Expert Systems with Applications 38 (2011) 15534–15539

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Comparison of regression tree data mining methods for prediction of mortality in head injury

Necdet Sut [a,*], Osman Simsek [b]

[a] Department of Biostatistics and Medical Informatics, Trakya University Medical Faculty, Edirne, Turkey
[b] Department of Neurosurgery, Trakya University Medical Faculty, Edirne, Turkey

ARTICLE INFO

Keywords:
Data mining
Regression tree
Classification
Head injury
Mortality

ABSTRACT

With this research, we sought to examine the performance of six different regression tree data mining methods to predict mortality in head injury. Using a data set consisting of 1603 head injury cases, we assessed the performance of: the Classification and Regression Trees (CART) method; the Chi-squared Automatic Interaction Detector (CHAID) method; the Exhaustive CHAID (E-CHAID) method; the Quick, Unbiased, Efficient Statistical Tree (QUEST) method; the Random Forest Regression and Classification (RFRC) method; and the Boosted Tree Classifiers and Regression (BTCR) method, in each case based on sensitivity, specificity, positive/negative predictive, and accuracy rates. Next, we compared their areas under the (Receiver Operating Characteristic) ROC curves. Finally, we examined whether they could be grouped in meaningful clusters with hierarchical cluster analysis. Areas under the ROC curves of regression tree data mining methods ranged from 0.801 to 0.954 ($p < 0.001$ for all). In predicting mortality in head injury under the ROC curve, the BTCR method achieved both the highest area (0.954) and accuracy rate (93.0%), while the CART method achieved both the lowest area (0.801) and accuracy rate (91.1%). All of the regression tree data mining methods were clustered in the same grouping, but the BTCR method was at the origin of the cluster while the CART and QUEST methods produced results that were least like the others. The BTCR, demonstrating a 93.0% accuracy rate and showing statistically significantly differences from the others, may be a helpful tool in medical decision-making for predicting mortality in head injury.

1. Introduction

There are many classification methods that can be employed in the task of data mining. One of them is regression tree, a method that is been widely used for classification tasks. The regression tree examines classification predictions of categorical variables which are the target of various areas. There are many applications for this method, including predicting business failure (Li, Sun, & Wu, 2010), essential hypertension (Ture, Kurt, Kurum, & Ozdamar, 2005), and aggressive prostate cancer on biopsy (Supurgeon et al., 2006).

Head injury is well known to be an important public health issue, affecting all age groups, throughout the world. It is one of the major causes of death and disability (Koskinen & Alaranta, 2008; Signorini, Andrews, Jones, Wardlaw, & Miller, 1999). As such, patients are generally anxious about the prognosis soon after sustaining a head injury. However, the current methods for determining final outcome predictions from head injury patients are imperfect, and present important questions to physician regarding the heterogeneity of patient data, the variety of trauma causes, and additional personal factors such as patient age and the prevalence of systemic disease (Schaan, Jaksche, & Boszczyk, 2002). Recent studies have attempted to show outcomes following head injury, and based their predictions on many factors, including demographics, epidemiologics, and clinic and radiologic findings such as: age, cause of injury, Glasgow Coma Scale score, pupil response, computerized tomography parameters, etc. (Schaan et al., 2002; Signorini et al., 1999).

Predicting the outcome of a head injury is a complex and cognitive process. Use of regression tree methods has proven helpful for medical decision-making in head injury. There exists a great number of studies that have compared risk assessment methods for predicting outcomes in head injury, such as that of Andrews et al. (2002), which compared the results of a decision tree and logistic regression analysis, of Rovlias and Kotsou (2004), which used the CART technique, and of Choi et al. (1991), which used a decision tree.

No study to date has attempted to consider the accuracy performances of: the Classification and Regression Trees (CART) method;

* Corresponding author. Address: Department of Biostatistics and Medical Informatics, Trakya University Medical Faculty, 22030 Edirne, Turkey. Tel.: +90 2842357641; fax: +90 2842357652.
E-mail addresses: nsut@trakya.edu.tr (N. Sut), gosimsek@trakya.edu.tr (O. Simsek).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.06.006

the Chi-squared Automatic Interaction Detector (CHAID) method; the Exhaustive CHAID (E-CHAID) method; the Quick, Unbiased, Efficient Statistical Tree (QUEST) method; the Random Forest Regression and Classification (RFRC) method; and the Boosted Tree Classifiers and Regression (BTCR) simultaneously in the clinical application of predicting mortality from head injuries. We sought to address this oversight in this study, with which we looked to compare the predictive accuracy of various regression tree data mining methods to this end.

2. Methods

2.1. Head injury data set

We applied the regression tree models to a real clinical data set of 1603 patients with head injuries. We selected prognostic risk factors on mortality in head injury by using logistic regression analysis with the backward stepwise method. This is explained in the following section.

2.1.1. Logistic regression analysis

Logistic regression analysis is commonly used when the independent variables include both numerical and nominal measures, and the outcome variable is binary or dichotomous and has only two values. It can also be used when the outcome has more than two values. The logistic model is expressed as follows:

$$p_{x(Event=dead)} = \frac{1}{1+\exp[-(b_0+b_1x_1+b_2x_2+\cdots)]}$$

where $x_1, x_2, \ldots$ are independent variables, b_0 is the intercept, $b_1, b_2, \ldots$ represent the regression coefficient and exp indicates that the base of the natural logarithm (exp = 2.718) is taken to the power shown in parentheses. The equation can be calculated using a stepwise method similar to the one for multiple regression; a chi-square test (instead of the t or F test) is used to determine whether a variable adds significantly to the prediction (Dawson-Saunders & Trapp, 1993).

2.1.2. Selection of risk factors by logistic regression analysis with backward stepwise method

The dependent variable of the regression tree models is mortality status (i.e., dead [1] or alive [0]).

Among the 19 prognostic variables, only eight were found to be significant influences on mortality in head injury, according to the backward stepwise logistic regression model. The flow chart of the research design is shown in Fig. 1.

2.2. Data analysis methods

Six different regression tree data mining methods were used in this study. In the following section, their properties are briefly explained.

- CART – Classification and Regression Trees
- CHAID – Chi-squared Automatic Interaction Detector
- E-CHAID – Exhaustive CHAID
- QUEST – Quick, Unbiased, Efficient Statistical Tree
- RFRC – Random Forest Regression and Classification
- BTCR – Boosted Tree Classifiers and Regression

2.2.1. Classification and Regression Trees (CART)

The CART method is a non-parametric technique that produces either classification or regression trees, depending on whether the

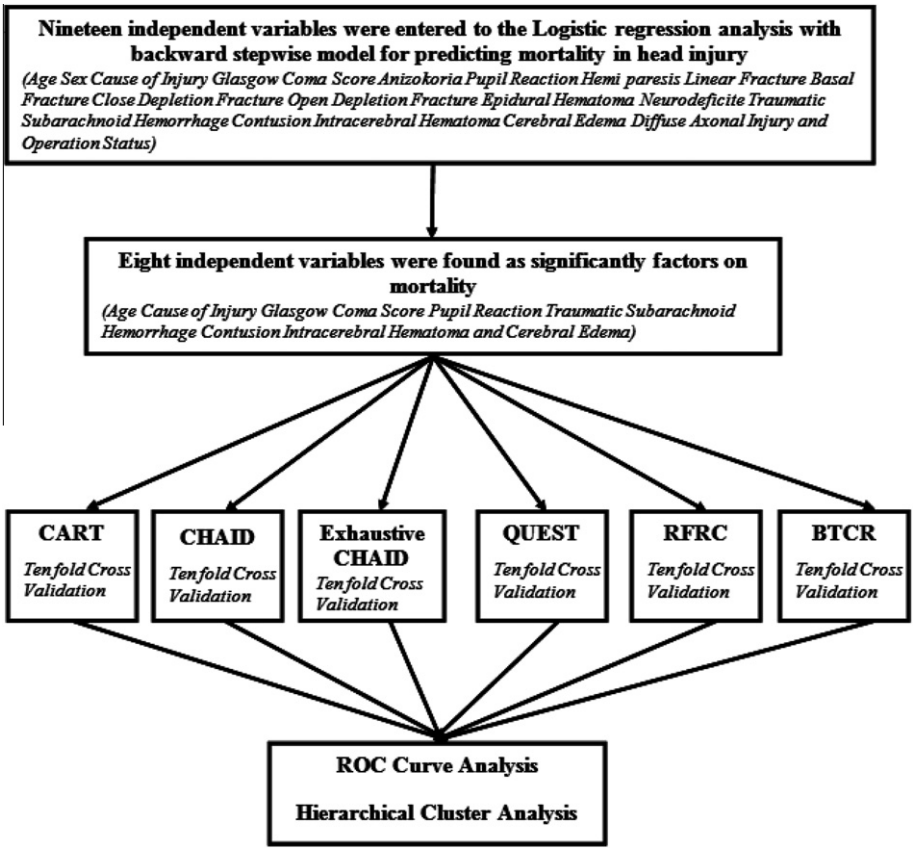

Fig. 1. Flow chart of the research design.