Neurocomputing 175 (2016) 935–947

Contents lists available at ScienceDirect

Neurocomputing

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

Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases[☆]

Octavio Loyola-González [a,b,*], José Fco. Martínez-Trinidad [b], Jesús Ariel Carrasco-Ochoa [b], Milton García-Borroto [c]

[a] *Centro de Bioplantas, Universidad de Ciego de Ávila, Carretera a Morón km 9, Ciego de Ávila, C.P. 69450, Cuba*
[b] *Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Sta. María Tonanzintla, Puebla, C.P. 72840, Mexico*
[c] *Instituto Superior Politécnico José Antonio Echeverría, Calle 114 No. 11901, Marianao, La Habana, C.P. 19390, Cuba*

ARTICLE INFO

Article history:
Received 28 November 2014
Received in revised form
19 March 2015
Accepted 20 April 2015
Available online 10 November 2015

Keywords:
Supervised classification
Resampling methods
Imbalanced databases
Contrast patterns

ABSTRACT

The class imbalance problem is a challenge in supervised classification, since many classifiers are sensitive to class distribution, biasing their prediction towards the majority class. Usually, in imbalanced databases, contrast pattern miners extract a very large collection of patterns from the majority class but only a few patterns (or none) from the minority class. It causes that minority class objects have low support and they could be identified as noise and consequently discarded by the contrast pattern based classifier biasing the results towards the majority class. In the literature, the class imbalance problem is commonly faced by applying resampling methods. Therefore, in this paper, we present a study about the impact of using resampling methods for improving the performance of contrast pattern based classifiers in class imbalance problems. Experimental results using standard imbalanced databases show that there are statistically significant differences between using the classifier before and after applying resampling methods. Moreover, from this study, we provide a guide based on the class imbalance ratio for selecting a resampling method that jointly with a contrast pattern based classifier allows us to have good results in a class imbalance problem.

1. Introduction

In many supervised classification tasks, a high accuracy is not the only desirable aim. Additionally, the classifier, as well as its results, should be understandable by the experts in the application domain [1]. For this reason, an option is to build supervised classifiers based on patterns extracted from a training set, since in this way the classification results can be explained through the patterns associated to each class.

In some supervised classification problems, the objects in each class are not equally distributed. Often, the most important class contains significantly less objects because it could be associated to rare cases [2]. This type of problems is known as the class imbalance problem or problems with imbalanced databases.

In recent years, the class imbalance problem has been addressed with high interest by the scientific community because it appears in several real-world applications. For example, online banking fraud detection is one of the problems with high class imbalance. In several online banking companies, their databases have five fraudulent transactions from approximately 300,000 transactions in a day [3]. Other examples of imbalanced databases can be found into the medical field for detection of microcalcifications in mammogram images [4], nosocomial infections [5], liver and pancreas disorders [6], and many others. Also, imbalanced datasets appear in other applications as prediction of protein sequences [7], forecasting of ozone levels [8], and face recognition [9].

Some contrast pattern based classifiers, which show good behavior in problems with balanced classes, do not necessarily achieve good performance in class imbalance problems. There are several reasons behind this behavior [10,11], some causes are:

(a) Patterns that predict the minority class are often highly specific and thus their support is very low, hence they are prone

[☆] This work was partly supported by National Council of Science and Technology of Mexico under the Project CB2008-106366 and Scholarship Grant 370272.

[*] Corresponding author at: Centro de Bioplantas, Universidad de Ciego de Ávila, Carretera a Morón km 9, Ciego de Ávila, C.P. 69450, Cuba. Tel.: +52 222 4646918, +53 54140710.

E-mail addresses: octavioloyola@bioplantas.cu (O. Loyola-González), fmartine@inaoep.mx (J.Fco. Martínez-Trinidad), ariel@inaoep.mx (J.A. Carrasco-Ochoa), mgarciab@ceis.cujae.edu.cu (M. García-Borroto).

http://dx.doi.org/10.1016/j.neucom.2015.04.120
0925-2312/© 2015 Elsevier B.V. All rights reserved.

to be discarded in favor of more general patterns that predict the majority class.

(b) Many approaches, like divide-and-conquer, produce object fragmentation into small partitions, which contain even less objects from the minority class, which makes more difficult the extraction of regularities.

(c) The use of global performance measures for guiding the learning process, such as the standard accuracy rate, may bias the classification results towards the majority class.

(d) Some minority class objects can be identified as noise, and therefore they could be wrongly discarded by the classifier. Conversely, some actual noisy objects can degrade the identification of the minority class, since it has only a few objects to train.

In this paper, our first goal is to come up with a study about the effects of the most used resampling methods for improving the accuracy achieved by a contrast pattern based classifier over several imbalanced databases. Our second goal is an analysis based on the experimental results that allows to choose a priori which resampling approach would be the best option regarding the class imbalance ratio.

Preliminary results of this study were reported in a conference paper [12]. The main differences between this paper and the conference paper are the following: this paper contains a more complete analysis of previous works since we have included many more resampling methods than in the conference paper; we add many more databases with higher imbalance ratio between the classes; and finally we include an analysis of the behavior of the resampling methods with the aim of providing a rough guide for selecting which would be the best resampling method regarding the class imbalance ratio.

This paper is organized as follows. Section 2 explains the difficulties that arise in contrast pattern based classifiers when they face class imbalance problems. Moreover, it contains a brief description of the main approaches reported in the literature to deal with class imbalance problems. After, our experimental methodology is presented in Section 3, whereas all results and discussion are shown in Section 4. Finally, Section 5 summarizes our conclusions and future work.

2. Related work

Mining contrast patterns and supervised classification using contrast patterns are challenging problems because of the high computational cost due to the exponential number of candidate patterns [13]. However, these problems become a major challenge on imbalanced databases because there are several problems that arise when contrast patterns are mined from imbalanced databases [14]. First, in this type of problems the number of objects belonging to a class (majority class) is significantly higher than the number of objects belonging to another class (minority class). Nevertheless, often the minority class is the most important but it is difficult to have enough objects due to the fact that it could be associated to exceptional cases [2]. Second, contrast pattern miners can be biased to the majority class and therefore they commonly extract several contrast patterns from this class but only a few patterns (or none) from the minority class. Therefore, it is important to develop some approaches for applying contrast pattern based classifiers in class imbalance problems.

Recently, Prati et al. [15] conduct an exhaustive experimental study on imbalanced classification. Although this study does not include the main contrast pattern based classifiers, the authors show that most of the studied classifiers, including C4.5 [16,17], present some loss of performance for all class distributions. This loss tends to be higher as the classes become more imbalanced. In this paper, we study classification on imbalanced databases by focusing on the application of resampling methods for contrast pattern based classifiers.

2.1. Contrast patterns

A *pattern* is an expression defined in a certain language that describes a collection of objects. For example, a pattern that describes a set of plants can be the following: $[PetalWidth \in [0.60, 1.60]] \wedge [PetalLength \leq 4.90] \wedge$ $[Stem = \text{“}Thick\text{”}]$. Then, a *contrast pattern* is a pattern appearing frequently in a class and infrequently in the remaining problem classes [18].

2.1.1. Contrast pattern miners

There are many paradigms for mining contrast patterns in class imbalance problems, each one using particular data structures and algorithms.

EPRC [19] is based on applying some improving stages to maximize the discriminating power of emerging patterns (EPs) in the minority class. The main idea is generating new undiscovered EPs in the minority class, pruning low-support EPs, and increasing the support of EPs in the minority class.

DEP [20] is based on dividing the mining process into a number of subprocesses and then combining the obtained EPs according to their strength. This approach involves a comparison between the final subset of minority class EPs and the majority class EPs. If an EP exists in both subsets, then it is eliminated from the final subset and the strongest EP in the pending subset is added to the final subset. The strength of an EP is computed through the quality measure proposed in [21].

Those algorithms based on decision trees deserve special attention because this paradigm does not include a global discretization step and it obtains a small collection of high quality patterns [1].

EPDT [22] aims at supporting decision trees in class imbalance problems. It creates new non-existing minority class instances. Then, the most important minority class instances are oversampled. This approach increases the performance of decision trees through balancing the classes by oversampling the minority class.

The Logical Complex Miner (LCMine) is a contrast pattern miner based on decision trees, which has reported good results in balanced distribution [1,23]. This miner extracts a representative collection of emerging patterns from a set of diverse decision trees. The tree induction procedure used in LCMine is similar to traditional methods for building decision trees, but in order to generate diverse decision trees, LCMine selects the best k splits in first levels and the best split in lower levels. This way, for $k = \{5, 4, 3, 2\}$ LCMine creates $(5 \cdot 4 \cdot 3 \cdot 2) = 120$ different trees. Each decision tree is pruned to allow obtaining non-pure leaf nodes, therefore extracted patterns range from pure contrast patterns to patterns with small difference of support between classes. Patterns are extracted from the paths from the root node to the leaves. For each extracted pattern, the class with highest support determines the class of the pattern. In LCMine, the authors include a filtering strategy to reduce pattern redundancy.

2.1.2. Contrast pattern based classifiers

Contrast pattern based classifiers have shown to make consistently more accurate predictions than popular classification models including decision trees, Naive Bayes, Nearest Neighbor, bagging, boosting, and even SVM [18,23].

Classification by Aggregating Emerging Patterns (CAEP) [21] was the first classifier using aggregation of support, which can use patterns with support in more than one class. CAEP computes, for