CrossMark

Multiclassifier Systems for Predicting Neurological Outcome of Patients with Severe Trauma and Polytrauma in Intensive Care Units

Javier González-Robledo[1] · Félix Martín-González[1] · Mercedes Sánchez-Barba[2] ·
Fernando Sánchez-Hernández[3] · María N. Moreno-García[4]

Abstract This paper presents an ensemble based classification proposal for predicting neurological outcome of severely traumatized patients. The study comprises both the whole group of patients and a subgroup containing those patients suffering traumatic brain injury (TBI). Data was gathered from patients hospitalized in the Intensive Care Unit (ICU) of the University Hospital in Salamanca. Predictive models were induced from both epidemiologic and clinical variables taken at the emergency room and along the stay in the ICU. The large number of variables leads to a low accuracy in the classifiers even when feature selection methods are used. In addition, the presence of a much larger number of instances of one of the classes in the subgroup of TBI patients produces a significantly lesser precision for the minority class. Usual ways of dealing with the last problem is to use undersampling and oversampling strategies, which can lead to the loss of valuable data and overfitting problems respectively. Our proposal for dealing with these problems is based in the use of ensemble multiclassifiers as well as in the use of an ensemble playing the role of base classifier in multiclassifiers. The proposed strategy gave the best values of the selected quality measures (accuracy, precision, sensitivity, specificity, F-measure and area under the Receiver Operator Characteristic curve) as well as the closest values of precision for the two classes under study in the case of the classification from imbalanced data.

Introduction

The care of polytraumatized patients represents a challenge for health professionals aiming at decreasing its negative social and economic impact, which is much greater than other pathologies. Severe trauma is the primary cause of mortality of young adults in the world and the most influential as regards the years of potential life lost (YPLL). On the other hand, severe trauma has an important economic impact due to the high cost of the traumatic injury treatment [14]. For instance, in the United States this cost is greater than cancer and cardiovascular diseases treatment [22].

Survival estimation of critical patients has been traditionally made from scoring systems as APACHE II (Acute Physiology and Chronic Health Evaluation II) and other severity classifications, however, some studies have proved that these scores are not suitable for individual clinical outcome prediction [8]. Scoring systems specifically developed for trauma and injury severity evaluation also present some drawbacks, especially for polytraumatized patients since popular systems such as Trauma and Injury Severity Score (TRISS) do not provide reliable results for patients with a large number of injuries [26].

✉ María N. Moreno-García
  mmg@usal.es

[1] Intensive Care Unit, University Hospital of Salamanca, Salamanca, Spain

[2] Department of Statistics, University of Salamanca, Salamanca, Spain

[3] School of Nursing and Physiotherapy, Prehospital Emergency Services, University of Salamanca, Salamanca, Spain

[4] Department of Computing and Automation, University of Salamanca, Salamanca, Spain

Nowadays, new estimation models based on data mining techniques are being developed given that current technology affords the possibility of storing huge amounts of medical data. However, a drawback for the treatment of this information is the diversity and quantity of involved variables regarding the number of available records of patients. In this context of high dimensionality and not too much data instances, the behavior of predictive models is not good. Other important problem is the imbalance of the data that takes place when there is a big difference between the number of instances of every class under study. In these situations the precision for the minority class is usually significantly lesser than the precision for the majority class; therefore, predictive models are not valid even when they present an acceptable accuracy. Thus, in this work the way of dealing with these problems has been addressed. The final goal is to build reliable models that allow to predict the outcome of severely traumatized patients who require ICU hospitalization, taking into account epidemiological and clinical variables as well as those related to the healthcare management. The study is carried out to determine mortality or poor neurological outcome of patients of all trauma patients, on the one hand, and just the patients suffering TBI, on the other hand.

Related work

Several studies have been reported in the literature about factors influencing the evolution of critical patients. They have derived in the proposals of different prognostic and mortality scales described in [5]. One of the most important is based in APACHE II, however, there are few works about polytraumatized patients where APACHE II is studied as a factor of seriousness and some of them are contradictory [10]. TRISS (Trauma and Injury Severity Score) is a scoring method that allows to predict the survival probability of injured patients [2], but this method does not give accurate results for polytraumatized patients. Glasgow Coma Scale (GCS) is other prognostic indicator in most of the scales. Some studies confirm that GCS is the factor most correlated with the neurological component [20] and its association with pupil alteration is an indicator of mortality [15].

Statistical techniques are the usual tools used in these works since they allow identifying key indicators of neurological outcome while data mining techniques are used to a lesser extent. Some classical data mining methods, such as Bayesian Networks (BN), Support Vector Machines (SVM) or Artificial Neural Network (ANN) have been applied for diagnostic decision support in several medical fields [1, 7, 27]. In the context of traumatized patients these classical methods are also used. Studies addressed to predict survival of this kind of patients is given in [17, 26] where trauma data are treated by means of Neural Networks and Bayesian Decision Trees

respectively. The last study proved that this method provides better accuracy than the TRISS system.

Ensemble multiclassifiers are less extended than classical machine learning techniques in the medical field, although several papers can be found in the literature, especially oriented to diagnosis. A sample of this kind of works is summarized next. Random Forest is used to improve cardiac arrhythmia diagnosis in [23]. In [6] a predictive model for diagnostics of diabetes mellitus is induced by means of the Bagging algorithm. Several ensemble multiclassifiers have been applied for predicting success/failure of noninvasive mechanical ventilation in [21]. Feature selection techniques were used in this work for reducing the number of attributes and improving prediction reliability. Early detection of multidrug-resistant tuberculosis risk was the target of a work in which Bagging and Boosting algorithms were used. [19]. The problem of building classifiers from imbalanced data is addressed in the work making use of sampling strategies.

In this work, it was necessary to tackle high dimensionality and imbalanced data problems. Given that in our experiments feature selection methods did not work as expected, we applied several ensemble multiclassifiers and we studied their behavior against single classifiers. On the other hand, taking into account the problems of sampling strategies, we have proposed an alternative approach also based in the use of ensembles.

Data mining study

The study includes data from 497 patients diagnosed with severe trauma and polytrauma who were hospitalized in the Intensive Care Unit (ICU) of the University Hospital in Salamanca. The analysis was carried out taking into account the variables described in Online Resource 1. Given that the objective is to make an early prediction of neurological outcome, we used only 120 variables whose values can be obtained along the first 24 hours of stay in ICU.

In this study several classification algorithms were applied to induce models that would allow the class of unclassified patients to be predicted from other attributes of these patients. Two kinds of algorithms were used for inducing both single classifiers and multiclassifiers, specifically, ensemble multiclassifiers.

As commented before single classifiers sometimes do not achieve an acceptable accuracy with some training sets as the one managed in this work, which contain many attributes and the imbalance between negative and positive instances is very high.

The usual way to deal with high dimensionality problem is the application of feature selection methods in order to select the most influential attributes (features) for the classification and obtain more reliable models. We used two types of algorithms, on the one hand, some variants of the methods based

only on the information gain provided by the attributes in an individual way and, on the other hand, the CFS (Correlation-based Feature Subset Selection) [13] algorithm, which also considers the correlation between attributes. CFS evaluates the significance of a subset of features taking into account the individual predictive ability of each feature and the degree of redundancy between them. This method selects the subsets of attributes that are highly correlated with the class while having low inter-correlation between them.

In most of the cases better accuracy is achieved when the classifiers are induced from the selected attributes, however, in the application domain of this work such objective was not achieved since accuracy, instead of improving, was worse in some cases.

The other problem to be addressed was the treatment of imbalanced data. Oversampling of the minority class instances or undersampling of the majority class ones are two common approaches to deal with imbalanced datasets, but they have important drawbacks. Undersampling may discard potentially valuable data, while oversampling artificially increases the size of the data set and, as a result, the computational cost of inducing the models. In addition, the replication of existing examples in the minority class causes overfitting problems [16]. The special characteristics of multiclassifiers make them suitable methods to potentially solve both important drawbacks we have to deal with.

The description of the algorithms used in the comparative study is provided in the next subsections.

Single classifiers

Firstly two tree induction algorithms were employed in the study, J48 and REPTree. J48 is an advanced version of C4.5 [25], one of the algorithms most used and well known. J48 is an information gain-based method with pruning procedures that use rules. REPTree is a fast decision tree learning algorithm, also based on information gain, which uses reduced-error pruning with back-fitting. As a single classifier we also applied a Bayesian network algorithm that provides a graph relating variables and including distributions of conditional probabilities associated with these variables. The learning process for a dataset lies in finding, from among all possible graphs, the one that best represents the set of dependences/independences between data. The problem does not have an exact solution and it is necessary to resort to heuristic search methods. In this study, the K2 search algorithm [9] was applied.

The study includes also a support vector machine (SVM) method, SMO (Sequential Minimal Optimization) algorithm for training a support vector classifier [24]. SVM are linear classifiers based on finding the hyperplane that gives the largest minimum distance to the training examples.

Multiclassifiers

In order to address the referred peculiarities characteristic of the datasets (high dimensionality and imbalanced distribution) some multiclassifiers were applied. These methods combine several individual classifiers induced with different basic methods or obtained from different training datasets with the aim of improving the accuracy of the predictions. Another additional advantage of these techniques is the reduction of the overfitting problem, which takes place when the learning process finds a regularity in the data that is distinctive of the training set but cannot be extended to other datasets [18].

The methods for building multiclassifiers can be divided in two groups. The first, named ensemble methods, such as Bagging [3], Boosting [11] and Random Forest [4], induce models that merge classifiers with the same learning algorithm, but introducing modifications in the training data set. The second type of methods, named hybrids, such as Stacking [28] and Cascading [12], create new hybrid learning techniques from different base learning algorithms.

In this study the methods used were Random Forest, Bagging, AdaBoost, a variant of the Boosting method, and Random Committee.

Bagging is the acronym for Bootstrap AGGregatING. The method induces a multiclassifier that consists on an ensemble of classifiers built on bootstrap replicates of the training set. Majority vote is used to choice the final prediction.

Boosting is a multiclassifier of the same kind of Bagging, however, this method assigns weights to the outputs of the induced single classifiers from different training sets (strategies). In an iterative process, the weights are updated by increasing the weight of strategies with the correct prediction and reducing the weight of strategies with incorrect predictions. In this way the multiclassifier is developed incrementally, adding one classifier at a time. The classifier that joins the ensemble at step k is trained on a data set selectively sampled from the training data set Z. The sampling distribution starts from uniform, and progresses in each k step towards increasing the likelihood of worst classified data points at step k − 1. This algorithm is called AdaBoost which comes from ADAptive BOOSTing. AdaBoost presents the advantage of driving the ensemble training error to zero in very little iteration [18].

Random Forest [4] can be considered a multiclassifier similar to Bagging since it involves the induction of an ensemble of tree classifiers, each of which produces its own output. The induction of each tree is produced from a subset of the original data set chosen independently (with replacement) and with the same distribution for all trees in the forest. For classification problems, the most popular class obtained by simple vote is chosen as the final outcome.

Random Committee constructs an ensemble where each base classifier is built using a different random number seed.

The final prediction is a straight average of the predictions generated by the individual base classifiers.

Validation of the classifiers

Ten-fold cross-validation was used in the validation of all classifiers. Thus, the available data was divided into ten disjoint subsets with the same size. This method involves ten trainings that are performed by taking into each of them a different subset as test set and building the model with the remainder subsets. The error rate is the average of the errors obtained in the trainings performed.

The quality metrics computed were accuracy, precision, sensitivity (recall), specificity, F-measure and area under the ROC (Receiver Operator Characteristic) curve.

In many research works the validation of classifiers is carried out only by examining their accuracy, that is, the percentage of correctly classified instances. However, that measure can be complemented with others that provide additional error perspectives, especially when evaluating binary decision problems. In these cases, the examples are classified as either positive or negative and the output of the classifier can belong to one of the following four categories: True positives (TP) are positive instances correctly classified, false positives (FP) are negative instances classified as positive, true negatives (TN) are negative instances correctly classified and false negatives (FN) are positive instances classified as negative. Given this information, it is possible to define some validation metrics such as precision, sensitivity, specificity, F-measure or AUC (area under the ROC curve).
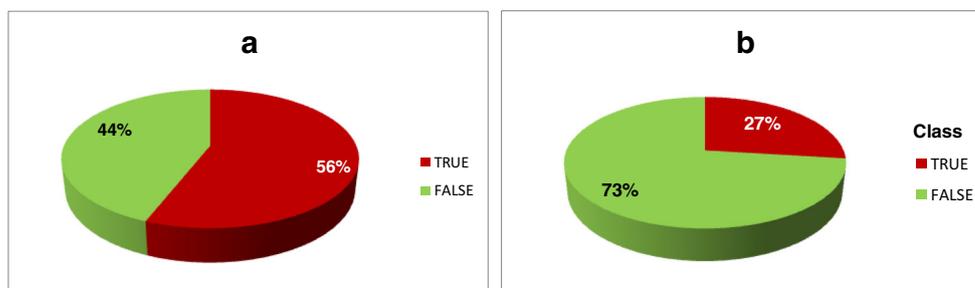
Precision is the probability that an example will be positive if the classifier classifies it as positive:

$$Precision = TP/(TP + FP) \qquad (1)$$

Sensitivity or recall refers to the probability of a positive example being classified as positive, which is the true positive rate (TPR).

$$TPR = TP/(TP + FN) \qquad (2)$$

Specificity however is referred to true negative rate (TNR).

$$TNR = TN/(TN + FP) \qquad (3)$$

A good classifier will provide precision and recall values as high as possible, but if this is not possible it must give the best possible balance between the two metrics. A metric that combines precision and recall is the F-measure.

$$F-measure = \alpha^* \left(\left(Precision^* Recall\right)/Precision + Recall\right)\right) \qquad (4)$$

The harmonic mean of precision and recall is obtained for $\alpha = 2$.

The ROC curve is the representation of the true positive rate (TPR) compared to the false positive rate (FPR). Point (0,0) of the ROC graph corresponds to a classifier that classifies all examples as negatives, and point (1,1) corresponds to a classifier that classifies all examples as positive. The best learning system will be the one that provides a set of classifiers with a greater area under the ROC curve (AUC).

Proposed strategy

The methodological approach proposed in this work is based in the use of multiclassifiers due to their special characteristics that make them suitable models for dealing with high dimensionality and imbalanced data drawbacks. For the last problem, the combination of two ensembles is proposed given that the combined random sampling strategies provided by both algorithms can help to balance the data without making use of undersampling or oversampling procedures.

Several reasons were taken into account in the choice of multiclassifiers to address the two problems found in our study. Firstly, the fact that multiclassifiers extend the hypothesis space with respect to single classifiers and achieve a better management of large number of attributes. On the other hand, multiclassifier algorithms reduce the overfitting of the induced models. This fact is essential to deal with imbalanced data problem.

In general, the key of the good behavior of classifiers ensemble is the diversity provided by different training sets.

Fig. 1 Percentage of instances of the *true* and *false* classes. **a**: All patients, **b**: Patients with TBI

Table 1 Quality metrics of the classifiers for the group of all patients

Algorithm	Accuracy	Sensitivity (recall)	Specificity	F-Measure	AUC
Decision tree J48	86,32%	0,903	0,813	0,863	0,853
Decision tree REPTree	85,71%	0,827	0,895	0,858	0,878
Bayes Net	75,86%	0,781	0,731	0,759	0,834
SVM	84,10%	0,845	0,836	0,841	0,840
Random Forest	77,87%	0,835	0,708	0,777	0,867
Bagging - J48	**86,92%**	**0,881**	**0,854**	**0,869**	**0,913**
Bagging - REPTree	**87,73%**	**0,860**	**0,909**	**0,878**	**0,915**
Bagging - SVM	84,91%	0,860	0,836	0,849	0,899
AdaBoost - J48	83,90%	0,860	0,813	0,839	0,909
AdaBoost - REPTree	85,31%	0,863	0,840	0,853	0,902
AdaBoost - SVM	81,29%	0,831	0,790	0,813	0,867
Random Commitee - Random Tree	79,48%	0,853	0,721	0,793	0,874
Random Commitee - Random Forest	83,50%	0,853	0,813	0,835	0,903

Bold entries represent the best results obtained from the methods evaluated

Significant improvements are achieved mainly when the base classifier is unstable, that is, small changes in the training set should lead to large changes in the classifier output [18]. This is the scenario that takes place when working with imbalanced data. In addition, majority vote properties assure the improvement of the single classifiers results.

Results and discussion

The study was divided in two parts corresponding on one hand to the whole group of patients and, on the other hand, to the subgroup of 282 patients having traumatic brain injury (TBI).

The class attribute to be predicted is "death or poor neurological outcome", which can take the values *true* and *false*. We considered a neurological outcome to be poor when the GOSE was equal to or less than 3.

As stated previously, owing to the large number of variables to be treated by the algorithms, several feature selection methods were applied in the preprocessing step, but accuracy did not improve. Consequently, we decided to use all attributes in the study.

An important aspect of the study is the distribution of examples belonging to each class (*true* and *false*) in the dataset (Fig. 1). When we considered the complete group of patients (Fig. 1-a) the proportion of instances of each class was similar, however, when we consider the subgroup of TBI patients this ratio was quite different. This unequal distribution can significantly influence the results of the data-mining algorithms, since high accuracy can be obtained but the rate of correctly classified instances of the minority class may be very low. In these cases, analysis of the results have to be focused not only on accuracy but also on the precision achieved for the two classes. In addition, the capture of other measures such as the F-measure and the ROC analysis can be useful.

Table 2 Quality metrics of the classifiers for the group of patients with TBI

Algorithm	Accuracy	Sensitivity (recall)	Specificity	F-Measure	AUC
Decision tree J48	79,79%	0,893	0,539	0,792	0,747
Decision tree REPTree	79,79%	0,922	0,461	0,784	0,747
Bayes Net	79,79%	0,854	**0,645**	0,799	0,821
SVM	80,85%	0,883	0,605	0,806	0,744
Random Forest	82,98%	0,966	0,461	0,812	0,834
Bagging - J48	79,79%	0,903	0,513	0,789	**0,866**
Bagging - REPTree	82,98%	0,971	0,447	0,810	0,826
Bagging - SVM	80,50%	0,893	0,566	0,800	0,823
AdaBoost - J48	79,43%	0,903	0,500	0,785	0,819
AdaBoost - REPTree	80,50%	0,908	0,526	0,797	0,815
AdaBoost - SVM	76,95%	0,825	0,618	0,773	0,767
Random Commitee - Random Tree	82,98%	0,947	0,513	0,817	0,810
Random Commitee - Random Forest	**84,40%**	**0,976**	0,487	**0,827**	**0,863**

Bold entries represent the best results obtained from the methods evaluated

Table 3 Results of the significance tests to compare accuracy of Bagging with J48 and Bagging with REPTree against the other classifiers for the dataset of all patients

Algorithm	Bagging - J48	Bagging - REPTree
Decision tree J48	**(1/0/0)**	**(1/0/0)**
Decision tree REPTree	(0/1/0)	**(1/0/0)**
Bayes Net	**(1/0/0)**	**(1/0/0)**
SVM	**(1/0/0)**	**(1/0/0)**
Random Forest	**(1/0/0)**	**(1/0/0)**
Bagging - J48	-	(0/1/0)
Bagging - REPTree	(0/1/0)	-
Bagging - SVM	**(1/0/0)**	**(1/0/0)**
AdaBoost - J48	**(1/0/0)**	**(1/0/0)**
AdaBoost - REPTree	**(1/0/0)**	**(1/0/0)**
AdaBoost - SVM	**(1/0/0)**	**(1/0/0)**
Random Committee - Random Tree	**(1/0/0)**	**(1/0/0)**
Random Committee - Random Forest	**(1/0/0)**	**(1/0/0)**

Bold entries represent the best results obtained from the methods evaluated

In order to do a comparative study about the performance of multiclassifiers against classical methods, we applied four simple classification algorithms, two decision trees, J48 and REPTree, a Bayes Net, and a SVM. The multiclassifiers used were Random Forest, Bagging, Adaboost and Random Committee, using several base classifiers.

The validation results obtained considering all patients are shown in Table 1. Table 2 contains the results of the subgroup of patients with TBI. In the first case the best values of all the selected quality metrics were provided by the Bagging algorithm using both J48 and REPTree as base classifiers, although with REPTree the results were slightly better. For the subgroup of patients with TBI, where the problem of imbalanced data is present, we proposed to apply Random Committee using other ensemble, Random Forest, as base classifier and,

as expected, the best values of the main metrics were provided by this combination.

The statistical significance of the results was evaluated by performing T-Tests using the experimenter tool of Weka. Tables 3 and 4 show a pair-wise comparison of the classifiers that yielded the best results against the other ones. The notation used is (x y z) where x = 1 indicates that the classifier in the column is significantly better than the one in the row, y = 1 means that both classifiers are similar and z = 1 indicates that the classifier in the column is worst. The established level of significance was 0.05.

For the dataset of all patients, T-Test results (Table 3) confirm that Bagging with J48 and Bagging with REPTree outperform the other classifiers and both have a similar behavior regarding accuracy, which is the most important measure when data are not imbalanced.

Table 4 Results of the significance tests for Random Committee with Random Forest against the other classifiers for the dataset of patients with TBI

Algorithm	Random Committee - Random Forest				
	Accuracy	Sensitivity (recall)	Specificity	F-Measure	AUC
Decision tree J48	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
Decision tree REPTree	(0/1/0)	**(1/0/0)**	(0/1/0)	**(1/0/0)**	**(1/0/0)**
Bayes Net	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
SVM	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
Random Forest	(0/1/0)	(0/1/0)	(0/1/0)	**(1/0/0)**	**(1/0/0)**
Bagging - J48	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	(0/1/0)
Bagging - REPTree	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
Bagging - SVM	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
AdaBoost - J48	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
AdaBoost - REPTree	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
AdaBoost - SVM	**(1/0/0)**	**(1/0/0)**	(0/0/1)	**(1/0/0)**	**(1/0/0)**
Random Committee - Random Tree	**(1/0/0)**	**(1/0/0)**	(0/1/0)	**(1/0/0)**	**(1/0/0)**

Bold entries represent the best results obtained from the methods evaluated

Fig. 2 Results obtained for the complete group of patients
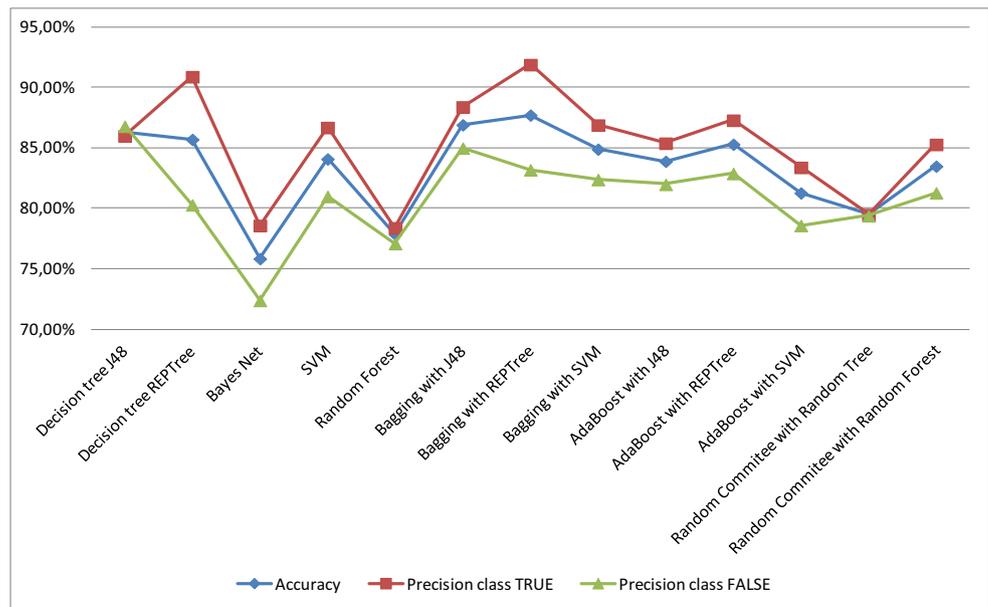
T-tests also evidence that Random Committee with Random Forest is the best classifier for the group of patients with TBI where there is data imbalance, as can be seen in Table 4. For all quality metrics, except specificity, this algorithm is significantly better than the other analyzed methods.

In order to analyze the behavior of the classifiers regarding both classes, true and false, we examined the precision of each of them. Figures 2 and 3 show the values of accuracy, precision of the class true and precision of the class false for the group of all patients and for the subgroup of patients with TBI respectively. The difference between the precision of both classes achieved by most of the classifiers is not very significant for the complete group (Fig. 2), nevertheless, this difference for the TBI subgroup (Fig. 3), which has a minority class, is very pronounced except for three ensembles: Random Forest, Bagging with REPTree and the ensemble of two multiclassifiers Random Committee with Random Forest.

Conclusions

This study addresses the problem of predicting death or a poor neurological outcome in patients affected by severe trauma and polytrauma, one of the pathologies with the greatest impact in today's society. Several data mining algorithms were applied to the data from 497 traumatized and polytraumatized patients

Fig. 3 Results obtained for the subgroup of patients with TBI
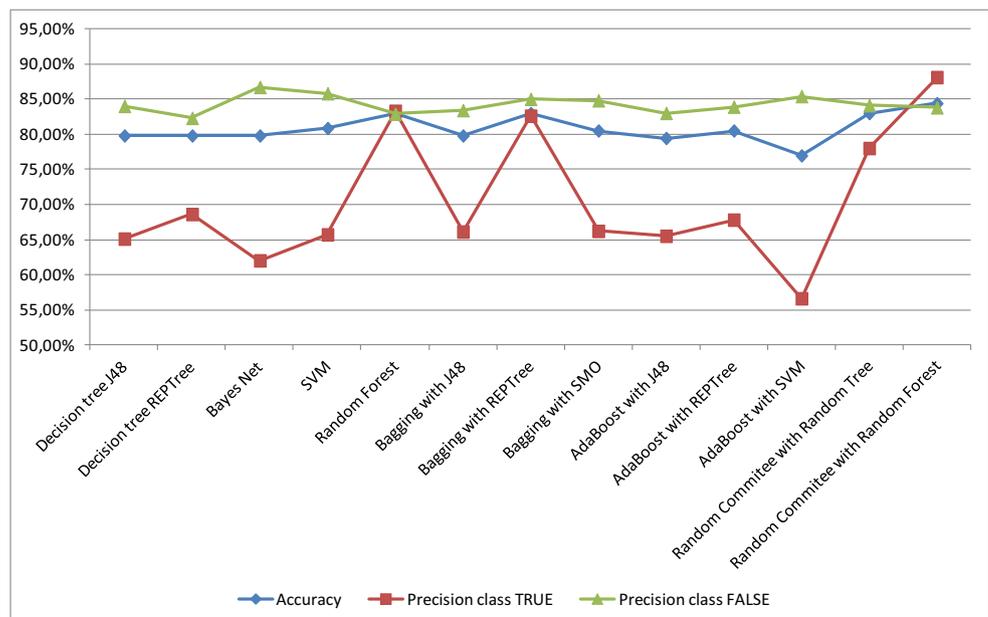
who required ICU hospitalization. The input variables to the algorithms were both epidemiologic and clinical and were taken at the emergency room and along the stay in the ICU.

The study was divided into two parts, corresponding to two different groups of patients, the first one containing all patients and the second one containing only the patients with TBI. In the first case, the best behavior was seen for the Bagging multiclassifier, achieving an accuracy of 86.92% when J48 was used as the base classifier and 87.73% when the base classifier was REPTree. In the second part of the study we detected the drawback of an imbalanced data distribution, which usually affords greater precision for the majority class than for the minority one. Accordingly, besides accuracy, it is necessary to consider the precision for all the classes in the study. In this situation, the best results were provided by Random Committee with Random Forest, our proposal of combining two multiclassifiers, which achieved an accuracy of 84.40% and very close values of precision for the true and false classes.

In general, the results point to the better behavior of multiclassifiers as compared with simple classifiers in contexts of high dimensionality and imbalanced datasets, without any need to resort to oversampling and undersampling strategies.

The high performance of the models induced by these algorithms from data obtained in the first hours of the stay in the ICU allows a reliable initial prediction of patient outcome. Therefore, the clinical repercussion of this study can be very useful since it involves the identification of new tools and factors to be incorporated in future studies about prognostic scores, which facilitate the early classification of patients according to their severity and help in that way in the decision making about the treatment levels that can be offered to such patients.

References

1. Berikol, G.B., Yildiz, O., and Özcan, I.T., Diagnosis of acute coronary syndrome with a support vector machine. *J. Med. Syst.* 40:84, 2016.
2. Boyd, C.R., Tolson, M.A., and Copes, W.S., Evaluating trauma care: The TRISS method. *J. Trauma.* 27:370–378, 1984.
3. Breiman, L., Bagging predictors. *Mach. Learn.* 24(2):123–140, 1996.
4. Breiman, L., Random forests. *Mach. Learn.* 45(1):5–32, 2001.
5. Chawda, M.N., Hildebrand, F., Pape, H.C., and Giannoudis, P.V., Predicting outcome after multiple trauma: Which scoring system. *Injury.* 35(4):347–358, 2004.
6. Chen, H., Tan, C., Lin, Z., and Wu, T., The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. *Comput. Biol. Med.* 50(2014):70–75, 2014.
7. Chen, D., Wang, H., Sheng, L., Hueman, M.T., Henson, D.E., Schwartz, A.M., and Patel, J.A., An algorithm for creating prognostic Systems for Cancer. *J. Med. Syst.* 40:160, 2016.
8. Cohen, N.H., The real reasons not to rely on severity scores. *Crit. Care Med.* 38(2010):334–335, 2010.
9. Cooper, G.F., and Herskovits, E., A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* 9:309–347, 1992.
10. Dossett, L.A., Redhage, L.A., Sawyer, R.G., and May, A.K., Revisiting the validity of APACHE II in the trauma ICU: Improved risk stratification in critically injured adults. *Injury.* 40(9):993–998, 2009.
11. Freund, Y., and Schapire, R.E., Experiments with a new boosting algorithm. *Proceedings 13th International Conference on Machine Learning*:148–156, 1996.
12. Gama, J., and Brazdil, P., Cascade Generalization. *Mach. Learn.* 41(3):315–343, 2000.
13. Hall, M. A. Correlation-based Feature Selection for Machine Learning. PhD Thesis, University of Waikato, Hamilton, Nueva Zelanda, 1999.
14. Hemmila, M. R., Jakubus, J. L., Maggio, P. M., et al. Real money: Complications and hospital costs in trauma patients. Surgery 144(2):307–316, 2008.
15. Hoffmann, M., Lefering, R., Rueger, J.M., Kolb, J.P., Izbicki, J.R., Ruecker, A.H., Rupprechtz, M., and Lehmann, W., Pupil evaluation in addition to Glasgow coma scale components in prediction of traumatic brain injury and mortality. *Br J Surg.* 99(1):122–130, 2012.
16. Hulse, J., Khoshgoftaar, T., and Napolitano, A., Experimental perspectives on learning from imbalanced data. *Proceedings of the 24th International Conference on Machine learning*:935–942, 2007.
17. Hunter, A., Kennedy, L., Henry, J., and Ferguson, I., Application of neural networks and sensitivity analysis to improved prediction of trauma survival. *Comput. Methods Prog. Biomed.* 62(2000):11–19, 2000.
18. Kuncheva, L.I., *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons, 2004.
19. Li, S., Tang, B., and He, H., An imbalanced learning based MDR-TB early warning system. *J. Med. Syst.* 40:164, 2016.
20. Marmarou, A., Lu, J., Butcher, I., et al., Prognostic value of the Glasgow coma scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: An IMPACT análisis. *J. Neurotrauma.* 24(2):270–280, 2007.
21. Martín, F., González, J., Sánchez, F., and Moreno, M.N., Success/failure prediction of noninvasive mechanical ventilation in intensive care units. Using multiclassifiers and feature selection methods. *Methods Inf. Med.* 55(3):234–241, 2016.
22. NAEMT, *PHTLS: Prehospital Trauma Life Support*. Mosby/JEMS, 2010. isbn:0-323-06502-3.
23. Özcift, A., Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Comput. Biol. Med.* 41(2011):265–271, 2011.
24. Platt, J. Fast. Training of support vector machines using sequential minimal optimization. In: Schoelkopf B, Burges C and Smola A (Eds.), Advances in Kernel Methods. Support Vector Learning. MIT Press, 1998.
25. Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA., USA, 1993.
26. Schetinin, V., Jakaite, L., Jakaitis, J., and Krzanowski, W., Bayesian decision trees for predicting survival of patients: A study on the US national trauma data bank. *Comput. Methods Prog. Biomed.* 111(2013):602–612, 2013.
27. Wagholikar, K.B., Sundararajan, V., and Deshpande, A.W., Modeling paradigms for medical diagnostic decision support: A survey and future directions. *J. Med. Syst.* 36:3029–3049, 2012.
28. Wolpert, D., and Stacked, H., Generalization. *Neural Netw.* 5:241–259, 1992.