables or prognostic factors in our case. We used Bayesnet and another Bayesian method AODE [10] from Weka [11]. AODE achieves highly accurate classification by averaging over all of a small space of alternative naïve-Bayes-like models that have weaker (and hence less detrimental) independence assumptions than naïve Bayes. The resulting algorithm is computationally efficient while delivering highly accurate classification on many learning tasks.

Decision Trees Decision trees [8] represent a supervised approach to classification. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcome. It can be used to explain why a question is being asked. Decision tree

is a map of the reasoning process. Decision trees are excellent tools for helping us to choose between several courses of action. They provide a highly effective structure within which we can lay out options and investigate the possible outcomes of choosing those options. They also help us to form a balanced picture of the risks and rewards associated with each possible course of action.

Logistic Regression Logistic regression (LR) is part of a category of statistical models called generalized linear models. Logistic regression allows one to predict discrete outcomes, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. In LR, univariate analyses are first performed to consider the significant risk factors. Then either backward or forward stepwise method is chosen. In the forward method, one factor is added at a time to increase the prediction performance; in the backward method, one factor is removed at a time to increase (or keep) the prediction performance. After each addition or removal, a beta coefficient or relative weight for that factor is defined. Odds ratios and risk ratios can then be calculated, which are very helpful for decision making.

Support Vector Machine Support vector machines (SVMs) [2] are statistical-learning-based methods for classification and regression. When used for classification, the SVM algorithm creates a hyperplane in a feature space with higher dimension that separates the data into two classes with the maximum-margin. Given training examples labeled either "yes" or "no", a maximum-margin hyperplane is identified which splits the "yes" from the "no" training examples, such that the distance between the hyperplane and the closest examples (the margin) is maximized.

Neural Networks Neural Network or Artificial Neural Network is an information processing technique inspired by the way biological brain system works. A neural network contains a number of interconnected processing nodes (or neurons) working in parallel to solve a particular problem.

Neural networks are powerful in deriving meanings from complex or imprecise data, which can be used to understand or recognize things that are too complex to be noticed by other methodologies. A neural network simulates human brains by learning expertise from examples, and stored knowledge in interneuron connection strengths known as synaptic weights. In our experiment, we applied multilayer perceptron (MLP) which is the most commonly used neural network architecture. MLP is a supervised network which requires a labeled training data for learning. Backpropagation is used to adjust the weights a small amount at a time in a way that reduces the error. The ultimate goal of the training process is to reach an optimal solution based on our performance measurement.

4 Model building

4.1 Evaluation measures

We define our prediction accuracy as total number of correctly predicted samples divided by the number of the total samples. We applied all together 6 machine learning algorithms (AODE, Bayesnet, Logistic Regression, Support Vector Machine, and Neural Network) to our data set, and we have 6 different ways to define the class labels:

1) 5 class labels. One for each GOS score: {death}, {vegetative state}, {severe disability},{moderate disability},{good recovery};
2) 3 class labels: {death}, {vegetative state, severe disability, moderate disability}, {good recovery};
3) 2 class labels: {death, vegetative state} and the rest;
4) 2 class labels: {death} and the rest;
5) 2 class labels: {good recovery} and the rest;
6) 2 class labels: {good recovery, moderate disability}, and the rest.

All together there are 36 experiments conducted. In each experiment, we applied 10-fold cross validation. Namely, we did training and testing for ten rounds; at each round, we randomly split data into 10 pieces; then we train our model using 9 pieces of them, and test it on the left 1 piece to get an accuracy; Finally we get the overall accuracy by taking the average from 10 rounds of testing results. We also tested our models on training data in each experiment.

4.2 Effects of different discretization methods on prediction results

To examine the effects of different discretization techniques on prediction accuracy, we tried three different methods: 1) without discretization; 2) the numeric variables are discretized into 5 equal-sized bins (an unsupervised method); and 3) the numeric variables are discretized with a supervised minimum-description-length (MDL)-based method proposed by Fayyad and Irani [5]. After (or without) discretization, models are built with decision tree method and the prediction accuracies are reported in Table 2.

Table 2 Prediction accuracy from decision tree method after different discretizations

Methods	Training	Testing
No discretization	67.42 %	58.07 %
Discretization using 5 equal-sized bins	73.23 %	58.07 %
Discretization with MDL method	69.97 %	62.18 %

The results show that unsupervised method (5 equal-sized bins here) can improve the accuracy on the training data; however, the accuracy from cross validation is the same as the results without discretization. The supervised MDL discretization method has a more stable performance, and it can well improve the accuracy both on training data and on cross validation. Therefore, we applied the MDL-based discretization to our data set for the following experiments. The variable "AGE" is discretized into two categories: either below or above 62. The Variable "preGCS" is discretized into three categories: one for original values 3-6, one for original values 7-11, and another one for original values 12-15. The Variable "rGCS" is discretized into three categories too: one for original values 3-5, one for original values 6-11, and another one for original values 12-15.

4.3 Results for 5 class labels – One for each GOS score.

Five class GOS prediction is generally a challenging problem which is mainly due to imbalance of the training data (See Figure 1 for GOS score distribution). By applying different algorithms to the dataset, we got the results as shown in Table 3. Among all the algorithms, neural networks achieved 89% accuracy on the training data, but with 52% accuracy on the testing data. It means that neural networks may memorize the training data but cannot generalize well. For the testing data, decision tree and SVM achieved the comparable results (around 62%).

4.4 Results for 3 class labels

In this set of experiments, we tried one way to group the 5 GOS outcomes into 3 categories, namely class 1 = {death}, class 2 = {vegetative state, Severe disability, moderate disability}, and class 3= {good recovery}. The prediction results are shown in Table 4. In this round, neural networks can achieve very good results on the training data (94%), but with poor prediction accuracy (59%). Here decision tree gives the best accuracy on testing data (65%).

4.5 Results for 2 class labels

There are a few ways to group GOS scores into 2 classes depending on different clinical purposes. We proposed the following four groups:

- Case 1: To predict for death & vegetative state: class 1 ={death, vegetative state}, and class 2= the rest
- Case 2: To predict for good recovery & moderate disability: class 2= {good recovery, moderate disability}, and class 1= the rest
- Case 3: To predict for good recovery: class 2= {good recovery} and class 1= the rest
- Case 4: To predict for death: class 1= {death} and Class 2= the rest

All together we have 4 groups of experiments as shown in Table 5 - Table 8. From the experiments we found that:

1) All the methods can achieve comparable prediction accuracy on the testing data (around 76% ~ 82%) under different assignments of the two GOS classes.

Table 3 Results for 5 class labels

Methods	Training	Testing
AODE	67.71 %	61.05%
Bayesnet	61.75 %	60.05%
Decision Tree	69.97 %	62.18%
LR	65.86 %	61.47%
SVM	64.73 %	62.46%
Neural Network	89.23 %	52.83%

Table 4 Results for 3 class labels

Methods	Training	Testing
AODE	68.83 %	64.16 %
Bayesnet	65.01 %	62.60 %
Decision Tree	75.49 %	65.15 %
LR	68.55 %	62.74 %
SVM	69.12 %	62.32 %
Neural Network	94.61 %	59.49 %

Table 5 Results for 2 class labels (death vs all others)

Methods	Training	Testing
AODE	82.29 %	80.31 %
Bayesnet	79.46 %	79.32 %
Decision Tree	85.12 %	82.15 %
LR	84.70 %	81.16 %
SVM	83.42 %	81.86 %
Neural Network	96.17 %	77.76 %

Table 6 Results for 2 class labels (death-vegetative vs others)

Methods	Training	Testing
AODE	82.72 %	81.30 %
Bayesnet	79.46 %	79.32 %
Decision Tree	87.54 %	80.59 %
LR	84.42 %	81.58 %
SVM	84.13 %	79.46 %
Neural Network	95.75 %	76.35 %

Table 7 Results for 2 class labels (good recovery & mild-disable vs others)

Methods	Training	Testing
AODE	82.44 %	79.60 %
Bayesnet	80.03 %	79.04 %
Decision Tree	82.86 %	79.75 %
LR	81.87 %	79.89 %
SVM	83.29 %	77.90 %
Neural Network	96.32 %	76.63 %

Table 8 Results for 2 class labels (good recovery vs others)

Methods	Training	Testing
AODE	82.01 %	78.61%
Bayesnet	79.60 %	78.75 %
Decision Tree	83.00 %	80.59 %
LR	81.73 %	79.04 %
SVM	83.29 %	80.31 %
Neural Network	96.46 %	77.76 %

2) Neural networks can always achieve good prediction accuracy on the training data. It means that neural networks can be a good method for summary. However, since neural networks never reach the best prediction accuracy on the testing data, it means that neural networks are not good methods for prediction in our experiment. This also tells us that an algorithm that performs well on training data does not necessarily mean that it will also perform well on unknown testing data. Thus it is very important to report results on testing data for a realistic performance evaluation.

3) The prediction accuracies on the training and testing data under two classes are not so different, compared with the results on the 5 class labels and 3 class labels. The reason is that the data with two GOS classes are evenly distributed, so the built model can well represent knowledge from training data. Thus the prediction accuracy on testing data can be very good, or even comparable to the testing result from training data.

5 Discussion and conclusion

In this paper, we have examined the strengths and limitations of different outcome analysis methods for head injury management in a uniform manner. Based on our experiment results, we have found that: 1) Different discretization methods have different effects on the head injury outcome analysis models. Particularly, the MDL-based discretization method has a more stable effect in improving prediction accuracy; 2) The evaluation of head injury outcome analysis models should be based on both the training and testing data. The results from the testing data are more realistic for prediction purpose; 3) Different class label assignments in classification will affect the prediction results significantly. Technically, a good set of class labels should ensure that data distribution is evenly balanced; and 4) performance of an algorithm varies with different training data set (data distribution, missing values etc). As we have seen from the experiments, no individual algorithm can always outperform the rest. By applying multiple algorithms in parallel, we can ensure to get the best possible performance among all.

For future work, we aim to design an automated framework combining different methodologies to facilitate outcome prediction in varying settings. The proposed framework will include adaptive capabilities to deal outcome analysis in complex, changing contexts, such as varying disease situations, patients from different regions, etc.

Acknowledgement

This work is supported by National University of Singapore. We also thank our colleagues R. Joshi, *et al.*, for their kind suggestions and sincere advice.

References

[1] P. Andrews, D. Sleeman, P. Statham, A. McQuatt, V. Corruble, P. Jones, T. Howells, C. Macmillan, Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression, Journal of neurosurgery 97 (2003) 326-336.

[2] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 2 (1998) 121-167.

[3] S. Choi, J. Muizelaar, T. Barnes, A. Marmarou, D. Brooks, H. Young, Prediction tree for severely head-injured patients, Journal of neurosurgery 75 (1991) 251-255.

[4] C.S. Dora, M. Sarkar, S. Sundaresh, D. Harmanec, T.T. Yeo, K.L. Poh, T.-Y. Leong, Building decision support systems for treating severe head injuries, in: IEEE International Conference on Systems, Man and Cybernetics, Vol. 5 (2001) 2952-2957.

[5] U.M. Fayyad, K.B. Irani, Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, in: Proceedings of 13th International Joint Conference on Artificial Intelligence (IJCAI) (1993) 1022-1027.

[6] D. Harmanec, T.Y. Leong, S. Sundaresh, K.L. Poh, T.T. Yeo, I. Ng, T.W.K. Lew, Decision analytic approach to severe head injury management, in: Proceedings of the 1999 AMIA Annual Symposium (1999) 271-275.

[7] J.J. Nissen, P.A. Jones, D.F. Signorini, L.S. Murray, G.M. Teasdale, J.D. Miller, Glasgow head injury outcome prediction program: an independent assessment, Journal of Neurology, Neurosurgery, and Psychiatry 67 (1999) 796-799.

[8] J.R. Quinlan, C4.5: programs for machine learning (Morgan Kaufmann, San Mateo, Calif., 1993).

[9] D.F. Signorini, P.J.D. Andrews, P.A. Jones, J.M. Wardlaw, J.D. Miller, Predicting survival using simple clinical variables: a case study in traumatic brain injury, Journal of Neurology, Neurosurgery, and Psychiatry 66 (1999) 20-25.

[10] G.I. Webb, J.R. Boughton, Z. Wang, Not So Naive Bayes: Aggregating One-Dependence Estimators, Machine Learning 58 (2005) 5-24.

[11] I.H. Witten, E. Frank, Data mining: practical machine learning tools and techniques with Java implementations (Morgan Kaufmann, San Francisco, 1999).