Neurocomputing 126 (2014) 15–28

Contents lists available at ScienceDirect

Neurocomputing

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

Addressing imbalanced classification with instance generation techniques: IPADE-ID

Victoria López [a,*], Isaac Triguero [a], Cristóbal J. Carmona [b], Salvador García [b], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, 18071 Granada, Spain
[b] Department of Computer Science, University of Jaén, 23071 Jaén, Spain

ARTICLE INFO

Article history:
Received 10 May 2012
Received in revised form
30 November 2012
Accepted 10 January 2013
Available online 26 July 2013

Keywords:
Differential evolution
Instance generation
Nearest neighbor
Decision tree
Imbalanced datasets

ABSTRACT

A wide number of real word applications presents a class distribution where examples belonging to one class heavily outnumber the examples in the other class. This is an arduous situation where standard classification techniques usually decrease their performance, creating a handicap to correctly identify the minority class, which is precisely the case under consideration in these applications.

In this work, we propose the usage of the Iterative Instance Adjustment for Imbalanced Domains (IPADE-ID) algorithm. It is an evolutionary framework, which uses an instance generation technique, designed to face the existing imbalance modifying the original training set. The method, iteratively learns the appropriate number of examples that represent the classes and their particular positioning. The learning process contains three key operations in its design: a customized initialization procedure, an evolutionary optimization of the positioning of the examples and a selection of the most representative examples for each class.

An experimental analysis is carried out with a wide range of highly imbalanced datasets over the proposal and recognized solutions to the problem. The results obtained, which have been contrasted through non-parametric statistical tests, show that our proposal outperforms previously proposed methods.

1. Introduction

Classification with imbalanced datasets is a challenging data mining problem that has attracted a lot of attention in the last years [1,2]. This problem is extremely important since it is predominant in many real-world data mining applications including, but not limited to, medical diagnosis, fraud detection, finances, network intrusion and so on. These applications feature samples from one class which are greatly outnumbered by the samples of the other class. Usually, the minority class is the most interesting class from the learning point of view and implies a higher cost of making errors [3,4].

Imbalanced datasets have become an important difficulty to most classifiers, which assume a nearly balanced class distribution [5]. Standard classifiers are developed to minimize a global measure of error, which is independent of the class distribution and causes a bias towards the majority class, paying less attention to the minority class. Consequently, classifying the minority class is more error prone than classifying the majority class, as a huge portion of errors are concentrated in the minority class [6]. Furthermore, the examples of the minority class can be treated as noise and they might be completely ignored by the classifier.

Numerous approaches have been suggested to tackle the problem of classification with imbalanced datasets [1,2,7]. These approaches are developed at both data and algorithm levels. Solutions at the algorithm level modify existing learning algorithms conducting its operations on the improvement of the learning on the minority class [8,9]. Solutions at the data level, also known as data sampling, try to modify the original class distribution in order to obtain a more or less balanced dataset that can be used to correctly identify each class with standard classifiers [10–12].

The use of instance reduction methods [13], which were originally designed for other preprocessing purposes (speed up, noise tolerance and reduction of storage requirements of learning methods [14]), can also be applied to imbalanced datasets [15,16] as a data level solution that is used to find a balance between the minority and the majority classes. It is important that instance reduction methods adapt their bias to this situation to obtain high performances.

An instance reduction process is devoted to find the best reduced set that represents the original training data with a lesser number of instances. This methodology can be divided into Instance Selection (IS) [13,17,18] and Instance Generation (IG) depending on how it creates the reduced set [19,20]. The former process attempts to choose an appropriate subset of the original

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: vlopez@decsai.ugr.es (V. López), triguero@decsai.ugr.es (I. Triguero), ccarmona@ujaen.es (C.J. Carmona), sglopez@ujaen.es (S. García), herrera@decsai.ugr.es (F. Herrera).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.01.050

training data, while the latter can also build new artificial instances to better adjust the decision boundaries of the classes. In this manner, the IG process fills some regions in the domain of the problem, which have no representative examples in the original dataset. IS methods have been applied to imbalanced datasets with promising results [15,16,21], however, to the best of our knowledge, IG techniques have not been used yet to deal with imbalanced classification problems.

Following the idea of IG techniques, we propose the usage of the Iterative Instance Adjustment for Imbalanced Domains (IPADE-ID) algorithm to deal with highly imbalanced datasets. IPADE-ID is a method inspired by the IG technique IPADE [22,23], that tries to obtain an adequate synthetic training set from the original training set following an incremental approach to determine the most appropriate number of instances per class. The proposal is based in three fundamental operations: a customized initialization procedure, an evolutionary adjustment of the prototypes and the selection of the most representative examples to define the classes. The initialization procedure should be befitting to the specific learning algorithm used with IPADE-ID.

In this work, we choose the Nearest Neighbor (NN) rule [24] and the C4.5 algorithm [25] as learning methods. In this way, we provide suitable initialization procedures for IPADE-ID that matches these learning approaches. At each step, an optimization procedure, based on an adaptive differential evolution algorithm [26–28], adjusts the positioning of the instances generated up to now, and a selection procedure adds new instances if needed. This selection procedure has been particularly designed to consider the existing imbalanced scenario focusing on the performance of the minority class. This informed and organized combination of techniques, leads us to a hybrid artificial intelligent system [29,30] that is able to cope with imbalanced datasets.

In order to analyze the performance of the proposal, we focus on highly imbalanced binary classification problems, having selected a benchmark of 44 problems from KEEL dataset repository[1] [31]. We will perform our experimental analysis focusing on the precision of the models using the Area Under the ROC curve (AUC) [32]. This study will be carried out using non-parametric statistical tests to check whether there are significant differences among the results [33,34].

The rest of the paper is organized as follows. In Section 2, some background about classification with imbalanced datasets and instance generation techniques is given. Next, Section 3 introduces the proposed approach. Sections 4 and 5 describe the experimental framework used and the analysis of results, respectively. Finally, the conclusions achieved in this work are shown in Section 6.

2. Background

This section purpose is to provide the background information needed to describe our proposal. It is divided in two parts: a description of instance generation techniques (Section 2.1) and an introduction to the problem of classification with imbalanced datasets (Section 2.2).

2.1. Instance generation techniques

This section presents the definition and notation for instance generation techniques.

A formal specification of the instance generation problem is the following: Let $\mathbf{x}_p$ be an example where $\mathbf{x}_p = (x_{p1}, x_{p2}, \ldots, x_{pD}, C_p)$, with $\mathbf{x}_p$ belonging to a class C_i given by C_p and a D-dimensional space in which x_{pj} is the value of the j-th feature of the p-th sample. Then, let us assume that there is a training set TR which consists of n instances $\mathbf{x}_p$ and a test set TS composed of t instances $\mathbf{x}_q$, with C_q unknown.

The original purpose of IG is to obtain an instance generated set (GS), which consists of r, $r < n$, instances $\mathbf{p}_u$ where $\mathbf{p}_u = (p_{u1}, p_{u2}, \ldots, p_{uD}, C_u)$, which are either selected or generated from the examples of TR. The instances of the generated set are determined to efficiently represent the distributions of the classes and to discriminate well when used to classify the training objects.

This methodology, also known as instance abstraction, has been widely studied in the specialized literature focusing on the NN rule [24] as target classifier. These techniques follow multiple mechanisms to generate an appropriate GS, such as interpolations between instances, movements of instances and artificial generation of new data. Using the taxonomy proposed in [20], they can be divided into several families depending on the main heuristic operation: positioning adjustment [35], class re-labeling [36], centroid-based [19] and space-splitting [37].

Among these families of methods, the algorithms that are based on the adjustment of the position of the instances were highlighted as outstanding methods in [20]. This methodology can be viewed as an optimization process of the positioning of the instances [38]. The precursor algorithm of this family is the learning vector quantization proposed by Kohonen [39]. One of the most recent and promising algorithms is the model presented in [22], called IPADE, which follows an incremental addition process of instances to determine which classes need more instances to be represented and their best locations.

More information about instance generation approaches (and instance reduction approaches in general) can be found at the SCI2S thematic public website on *Prototype Reduction in Nearest Neighbor Classification: Prototype Selection and Prototype Generation*.[2]

2.2. Imbalanced datasets in classification

In this section we delimit the context in which this work is content, briefly introducing the problem of imbalanced classification. Then, we will describe which approaches are used to deal with this problem, giving special importance to data level approaches that modify the class distribution. We finish this section describing the evaluation metrics that are used in this specific problem with respect to the most common ones in classification.

2.2.1. The problem of imbalanced datasets

In some classification problems, the number of instances that belong to each class is radically different [1,2]. The problem of classification with imbalanced datasets has acquired much relevance in the last time due to its presence in abundant real-world applications such as medical diagnosis [40], finances [41,42], anomaly detection [43] or bioinformatics [44] just naming some of them. Furthermore, the underrepresented class is usually the most interesting class from the learning point of view incorporating high costs when it is not correctly identified [3,4].

In this paper, we focus on two-class imbalanced datasets, where there is a positive (minority) class, with the lowest number of instances, and a negative (majority) class, with the highest number of instances. Although this class distribution is frequent in real data mining problems, standard classifiers are not usually able to cope with the correct identification of positive samples. Frequently, standard classifiers are biased towards the majority class as they are guided by global performance measures, selecting more general rules that cover as many samples as possible and

[1] http://www.keel.es/datasets.php.

[2] http://sci2s.ugr.es/pr/.