Original Article

Comparison of four variable selection methods to determine the important variables in predicting the prognosis of traumatic brain injury patients by support vector machine

Saeedeh Pourahmad[1,2], Soheila Rasouli-Emadi[2], Fatemeh Moayyedi[3], Hosseinali Khalili[4]

[1]Bioinformatics and Computational Biology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran, [2]Department of Biostatistics, Medical School, Shiraz University of Medical Sciences, Shiraz, Iran, [3]Department of Computer Engineering, Larestan University, Lar, Iran, [4]Trauma Research Center, Shahid Rajaee (Emtiaz) Hospital, Shiraz University of Medical Sciences, Shiraz, Iran

Background: Large amounts of information have called for increased computational complexity. Data dimension reduction is therefore critical to preliminary analysis. In this research, four variable selection (VS) methods are compared to obtain the important variables in predicting the prognosis of traumatic brain injury (TBI) patients. **Materials and Methods:** In a retrospective follow-up study, 741 TBI patients who were hospitalized for at least 2 days and had a Glasgow Coma Scale score of at least one were followed. Their clinical data recorded during intensive care unit (ICU) admission and eight-category extended GOS conditions 6 months after discharge were utilized here. Two filter- and two wrapper-based VS methods were applied for comparison. A support vector machine (SVM) classifier was then used, and the sensitivity, specificity, accuracy, and the area under the receiver characteristic curve (AUC) values were calculated. **Results:** Theoretically, the variables selected by sequential forward selection (SFS) method would better predict the prognosis (AUC = 0.737, 95% confidence interval [0.701, 0.772], specificity = 89.2%, sensitivity = 58.9% and accuracy = 79.1%) than the others. Genetic algorithm (GA), minimum redundancy maximum relevance (MRMR), and mutual information method were in the next orders, respectively. **Conclusion:** The use of an SVM classifier on optimal subsets given by GA and SFS reveals that wrapper-based methods perform better than filter-based methods in our data set, although all selected subsets, except for the MRMR, were clinically accepted. In addition, for prognosis prediction of TBI patients, a small subset of clinical records during ICU admission is enough to achieve an accepted accuracy.

Key words: Variable selection, traumatic brain injury, support vector machine, filter, wrapper, prediction, prognosis

How to cite this article: Pourahmad S, Rasouli-Emadi S, Moayyedi F, Khalili H. Comparison of four variable selection methods to determine the important variables in predicting the prognosis of traumatic brain injury patients by support vector machine. J Res Med Sci 2019;24:97.

INTRODUCTION

Advances on data collection and storage capacity in recent decades have resulted in a flood of information in many sciences, including medicine. Although large datasets are very valuable, they create a lot of computational complexity, which is why reducing data dimensionality is critical. The dimensionality reduction methods are divided into two general categories of variable selection (VS) and variable extraction. In this research, the VS method was utilized to reduce data dimensionality. The VS process is not only time- and cost-efficient but also can improve the prediction performance and better understand the data.[1]

Dimensionality reduction is necessary to achieve certain goals. For instance, Rosati and Balestra divided 80 migraine patients into two groups (mild and severe) to better reduce data dimensionality and improve classification accuracy. They compared three different

Access this article online

Quick Response Code:

Website: www.jmsjournal.net

DOI: 10.4103/jrms.JRMS_89_18

Address for correspondence: Dr. Hosseinali Khalili, Trauma Research Center, Shahid Rajaee (Emtiaz) Hospital, Shiraz University of Medical Sciences, Shiraz, Iran. E-mail: khalili_h@sums.ac.ir

Received: 17-02-2018; **Revised:** 22-03-2019; **Accepted:** 13-08-2019

VS.[2] Luque-Beana *et al*. performed a study in 2014 to identify the most effective genes that generate cancer. Using the genetic algorithm (GA) and the sequential forward selection (SFS) algorithm, they concluded that the GA method had provided a smaller and more efficient subset of genes than the SFS method.[3] Drotár *et al*. conducted a comparative study in 2015 to compare ten VS methods and consequently arrive at a reduction method in simulated biomedical data suitable for the diagnosis of Parkinson's disease.[4] Güney and Kalkan also conducted a study in 2013 to reduce the dimensions of simulated gene expression data for diseases such as leukemia, breast cancer, prostate cancer, and colon cancer. To achieve this, they compared the three methods of nonlinear and uncontrolled dimensionality reduction methods.[5] Obviously, since each VS algorithm may behave differently on different datasets,[1] scientists continue their analysis of VS methods on different datasets.

In the clinical researches, brain trauma, caused by a damage or injury to the brain by an external factor, is one of the most common causes of death throughout the world.[6] Since these patients have highly unstable conditions during the first weeks after traumatic brain injury (TBI), accurate medical evaluations are needed to predict the prognosis of these patients. Prognosis is a medical term used to describe a patient's future status. The only diagnostic criterion used for assessing the status of TBI patients is the Glasgow Outcome Scale (GOS), which is an ordering scale with 5 classes. Furthermore, the eight-category extended GOS (GOSE) is used to increase the sensitivity of TBI result evaluations.[6]

There are many studies on predicting the prognosis of TBI patients. While some of these researches deal with prognostic data modeling,[6-9] others aim to select the most important variables and compare different analytical methods.[6,10-13] In our previous study,[6] we predicted the prognosis of favorable outcome and ranked all intensive care unit (ICU) findings of TBI patients through a hybrid method combining decision tree and artificial neural network. However, VS, a preliminary analysis of modeling that improves the results, was not applied for that study. Accordingly, this study is conducted to determine an optimal subset of the variables affecting the prognosis of these patients, 6 months after being discharged from the ICU.

MATERIALS AND METHODS

Study design and participants

In a retrospective follow-up study, all TBI patients admitted to the ICU of Shahid Rajaee Hospital (Shiraz, south of Iran) during 2013–2014 were included in the study. These patients suffered brain injuries from various causes of motor vehicle accidents, falls, assault injuries, and pedestrian phenomena. The inclusion criteria were hospitalization for at least 2 days and the patients without any follow-up information for 6 months after discharge. Seven hundred and forty-one TBI patients were then admitted. All data were anonymized from TBI database, and whole study was approved by research committee of Shiraz University of Medical Sciences in a previous thesis numbered 91-6166. All patients are regularly given consent to use their anonymized data on admission by hospital authorities.

Procedures and assessment variables

The patient records during ICU admission were collected. Twenty-nine variables were applied in the study, all of which may affect a patient's GOSE condition on admission including marriage, gender, having meningitis, pupil reaction (two eye response [R], two eye fix [F], noncheckable [N], one response, and one fix [O]), cerebrospinal fluid (CSF) culture, mechanism of injury (four general reasons for TBI), age, systolic blood pressure on admission, ABG finding on admission (arterial oxygen pressure, potential of hydrogen, arterial carbon dioxide pressure, base excess), platelet count, pulse rate, blood hemoglobin level, respiratory rate, fibrinogen level, CT scan Rotterdam score, number of transfused packet cell, number of active CSF samples (PMN > 5), number of surgical site positive cultures, number of positive blood cultures, number of positive sputum cultures, fresh frozen plasma transfusion, thiopental infusion (vial 500 mg), number of neurosurgical procedures, number of ventriculostomy, Glasgow Coma Scale (GCS), and motor component of GCS [Table 1].

Neurosurgeon specialists reassessed the patients' general conditions 6 months after their discharge and classified the patients into two categories: favorable (GOSE ≥5) and unfavorable (GOSE ≤4).[6]

Statistical analysis

Quantitative variables were presented as mean ± standard deviation, and qualitative variables were presented as frequencies (percentage) [Table 1]. Indicators (binary variables coded as 0 or 1) were defined for each category of qualitative variables (the variables increased to 35, then). To find a best method for detecting prognostic factors in TBI patients, four VS methods were applied, two filter-based (mutual information [MI] and minimum redundancy maximum relevance [MRMR]) and two wrapper-based (SFS and GA) methods. In addition, to evaluate the selected subsets, SVM classifier was used. In order to substitute the missing values, the data set was grouped into 30 clusters based on bivariate Euclidean distance among the points and for quantitative variable the mean and for qualitative variable the mode value of that variable in each cluster was

Table 1: Description of the variables for the patients with traumatic brain injury on intensive care unit study admission

Qualitative variables	Frequency (%)
Marriage	
Single	287 (38.7)
Married	454 (61.3)
Gender	
Men	637 (86)
Women	104 (14)
Meningitis	
Have	125 (16.9)
Not have	616 (83.1)
Pupil reaction	
Two eyes responded	500 (67.5)
Two eyes fixed	100 (13.5)
Noncheckable	100 (13.5)
One responded and one fixed	41 (5.5)
CSF culture	
Positive	595 (80.3)
Negative	146 (19.7)
Mechanism of injury	
Motor vehicle	408 (55.1)
Assault	25 (3.4)
Falling	161 (21.7)
Pedestrian	147 (19.8)
GOSE (output: decision classes)	
GOSE equal or <4	248 (33.5)
GOSE >4	493 (66.5)
Quantitative variables	**Mean±SD**
Age (year)	37.7±18.5
Systolic blood pressure (mmHg)	124.17±19.3
Arterial oxygen pressure (mmHg) PaO$_2$	110.74±90.3
PH	7.3±0.54
Arterial carbon dioxide pressure (mmHg) PaCO$_2$	63.62±82.02
Platelet count	213.7±68.6
Pulse rate	96.61±18.06
Blood hemoglobin level	12.58±2.19
Base excess in ABG (mEq/L)	18.8±35.42
Respiratory rate	19.03±4.18
Fibrinogen level	114.05±94.02
CT scan Rotterdam score	39.2±0.98
Number of transfused packet cell	1.35±0.91
Number of active CSF analysis	2.94±0.83
Number of surgical site-positive cultures	0.14±0.43
Number of positive blood cultures	0.15±0.42
Number of positive sputum cultures	0.44±0.71
GCS	9.16±3.64
GCS (motor part)	4.75±1.34
Number of fresh frozen plasma transfusion	5.73±7.75
Number of thiopental vial 500 mg infusion	3.05±8.09
Number of neurosurgery procedures	0.7±0.91
Number of ventriculostomy needed	0.16±0.36

GOSE=Extended Glasgow Outcome Scale; PH=Potential of hydrogen; CT=Computed tomography; GCS=Glasgow Coma Scale; ABG=Arterial blood gas; CSF=Cerebrospinal fluid; SD=Standard deviation

substituted. The following section briefly reviews each method.

Mutual information

Intuitively, MI measures the information shared between *X* and *Y*. It measures the uncertainty's reduction in one variable by knowing the values of the other one. The *X*s (input variables) in this method are ranked on the basis of their relationship with the *Y* (the output variable). The highest ranked variables will be nominated to enter the optimal subset.[14] To assess the relationship between the variables, variable entropy is calculated through a formula based on their joint probability distribution. Then, the amount of entropy's reduction in *Y* accounting the values of *X* is considered as MI value for *X*.

Minimum redundancy maximum relevance

The relation with the response variable and the minimum redundancy are both considered in the extended MI method to prevent the entrance of variables with overlapping.[15]

Sequential forward selection

Since the SFS algorithm is a wrapper-based method, the VSs coincide with the classification procedure. Initially, the SVM classifier categorizes the data without input variables in the model. Then, the variable that increases the categorization accuracy is entered step by step. To add another variable in the next step, the presence of the previous entered variable (s) is considered. This process is repeated until the entered variable does not affect the categorization accuracy.[1]

Genetic algorithm

The smallest constituent unit in the GA, as inspired by the natural cycle, is the gene. The variables in the present research define a set of genes called chromosomes. The GA is executed through selection, composition, mutation, fitness, and similar functions. Research's variables are called the population. This algorithm randomly selects a number of variables referred to as the initial population. Using the selection operator, two chromosomes are selected among the initial population to serve as the parent. The two are then combined by a mutation operator and form the child chromosomes. Then, the suitability of these chromosomes is measured by the fitness function in each stage. This process is continued until a steady number of generations, which were in fact the same selected variables,[16] is chosen. In the present study, Roulette Wheel method for parent selection, flip bit method for mutation with a mutation rate of 0.1, and crossover method for evaluation were applied in the GA analysis.

Support vector machine

An SVM is actually a binary classifier that separates two classes using a linear boundary. In this method, the samples that form the boundaries of the classes are obtained using an optimization algorithm (support vectors). The two classes are separated by a linear classifier function called the hyperplane. Assuming that the classes are linearly separable, maximum margin hyperplanes are achieved by this method to separate the classes. Kernels are used in cases where the data are not linearly separated. For further examples see the reference.[17]

Performance evaluation

The sensitivity and specificity are two main quantities that are used to evaluate the performance of the classification models. While sensitivity is a proportion of positive cases that are correctly introduced by the model as the positive rate, specificity is a proportion of negative cases that are correctly introduced as the negative rate. Various indices have been proposed to evaluate the performance of a classifier based on these two quantities. One of these indices is the area under the receiver characteristic curve (AUC) (based on the sensitivity versus 1-specificity). The large AUC indicates that the probability of being positive is higher for an individual with a positive result than one with a negative result, which is the indication of better performance for the considered model than the others.[18]

In this study, the sensitivity and specificity values, model accuracy, and the AUC were all used for comparison among the methods. 10-fold cross-validation method was also used in the result evaluation.

RESULTS

A total of 741 patients who were suffered TBI according to TBI database inclusion criteria (head abbreviated injury severity ≥3) were participated in the present study. Their ages ranged from 14 to 91 years (mean = 33.7, standard deviation = 18.5) with 86% identified as male and 14% as female [Table 1]. The youngest patients were two boys, who were traumatized one as pedestrian phenomena and another as motor vehicle accidents. They had a GCS of 5 and 13, respectively. The oldest patient was a man who had been traumatized in a fall and had a GCS of 12.

Most of the patients' characteristics such as ABG, CBC, number of pack cell and FFP, fibrinogen level, GCS, pupil reaction, pulse rate and systolic blood pressure, and the admission data were applied. Other variables (cultures, thiopental, and procedures) are collected during hospital stay [Table 1].

The missing values were substituted by k-means clustering method (section 2.3), and four VS procedures were applied on dataset. Table 2 represents the important variables in prognosis prediction in the order they had been selected by each method.

As shown, four different VS methods determine different collection of clinical, laboratory, and demographic parameters as the important variables in prognosis [Table 2]. Age, coagulation profile, and hemodynamic stability indexes as well as primary neurologic condition play a major role in prediction of prognosis of the patients in all four methods with different priorities.

To compare the methods statistically, the SVM modeling method was applied on all subsets selected by four methods, and the results are summarized on Table 3. As seen, MRMR method has the most specificity. That means, the subset selected by this method can best predict the favorable prognosis, while it has the lowest sensitivity and is weak in predicting unfavorable prognosis. According to accuracy as a combined index and AUC as the most popular index for evaluating the binary classifiers, the best subset is selected by SFS method. This subset has even better performance than the set of all variables. As it is known, the far values of AUC from 0.5 indicate good performance, and if its confidence interval does not contain the value 0.5, it is statistically significant.

DISCUSSION

The main objective of this study was to find the best method which determines the important factors in prognosis of TBI patients. Therefore, all the patients considered in this study were selected from TBI database and patients with no experience of TBI were not important for comparison. Accordingly, four VS procedures including MI, MRMR, SFS, and GA were compared in the present study on a dataset of 741 patients admitted to the ICU. The results revealed the superiority of wrapper-based methods (GA and SFS) in this dataset. The application of SVM on the optimal subset of variables using these two methods resulted in even better results than the SVM on all variables. The SFS method had higher accuracy (79.1%) and AUC (area = 0.737, [0.701, 0.772]) than the other methods. GA, MRMR, and MI followed it, respectively. These results may refer to VS simultaneity and classification in wrapper-based methods. The SFS and GA methods had a simultaneous presence of variables during the modeling process, while the selection of variables in filter-based methods occur independent of the classification. Slightly weaker performance of GA may be due to user selected criteria, such as population size, combination type, selection type, and the fitness function. The superiority of the SFS algorithm is also confirmed by other studies.[1] According to a previous study, the variables selected by GA were more discriminant than those selected by the WEKA_CFS and WEKA ranker (IG).[16] Furthermore, Yang *et al.*'s research showed that MI calculation is influenced by the density's approximation of the variables, which results in lesser accuracy of the MI method. Low performance of MI is also confirmed by our results.[19]

Comparing optimal subset of variables obtained by each method is interesting clinically. The selected subset of

Table 2: Selected variables by four variable selection methods

Forward selection	GA	MRMR	MI
Platelet count	Age	Platelet count	Platelet count
Pupil reaction	Number of CSF	Mechanism of injury	Arterial oxygen pressure (mmHg)
Mechanism of injury	Number of positive Sputum cultures	Pupil reaction=Two eyes responded	Base excess in ABG (mEq/L)
Age	Fresh frozen plasma pretention transfusion	PH	Age
Fresh frozen plasma pretention transfusion	Baseline GCS	Number of ventriculostomy surgery	Fibrinogen level
Baseline GCS	Number of neurosurgery procedures	Pupil reaction=Two eyes fixed	Systolic blood pressure (mmHg)
Number of CSF	Number of thiopental vial 500 mg infusion	Number of surgical site positive cultures	Pulse rate

MRMR=Minimum redundancy maximum relevance; GCS=Glasgow coma scale; GA=Genetic algorithm; MI=Mutual information; CSF=Cerebrospinal fluid; ABG=Arterial blood gas

Table 3: The results of support vector machine classifier on different subsets of variables

Input variables	AUC	95% CI for AUC	Specificity (%)	Sensitivity (%)	Accuracy (%)
All variables	0.719	0.683-0.756	86.6	58.1	77.1
Selected by SFS method	0.737	0.701-0.772	89.2	58.9	79.1
Selected by GA method	0.715	0.678-0.752	88.4	55.2	77.3
Selected by MRMR method	0.607	0.565-0.629	89.4	32.7	70.4
Selected by MI method	0.569	0.526-0.612	79.7	34.3	64.5

SFS=Sequential forward selection; GA=Genetic algorithm; MRMR=Minimum redundancy maximum relevance; MI=Mutual information; AUC= Area under the receiver operating characteristic curve; CI=Confidence interval

predictors for the prognosis of TBI patients by MRMR was not clinically suitable. whereas, for the MI method, the selected subset was superior since more treatment-related variables were selected. The optimal subset of variables by the SFS and GA methods were also verified by the fourth author of the present study as a clinical expert [Table 3]. Indeed, the rank of variables obtained by GA was closer to the clinical literature. This method determined the age of TBI patients as the most important variable in prediction of his/her prognosis 6 months after discharge. Number of active CSF, number of positive sputum cultures, fresh frozen plasma transfusion, admission GCS, number of neurosurgery procedures, and number of thiopental vial 500 mg infusion are at the next orders. All these variables are clinically effective variables on the prognosis. However, the range of glucose during ICU stay which was determined as an important risk factor on the survival of TBI patients by the previous study, had not been included in our data set.[10]

Comparing to our previous study, the rank of the variables obtained by these four VS methods does not match the order determined by the hybrid modeling method.[6] It seems that further research is needed on this topic using different VS methods and classifier, since the selected variables can be used for clinical follow-up researches aiming to reduce occurrence of unfavorable prognosis in TBI patients.

Acknowledgments

This work was supported by the grant number 95-11511 from Shiraz University of Medical Sciences Research Council. This article was extracted from Soheyla Rasouli-Emadi's Master of Science Thesis.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40:16-28.
2. Rosati S, Balestra G. A comparison of performances of different feature selection methods applied to biomedical data. Int J Bioelectromagn 2012;14:198-203.
3. Luque-Baena RM, Urda D, Subirats JL, Franco L, Jerez JM. Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data. Theor Biol Med Model 2014;11 Suppl 1:S7.
4. Drotár P, Gazda J, Smékal Z. An experimental comparison of feature selection methods on two-class biomedical datasets. Comput Biol Med 2015;66:1-10.
5. Güney F, Kalkan K. Comparison of Non-linear Dimensionality Reduction Techniques for Classification with Gene Expression Microarray Data; 2013.
6. Pourahmad S, Hafizi-Rastani I, Khalili H, Paydar S. Identifying important attributes for prognostic prediction in traumatic brain injury patients. A hybrid method of decision tree and neural network. Methods Inf Med 2016;55:440-9.
7. Subaiya S, Roberts I, Komolafe E, Perel P. Predicting intracranial hemorrhage after traumatic brain injury in low and middle-income countries: A prognostic model based on a large, multi-center, international cohort. BMC Emerg Med 2012;12:17.
8. Lesko MM, Jenks T, O'Brien SJ, Childs C, Bouamra O, Woodford M,

et al. Comparing model performance for survival prediction using total glasgow coma scale and its components in traumatic brain injury. J Neurotrauma 2013;30:17-22.

9. Sobuwa S, Hartzenberg HB, Geduld H, Uys C. Predicting outcome in severe traumatic brain injury using a simple prognostic model. S Afr Med J 2014;104:492-4.
10. Terzioglu B, Ekinci O, Berkman Z. Hyperglycemia is a predictor of prognosis in traumatic brain injury: Tertiary intensive care unit study. J Res Med Sci 2015;20:1166-71.
11. Oh H, Seo W. Functional and cognitive recovery of patients with traumatic brain injury: Prediction tree model versus general model. Crit Care Nurse 2009;29:12-22.
12. Shibuya TY, Karam AM, Doerr T, Stachler RJ, Zormeier M, Mathog RH, *et al.* Facial fracture repair in the traumatic brain injury patient. J Oral Maxillofac Surg 2007;65:1693-9.
13. Lingsma HF, Roozenbeek B, Steyerberg EW, Murray GD, Maas AI. Early prognosis in traumatic brain injury: From prophecies to predictions. Lancet Neurol 2010;9:543-54.
14. Cover TM, Thomas JA. Elements of Information Theory. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2012.
15. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226-38.
16. Babatunde O, Armstrong L, Leng J, Diepeveen D. A genetic algorithm-based feature selection. Br J Math Comput Sci 2014;4:889-905.
17. Jakkula V. Tutorial on Support Vector Machine (SVM). School of Electrical Engineering and Computer Science, Washington State University; 2006.
18. Pourahmad S, Kharrati-Kopaei M, Hosseini-Moghadam Z, Nouri S, Jahromi MA. An optimal combination of several binary markers based on Neyman-Pearson lemma: with application to immunohistochemical markers for the diagnosis of classical Hodgkin. Adv Appl Stat 2016;48:369.
19. Yang Y, Pedersen JO, editors. A Comparative Study on Feature Selection in Text Categorization. USA: International Conference on Machine Learning; 1997.