Neurocomputing 135 (2014) 32–41

Contents lists available at ScienceDirect

Neurocomputing

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

A method for resampling imbalanced datasets in binary classification tasks for real-world problems

Silvia Cateni *, Valentina Colla, Marco Vannucci

TeCIP Institute, Scuola Superiore Sant'Anna, Pisa, Italy

ARTICLE INFO

Article history:
Received 25 October 2012
Received in revised form
19 February 2013
Accepted 30 May 2013
Available online 3 January 2014

Keywords:
Oversampling
Undersampling
Imbalanced dataset

ABSTRACT

The paper presents a novel resampling method for binary classification problems on imbalanced datasets. Imbalanced datasets are frequently found in many industrial applications: for instance, the occurrence of particular product defects, the diagnosis of severe diseases in a series of patients or machine faults are rare events whose detection is of utmost importance. In this paper a new resampling method is proposed combining an oversampling and an undersampling technique. Several tests have been developed aiming at assessing the efficiency of the proposed method. Four classifiers based, respectively, on Support Vector Machine, Decision Tree, labelled Self-Organizing Map and Bayesian Classifiers have been developed and applied for binary classification on the following four datasets: a synthetic dataset, a widely used public dataset and two datasets coming from industrial applications. The results that have been obtained in the tests are presented and discussed in the paper; in particular, the performances that are achieved by the four classifiers through the proposed novel resampling approach have been compared to the ones that are obtained, without any resampling, through a widely applied and well known resampling technique, i.e. the classical SMOTE approach, and through another approach coupling informed SMOTE-based oversampling and informed clustering-based undersampling.

1. Introduction

In the literature and in real-world problems, binary classification often needs to be performed on a so-called *imbalanced dataset*, where the samples belonging to one of the two classes to detect are far less numerous than the other ones. Often, the rare patterns are also difficult to separate from the most frequent ones. However, in many cases, rare events within a dataset are the most important ones to detect, such as, for instance, in the medical field for disease diagnosis or in the industrial field for fault diagnosis.

The class imbalance negatively affects the performances of all the common classifiers, such as support vector machines, decision trees and neural networks. Actually, the standard versions of these classifiers are designed to optimize the overall performance [1] on the whole dataset. As a consequence, they correctly classify most of the patterns belonging to the frequent class but they obtain a very poor performance on the other ones. This undesirable result is caused by the rarity of the minority class, which compromises the correct characterization of its patterns and prevents the correct separation from the majority class. As shown in many review works in the literature [2,3], any classifier is sensitive to the effect of imbalance in datasets, which affects not only the correct identification of rare patterns but also the rate of overall correct classifications.

Due to its significance, the problem of classifying imbalanced datasets has been faced in many literature works. The methods developed for coping with class imbalance can be divided into two main groups on the basis of the different approaches to the problem: the *internal methods* deal with the development of new algorithms expressly designed to cope with uneven datasets, while the *external methods* operate on the dataset to be used for the training of the classifier in order to attenuate the imbalance of the dataset.

Among internal methods it is worth to mention a class of approaches based on the use of different misclassification weights during the training of the classifiers [4]. In this framework the costs of misclassifying a rare pattern are higher with respect to other kinds of errors in order to encourage their correct detection. Support Vector Machines (SVMs) have also been employed for facing imbalanced datasets: in [5] an ensemble of SVM is used, while in [6] a variation of traditional SVM architecture (the so-called v-SVM) is used for the detection of rare samples: in this latter case the v-SVM is trained by using only rare patterns so as to allow their sole identification in the classification stage. A particular kind of radial basis function has been developed and tested in [7] where hyper-rectangular activation function neurons are used in the hidden layer in order to achieve more precision in the

* Corresponding author at: Valentina Colla, PERCRO laboratory, Via Alamanni 13D, 56010 San Giuliano Terme, Pisa, Italy. Tel.: +39 050 882507.
E-mail addresses: s.cateni@sssup.it (S. Cateni), colla@sssup.it (V. Colla), mvannucci@sssup.it (M. Vannucci).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.05.059

detection of the boundary of the input space regions reserved to each class. Finally in [8] the approach based on the use of an uneven cost matrix is combined to a self-organizing map and a fuzzy inference system obtaining interesting results in the literature and industrial datasets.

The external methods modify the distribution of rare and frequent patterns within the database in order to favour the detection of the rare ones. This operation is called *resampling* and aims at increasing the rate of rare samples by creating a new dataset from the original one. This rebalance can be obtained by both removing samples belonging to the frequent class and adding samples to the rare one. These two techniques, which can be eventually combined such as in [9], are called *undersampling* and *oversampling*. Through undersampling, some frequent samples are removed from the original dataset until a predetermined balance ratio is reached. Data to be removed can be selected in a random way or, more fruitfully, according to particular criteria such as, for instance, the removal of those patterns lying on the external regions of the input space. This latter technique effectively reduces the input space area assigned to the frequent class by the classifier and, as a consequence, favours the correct classification of rare patterns [10].

On the other hand, the oversampling methods counteract the original database unbalance by adding new samples of the rare class. This operation can be done by both replicating existing rare patterns and creating new ones in a particular region of the input space. The replication can be performed in a random way or by selecting those patterns lying on the boundary between rare and frequent samples in order to force the classifier to allocate such regions of the space to the rare class. The main criticality of the replication-based oversampling is that it does not add any new informative content to the dataset but it simply rebalances the ratio between rare and frequent samples. In order to overcome this drawback, several techniques creating new synthetic rare samples in those regions where they likely could be have been developed.

In this paper a new resampling method combining oversampling and undersampling is presented which is named *Similarity-based UnderSampling and Normal Distribution-based Oversampling* (SUNDO) methods. This method combines the two approaches: for the oversampling phase, it places new samples where they likely could be and avoids to place them close to frequent samples; moreover it employs an innovative undersampling technique.

The paper is organized as follows. Section 2 presents an overview of the literature survey, the general method is described in Section 3 and the details of oversampling and undersampling are provided in Sections 3.1 and 3.2, respectively, while Section 3.3 depicts through a simple example the advantages of SUNDO over some widely applied traditional approaches. The results achieved by SUNDO on several public and industrial imbalanced datasets are reported and discussed in Section 4, where they are compared to the ones achieved by other standard resampling techniques. Finally, in Section 5 some conclusions are drawn.

This paper is an extended version of the paper entitled “Novel resampling method for the classification of imbalanced datasets for industrial and other real-world problems” which was presented by the same authors at the 11th International Conference on Intelligent Systems Design and Applications ISDA 2011.

2. State of art

Imbalanced datasets occur in several real-world applications where the decision system is aimed to detect rare cases such as telecommunication customers, detection of oil spills in radar images, text classification and many others [11–13]. There are many different forms of resampling such as random oversampling, random undersampling, directed oversampling, directed undersampling and combinations of the two techniques.

Random undersampling [14] balances the minority class through the random elimination of some samples belonging to the majority class. This approach has been criticized in some papers because potentially useful data could be removed leading to a huge information loss [15–19]. Another disadvantage with this method is that often the aim of machine learning is for the classifier to assess the probability distribution of the target.

Kubat and Matwin [20] proposed a selective undersampling method keeping the original samples belonging to the minority class. They used a geometric mean as a performance test for the classifier. In particular the samples belonging to the minority class were categorized into four parts: noise overlapping the minority class, borderline samples, redundant samples and safe samples. Borderline samples are selected through the *Tomek link* notion [21]. Given two samples s_i and s_j belonging to different classes, a pair (s_i, s_j) is called a Tomek link if there is not a sample s_w such that $d(s_i, s_w) < d(s_i, s_j)$ or $d(s_j, s_w) < d(s_i, s_j)$ where $d(\cdot,\cdot)$ is the *distance* between the two considered samples. When two samples form a Tomek link, one of these samples is noise or both samples are borderline. The idea of this undersampling method is to delete samples belonging to the majority class that are distant from the decision border in order to eliminate samples less relevant for learning.

Random oversampling consists in balancing the minority class by randomly replicating its samples. The main disadvantage of this approach is the enhancement of occurring overfitting, since it creates coincident samples. Moreover the overfitting due to the addition of new samples increases the computational requirements when the dataset is already large and imbalanced.

SMOTE [22] is the most famous oversampling method. The basic idea of this algorithm is to place the new samples along the lines connecting existing rare samples. SMOTE has been widely used in literature works and often combined with other techniques such as in [23], where it is coupled with an ad hoc version of Support Vector Machine (SVM). Moreover several other methods follow the basic idea of recreating new rare samples in those regions of the space where they would be more likely to be found. For instance in [24] new samples are created in the surrounding areas of rare observations.

3. The proposed resampling method

The method that is proposed here combines undersampling and oversampling techniques in order to obtain a balanced dataset without significant loss of information and without the addition of a great number of synthetic patterns.

Firstly the imbalanced dataset is divided into a training set (75%) and a validation set (25%) maintaining the same proportion between the two classes. Then the training set is divided into two subsets containing the samples belonging to the minority class and to the majority class. Let us suppose that n_1 is the number of samples belonging to the minority class subset of the training dataset and n_0 is the number of the samples included in the majority class subset of the training dataset. Moreover let us suppose that k_0 and k_1 are, respectively, the target percentages of samples to be achieved by the majority and minority classes: obviously $k_0 + k_1 = 1$. These parameters may vary from their initial value of imbalance until to obtain a perfect equilibrium between the two classes, which means that they have the same number of samples. The target of 50% of samples has been chosen as a limit value for the minority class, as the assumption has been made that the majority class should not lose its primacy, which reflects