Decision Tree Induction to Prediction of Prognosis in Severe Traumatic Brain Injury of Brazilian Patients from Florianopolis City

Merisandra C. M. Garcia, Evandro T. Martins, and Fernando M. Azevedo

Abstract—**The data mining consists in identification of characteristics and relationships between data, aiming the transformation of these into useful knowledge. The obtainment of knowledge occurs trough tasks, methods and algorithms that have specific purposes, and that are applied according to the goals of analysis. The analysis showed in this article consists in application of data mining by task of classification by the method of decision tree induction by the C4.5 algorithm for the prediction of prognosis in severe traumatic brain injury. The traumatic brain injury is a public health problem, constituting in one of the main causes of morbidity and mortality. In the development were performed the steps of preprocessing, the application of data mining and the evaluation of generated model, obtaining the accuracy of 87%.**

I. Introduction

STATISTICS tools, mathematical models, structured queries are commonly used to assist in obtaining information from databases. However, these tools have limitations that can compromise the accuracy of the information generated. In this context arises the data mining that gathers techniques of different areas as computational intelligence, recognition of standards, machine learning, statistical methods and databases, to perform knowledge discovery in databases.

The data mining is a tool of analysis of data which has been used for explore the information in search of knowledge at the most different areas, for example, in biomedical for predicting the rate of survival in cases of breast cancer, identify predictors of urinary tract infections, results of traumatic brain injury, among other[1].

The Traumatic Brain Injury (TBI) is a public health problem which is a leading cause of morbidity and mortality on Brazil and World, affecting the most productive age victims [2].

To suffer this type of injury patients and familiars are concerned about the prognosis, but current methods for determining this outcome are imperfect and present to doctors important questions regarding the heterogeneity of patient data, variety of causes of traumas, and other factors such as age and prevalence of systemic diseases. Therefore, predicting the outcome of a TBI is a complex and cognitive process [3].

It's traditionally employed techniques that are not ideal for working with complex biological data, sometimes multidimensional and stored in large data repositories, making it time-consuming process of analysis. Thus, due the fact that there is no consensus on an optimal method, it's interesting to explore different methods [4].

This research was carried out applying of task of classification by middle of decision tree induction method by the C4.5 algorithm in analysis of severe traumatic brain injury, for the purposes of identify the model of death prediction.

The classification is responsible for learning a target function, which is known as the classification model, which can be useful for the purpose of a predictive modeling. Thus, when the set of attributes of a record is unknown submitted to this predictive model, it is able to automatically assign it to a class label.

Among the methods used for data classification, this research applied to decisions tree induction, which are forms of classification that have been employed to support the decision by health professionals, as they have been shown to be suitable for conducting medical predictions [6]. The C4.5 algorithm and the construction of the decision tree implements its simplification rules excluding those who do not have significant values for the accuracy of the model. This reduces the complexity of the tree, which allows a classification faster and more efficient [7].

In this research we opted for the use of a classical model in data mining, in this case the C4.5 algorithm, because according to [8] and [9], respectively, in the article and book titled "Top 10 algorithms in data mining", this algorithm is identified as a most influential and used by the data mining community.

II. C4.5 Algorithms

The C4.5 is a decision tree algorithm, developed by John Ross Quinlan, who finds hypotheses precision as it carries out pruning rules originated in the construction of the tree, but has higher computational cost than its predecessor, ID3, in terms of time and search space [10]. The C4.5 algorithm generates a classifier model, running on two phases: construction and simplification of the decision tree.

The C4.5, as most of the induction algorithms decision tree algorithm is based on Hunt algorithm, it employs a

M. C. M. Garcia is with the Institute of Biomedical Engineering, Electrical Engineering Department, Federal University of Santa Catarina, Florianopolis, SC, Brazil and Computer Science Course, Unit of Sciences, Engineering and Technologies, Extreme South University of Santa Catarina, Criciuma, SC, Brazil (corresponding author to provide phone: 55 48 3431-2553; e-mail: merimattos@gmail.com).

E. T. Martins is with Intensive Care Unit, Governador Celso Ramos Hospital, Florianopolis, SC, Brazil.

F. M. Azevedo is with Institute of Biomedical Engineering, Electrical Engineering Department, Federal University of Santa Catarina, Florianopolis, SC, Brazil (e-mail: azevedo@ieb.ufsc.br).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

strategy that grows a tree considering a series of locally optimal decisions about which attribute to use to partition the data [11].

Points of partitioning for each node of the tree are calculated by measures of information gain [12]. In equation (1) is the calculation of the entropy of the entire data set, S, of the class labels.

$$Info(S) = -\sum_{j=1}^{k}\left(\left(\frac{freq(C_j, S)}{|S|}\right).\log_2\left(\frac{freq(C_j, S)}{|S|}\right)\right) \quad (1)$$

Where: *S* is any set of samples, which may represent the complete base (in the case of the root node) or partitions of the database, is the number of times the class *Cj* happens in *S*, $|S|$ is the number of samples set *S, k* the number of possible classes.

Further, considering attributes of the data set *S*, one must divide it into subsets T_1, T_2 and T_3 representing the possible values of an attribute X. Test the expected information can be found by equation (2), having finally the total gain for the attribute equation (3), and one that maximizes the gain.

$$Info_x(T) = \sum_{i=1}^{n}\left(\left(\frac{|T_i|}{|T|}\right).\inf o(T_i)\right) \quad (2)$$

Where T is the number of occurrences in the partition being analyzed and T_i the number of instances of a class contained in the set T.

$$Gain(X) = Info(T) - Info_x(T) \quad (3)$$

After the initial split, each child node has several samples of the database and the whole process of test selection and optimization is repeated for each child node [13].

In C4.5 has a specified parameter in equation (4) so as to have an additional call information gain ratio [13], defined in Equation (5).

$$SplitInfo(X) = -\sum_{i=1}^{n}\left(\left(\frac{|T_i|}{|T|}\right).\log_2\left(\frac{|T_i|}{|T|}\right)\right) \quad (4)$$

$$Gainratio(X) = \frac{gain(X)}{SplitInfo(X)} \quad (5)$$

The ratio of the gain, equation (5) expresses the percentage of the information generated by the division that appears to be most useful for the classification. After calculating the ratio of the gain for all attributes of the data set, who submit the greater will be the root of the tree, the database being partitioned and the process repeated for each new node generated [7].

The second stage of implementation of the C4.5 algorithm is the simplification of the decision tree, the method of post-pruning, reducing some subtrees to leaves. For that, guided by a statistical test that evaluates the importance of rules generated by the tree, so those do not add knowledge are pruned, originating from a tree with better ranking [7].

The C4.5 algorithm uses the pessimistic pruning, which is calculated for each node of the tree an estimate of the upper confidence limit, U_{cf}. The C4.5 uses the standard confidence interval of 25% and compares $U_{25\%}(|T_i|/E)$ for a given node T_i with confidence weighted their leaves, taking as weight the total number of cases for each sheet. If the error predicted a root node of a subtree is less than the weighted sum of $U_{25\%}$ to the leaves, then the subtree can be replaced by that root node becomes a leaf of a tree pruned [11] [12].

III. DATABASES

The TBI is an assault on the brain, causing injury or anatomical functional impairment of the skull, meninges, or brain [13]. This aggression is caused by an external physical force that may diminish or alter the state of consciousness and impair motor or cognitive abilities [14].

The database used in this study consists of records of patients with severe TBI who were admitted to the Intensive Care Unit of the Governador Celso Ramos Hospital, in the city of Florianopolis, in the period from January 1994 to December 2003. According to [15] this is a public hospital for TBI serving a population of approximately one million, making up the metropolitan region of Florianopolis.

The database consists of 748 records, each of which has 18 attributes that represent the characteristics related to TBI (Table 1).

TABLE I
DATABASE OF SEVERE TBI

Attributes	Values
Sex	Male, female
Age	≥ 12 years
Glucose	≤ 60 to > 300
Year of Service	1994 to 2003
Cause of TBI	Accidents: highway, car, bike; fall, agression, other
Marshall Classification	Type I, II, III and IV
Subarachnoid Hemorrhage	Yes, no
Associated Trauma	Yes, no
Type of associated trauma	
Face	Yes, no
Cervical spine	Yes, no
Dorso-lumbar spine	Yes, no
Chest	Yes, no
Abdominal	Yes, no
Members	Yes, no
Others	Yes, no
Glasgow Coma Scale	3 to 8
Pupils	Isochoric, miotic, anisochoric, mydriatic
Denouement	Survival, death

IV. METHODOLOGY

The development process began with the definition of the problem in data mining has been applied, the observance-to-

gross database regarding the structure of the data set in terms of their attributes and records.

This stage was also characterized by the definition of an expert in the field of application that holds knowledge about the problem, is fundamental in the process, as it helps in identifying the objectives of data mining and evaluation of results. The objectives of the application also make this phase and include the expected characteristics of the knowledge model.

The modeling includes the preprocessing of data and execution of data mining. The preprocessing consisted in organizing and processing the data, preparing them for submission to the algorithm. Data mining refers to the step of applying the algorithms to identify the patterns present in these data, and in this study we used the induction of decision trees by C4.5 algorithm.

The evaluation included the analysis of the model identified by the algorithm above, using quality measures, which consist of statistical indexes to evaluate the performance of the models discovered.

A. Preprocessing

Initially was performed the preprocessing the data in order to improve the application of data mining in terms of time, cost and quality. This step is performed the functions of data selection, aggregation, cleansing and transformation of variables.

The data used in this research were already organized in a single table, so the selection of the consisted of the choice of attributes to be considered in data mining. The selection data for vertical reduction was implemented by removal of attributes whose content is not considered relevant to the problem, such as, for example, the identifier code. Besides this, other attributes were also disposed of some analyzes of data mining, according to guidelines of the specialist field of application, such as: year of service and type of associated trauma. The reduction of vertical data, according to [10], can assist in obtaining knowledge models with higher accuracy and brevity because they eliminate irrelevant characteristics and reduce noise.

The aggregation was employed to reduce the number of possible values of a certain attribute, for example, year of treatment, age and glucose.

The cleaning of the data included the elimination of missing values in the data set, through the exclusion of cases that had attributes with missing values. The method of handling missing data was employed Listwise Deletion (LD) which discards all objects with some attribute missing.

The transformation of the variables must meet the needs of data mining algorithms in the case of this study, the nominal attributes were converted to numeric.

After preprocessing the database now has 728 records, which were submitted to data mining.

B. Application of Data Mining

This step involved the application of induction of decision trees by C4.5 algorithm. In implementing C4.5 defined a confidence level of 25% and standard cross validation method 10 parts, as indicated by the literature in the area, [11] and [12] as being the ideal.

In implementing data mining tool was used Waikato Environment for Knowledge Analysis (Weka) which is distributed under the terms of this General Public License (GNU). The Weka was developed at the University of Waikato in New Zealand, being used in research in the area of data mining. The Weka implements in Java several data mining algorithms, allowing them to run on different platforms [6].

The tool is available for free at: http://www.cs.waikato.ac.nz/~ml/weka/.

The C4.5 algorithm in all analyzes performed using as output attribute (class label) death selected attribute as one pupil more relevant in the database to determine the death of the patients had the highest gain information, thus constituting the root node of the decision tree, as you can see in Fig. 1. Considering the 17 attributes of the database, besides the pupil other attributes are relevant to the determination of death as the value of the Glasgow Coma Scale, subarachnoid hemorrhage, the type of associated trauma, blood glucose and cause of TBI, which were also considered in the generated tree (Fig. 1). The tree can be intuitively translated into a set of rules, because each node corresponds to a comparison of the attribute value.

```
=== Classifier mode l ===
C4.5 pruned tree
------------------
Puplis = Anisocórica
| CATEG = 7-8
| | HAS = yes
| | | MEMBERS = yes
| | | | CATGLIC2 = 221-300: yes (2.0)
| | | | CATGLIC2 = 61-110: no (2.0)
| | | | CATGLIC2 = 111-220: no (1 3.0/2.0)
| | | | CATGLIC2 => 300: no (0.0)
| | | | CATGLIC2 =< 60: no (0.0)
| | | MEMBERS = no: yes (33.0/13.0)
| | HSA = no: no (83.0/12.0)
| CATEG = 5-6
| | CHEST = no
| | | CATIDADE => 60: yes (7.0/1.0)
| | | CATIDADE = 46-60: yes (7.0/ 3.0)
| | | CATIDADE = 12-30
| | | | BIENNIUM2 = 2002-2003-2004: (5.0)
| | | | BIENNIUM2 = 2000-2001: (8.0/2.0)
| | | | BIENNIUM2 = 1998-1999: (14.0/2.0)
| | | | BIENNIUM2 = 1996-1997: (11.0/5.0)
| | | | BIENNIUM2 = 1994-1995: (7.0/1.0)
| | | CATIDADE = 31-45
| | | | FACE = no: no (19.0/5.0)
| | | | FACE: yes: yes (3.0)
| | CHEST = yes: no (15.0/2.0)
| CATEG = 3.4
| | CAUSE = fall: yes (25.0/11.0)
| | CAUSE = running over
| | | MEMBERS = yes: no (9.0/3.0)
| | | MEMBERS = no: yes (22.0/7.0)
| | CAUSE = motorbike: yes (22.0/10.0)
| | CAUSE = driver
| | | STOMACH = no: no (14.0/2.0)
| | | STOMACH = yes: yes (6.0/1.0)
| | CAUSE = passenger: yes (6.0/1.0)
| | CAUSE = bike: no (1.0)
| | CAUSE = agression: yes (5.0/2.0)
| | CAUSE = other: no (4.0/1.0)
| | CAUSE = fire weapon: yes (0/0)
PUPIL = Isochoric: no (278.0/43.0)
PUPIL = mydriatic: yes (79.0/17.0)
PUPIL = miotic: no (28.0/7.0)
```

Fig.1 Classification model generated

C. Evaluation of Model

The models generated in step data mining were evaluated by applying the following metrics for evaluating performance: sensitivity, specificity, accuracy, error rate of the overall model, positive and reliability kappa index. Some results even being identified after the execution of data mining, can be used for parameters that informed the algorithms are redefined, rerunning them in order to get new results, which are revalued. This is quite common in the process, since it is interactive and iterative.

V. CONCLUSION

The TBI is one of the types of trauma that most affects the population, occurring severe injuries leading to hospitalization. Among the prognostic indicators of severe TBI, as the execution of data mining, it can be seen that one of the major causes that lead to death are traffic accidents. In this study it was concluded that some indicators are more related to death as the pupils, the low Glasgow Coma Scale, the presence of subarachnoid hemorrhage, absence of chest trauma and systemic diseases.

After the construction of a classification model, is evaluated by the test set the number of instances correctly classified, obtaining the accuracy. Outcome prediction was classified by C4.5 algorithm with an accuracy of 87%, including combinations of indicators that lead to survival and death.

Compared with other works, the study [13] performed with the same database of Brazilian patients in the city of Florianopolis, but using logistic regression achieved 76.9% of right predictions. While international studies, using the C4.5 algorithm for prediction of trauma as [16] and [4], obtained respectively 77% and 98.97% of accuracy.

In addition to C4.5, is intended to include other classifiers which will have their results compared with the C4.5, in order to obtain a tool for the prediction of prognosis in severe TBI.

ACKNOWLEDGMENT

The Institute of Biomedical Engineering, Department of Electrical Engineering, Federal University of Santa Catarina and the Course of Computer Science, University of the Extreme South of Santa Catarina.

REFERENCES

[1] T.T. Lee et al, “Application of data mining to the identification of critical factors in patient falls using a web-based reporting system”, International Journal of Medical Informatics, vol.80, pp.141–150, feb. 2011.

[2] K.E. Saatman et al, “ Classification of traumatic brain injury for targeted therapies”, Journal of Neurotrauma, vol. 25, pp. 719-738, jul. 2008.

[3] N. Sut and O. Simsek, “Comparison of regression tree data mining methods for prediction of mortality in head injury”, Expert Systems with Applications, vol. 38, pp. 15534-15539, nov./dec. 2011.

[4] E. M. Theodoraki et al, “Innovative data mining approaches for outcome prediction of trauma patients”, Journal Biomedical Science and Engineering, vol. 3, pp. 791-798, aug. 2010.

[5] G. Dolce et al, “Clinical signs and early prognosis in vegetative state: a decisional tree, data mining study”, Brain injury, vol. 22, pp. 617-623, jul. 2008.

[6] I. H. Witten, E. Frank and M. Hall, *Data mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, 2011.

[7] J. Han, M. Kamber and J. Pei, *Data mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2011.

[8] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. Chapman and Hall, New York, 2009.

[9] G. Dimitoglou, J.A. Adams and C. M. Jim, “ Comparison of the C4.5 and Naive Bayes classifier for the prediction of lung cancer survivability”, Journal of Computing, vol. 4, aug. 2012.

[10] P.N. Tan, M. Steinbach and V. Kumar, *Introdução ao Data Mining*. Ciência Moderna, Rio de Janeiro, 2009.

[11] J.R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 1993.

[12] M. Kantardzic, *Data mining: Concepts, Models, Methods and Algorithms*. John Wiley & Sons, New Jersey, 2011.

[13] A. Diament and S. Cypel, *Neurologia Infantil*. Atheneu, São Paulo.

[14] D.P. Graham and A.L. Cardon, “An update on substance use and treatment following traumatic brain injury”, Annals of the New York Academy of Sciences, vol. 1141, pp.148-162, oct. 2008.

[15] E.T. Martins et al, “Mortality in severe traumatic brain injury: a multivariated analysis of 748 Brazilian patients from Florianopolis City”, The Journal of Trauma, vol. 67, pp. 85-90, jul. 2009.

[16] T. Chesney et al, “Data mining trauma injury data using C5.0 and logistic regression to determine factors associated with death”, International Journal of Healthcare Technology and Management, vol. 10, p. 16-26, 2009.