Applying Resampling Methods for Imbalanced Datasets to Not So Imbalanced Datasets

Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, and Jesús María Pérez

Computer Science Faculty, University of the Basque Country (UPV/EHU),
Manuel Lardizabal 1, 20018 Donostia, Spain
{olatz.arbelaitz,i.gurrutxaga,j.muguerza,txus.perez}@ehu.es
http://www.sc.ehu.es/aldapa

Abstract. Many efforts have been done recently proposing new intelligent resampling methods as a way to solve class imbalance problems; one of the main challenges of the machine learning community nowadays. Usually the purpose of these methods is to balance the classes. However, there are works in the literature showing that those methods can also be suitable to change the class distribution of not so imbalanced and even balanced databases, to a distribution different to 50% and significantly improve the outcome of the learning process. The aim of this paper is to analyse which resampling methods are the most competitive in this context. Experiments have been performed using 29 databases, 8 different resampling methods and two learning algorithms, and have been evaluated using AUC performance metric and statistical tests. The results show that SMOTE, the well-known intelligent resampling method, is one of the best candidates to be used, improving the results obtained by some of its variants that are successful in the context of class imbalance.

Keywords: Optimal class distribution, class imbalance problems, resampling methods, SMOTE.

1 Introduction

Class imbalance problem is considered one of the emerging challenges in the machine learning area [11, 17, 23]. In class imbalance problems, the number of examples of one class (minority class) is much smaller than the number of examples of the other classes, with the minority class being the class of greatest interest and that with the biggest error cost from the point of view of learning.

One of the approaches used to deal with class imbalance problems, called data approach, consists of resampling (subsampling or oversampling) the data in order to balance the classes before building the classifier. This approach is independent of the learning algorithm used and most of the research has been done in this direction [5, 9, 18]. One of the most popular data approaches is SMOTE [6]: an intelligent oversampling technique to synthetically generate more minority class examples. A broad analysis and comparison of some variants can be found in [4, 13].

C. Bielza et al. (Eds.): CAEPIA 2013, LNAI 8109, pp. 111–120, 2013.

Although resampling methods are usually addressed to solve class imbalance problems, Weiss and Provost [20] showed that there is usually a class distribution different to that appearing in the data set, with which better results are obtained.

Based on Weiss and Provost's work, Albisua et al. [1] confirmed that changes in the class distribution of the training samples improve the performance of the classifiers. However, in contrast to what Weiss and Provost pointed out in their work, they found that the optimal class distribution depends on the learning algorithm used (even if there are decision tree learners using the same split criteria, such as C4.5 and CTC) and also on whether or not the trees are pruned. Later, the same authors proposed in [2] an approach for enhancing the effectiveness of the learning process that combines the use of resampling methods with the optimal class distribution (instead of balancing the classes). It should be noted that the use of this approach is not restricted to imbalanced data sets but can be applied to any data set (imbalanced or not) in order to improve the results of the learning process. The authors demonstrated that 50% is not always the optimal class distribution even when intelligent resampling methods are used. The authors proposed a methodology able to find a class distribution that obtains better results than the balanced one with statistically significant differences (in many cases) for eight resampling methods and two learning algorithms. The experiments described in their work confirm that an optimal class distribution exists, but that it depends not only on the data's characteristics but also on the algorithm and on the resampling method used.

However, in the mentioned work there is a question that remains unanswered and we will try to answer in this work: when using C4.5 and PART to solve a real world problem, which of the 8 evaluated resampling methods and class distribution are the best for the concrete problem?

The work presented in this paper tries to answer the previous question based on experiments performed with 29 real problems (balanced and imbalanced ones) extracted from the UCI Repository benchmark [3] using 8 different resampling methods, C4.5 and PART algorithms and the AUC performance measure. For estimating performance we used a 10-fold cross-validation methodology executed five times (5x10CV). Finally, we used the non-parametric statistical tests proposed by Demšar in [8] and García et al. in [14] and [15] to evaluate the statistical significance of the results.

Section 2 provides a brief description of the resampling methods, algorithms and performance metric to be used. In Section 3 we describe the experimental methodology used to corroborate the previously mentioned hypothesis and in Section 4 we present an analysis of the experimental results. Finally, in Section 5 we summarize the conclusions and suggest further work.

2 Resampling Methods, Algorithms and Performance Metrics

In this section we briefly describe some of most popular and interesting resampling methods used to tackle the class imbalance problem found in bibliography, the two algoritms (C4.5 and PART) and the performance metric used to evaluate the classifiers.