Machine Learning Methods for Mortality Prediction of Polytraumatized Patients in Intensive Care Units – Dealing with Imbalanced and High-Dimensional Data

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

[1] Department of Computing and Automation, University of Salamanca, Salamanca, Spain
Department of Computing and Automation, Plaza de los Caídos s/n, 37008 Salamanca
mmg@usal.es
[2] Intensive Care Unit, University Hospital of Salamanca, Salamanca, Spain
[3] School of Nursing and Physiotherapy, University of Salamanca,
Prehospital Emergency Services, Salamanca, Spain
[4] Department of Statistics, University of Salamanca, Salamanca, Spain

Abstract. The aim of this study is the prediction of death of polytraumatized patients based on epidemiological, clinical and health treatment variables by means of data-mining methods. The main problems to be addressed were high dimensionality and imbalanced data. Since the techniques usually used to deal with these drawbacks, as feature selection methods and sampling strategies respectively, did not provided satisfactory results, the aim of the study was to find out the data mining algorithms showing the best behavior in this kind of scenarios. The study was carried out with data from 497 patients diagnosed with severe trauma who were hospitalized in the Intensive Care Unit (ICU) of the University Hospital of Salamanca. The results of the study reveal the better behavior of multiclassifiers as compared with simple classifiers in contexts of high dimensionality and imbalanced datasets, without the need to resort to undersampling and oversampling strategies, which can lead to the loss of valuable data and overfitting problems respectively.

Keywords: Severe trauma, polytrauma, mortality, data mining, classifiers, multiclassifiers.

1 Introduction

Severe trauma is considered to be one of the pathologies with the greatest impact on current society from the point of view of health as well as from the economic perspective. It 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). Regarding the economical aspect, it has been reported that the average economic cost for the treatment of traumatic injury in the United States is greater than the treatment of cancer and cardiovascular diseases [8].

* Corresponding author.

E. Corchado et al. (Eds.): IDEAL 2014, LNCS 8669, pp. 309–317, 2014.

The care of polytraumatized patients represents a challenge for current society and, in particular, for health professionals seeking to decrease its negative social and economical impact as well as personal consequences for the patient as much as possible. Current technology affords the possibility of storing huge amounts of medical data as electronic health records (EHRs), which can be processed by advanced techniques, such as data mining algorithms, to obtain useful knowledge on which decisions can be based. However, an important problem to be addressed is the great quantity and variety of variables that is necessary to take into account, from demographic data to clinical variables as well as those related to the healthcare management.

One usual way to deal with that drawback is the application of feature selection methods to know the best attributes for classification in order to use them for building the predictive models, discarding the remainder ones. In most of the cases a better accuracy is achieved when the algorithms work with the selected attributes, however, in the application domain considered in this study these techniques did not provided good results since the accuracy was not improved and in some experiments it became even worse. Two well-known and widely used algorithms whose efficacy has been demonstrated have been applied: CFS (Correlation-based Feature Subset Selection) [6] and a method based on information gain (IG) with respect to the class [7]. On the other hand, an additional problem to be addressed is the treatment of imbalanced data since the number of records belonging to one class is much greater than the number of records of the other class. In this kind of scenarios machine learning algorithms can achieve an acceptable global accuracy but usually the precision for the minority class is low. Oversampling of the minority class records or undersampling of the majority class records 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 [9].

The aim of the present study is to apply suitable machine learning algorithms, paying a special attention to multiclassifiers, in order to overcome the mentioned problems. Data mining techniques have been successfully used to infer knowledge in very diverse medical areas; however, in spite of their great interest and promising results, these methods have not yet been exploited in the specific domain of politraumatized patient treatment in intensive care units. In this study they are used to predict the final outcome of these patients.

2 Background

This section is devoted to expose some basic aspects of multiclassifiers that can help to understand their general better behavior against single classifiers. The book of Kuncheva [10] has been taken as reference to develop the contents of sections 2 and 3.

Multiclasifiers 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. The methods for building multiclassifiers can be divided