Knowledge-Based Systems 73 (2015) 1–17

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets

Saleh Alshomrani [a], Abdullah Bawakid [a], Seong-O Shim [a], Alberto Fernández [b,*], Francisco Herrera [a,c]

[a] *Faculty of Computing and Information Technology – North Jeddah, King Abdulaziz University (KAU), Jeddah, Saudi Arabia*
[b] *Department of Computer Science, University of Jaén, Jaén, Spain*
[c] *Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, Granada, Spain*

ARTICLE INFO

Article history:
Received 7 February 2014
Received in revised form 1 August 2014
Accepted 7 September 2014
Available online 16 September 2014

Keywords:
Imbalanced datasets
Overlapping
Feature weighting
Evolutionary fuzzy systems
Fuzzy rule based classification systems

ABSTRACT

In a general scenario of classification, one of the main drawbacks for the achievement of accurate models is the presence of high overlapping among the concepts to be learnt. This drawback becomes more severe when we are addressing problems with an imbalanced class distribution. In such cases, the minority class usually represents the most important target of the classification. The failure to correctly identify the minority class instances is often related to those boundary areas in which they are outnumbered by the majority class examples.

Throughout the learning stage of the most common rule learning methodologies, the process is often biased to obtain rules that cover the largest areas of the problem. The reason for this behavior is that these types of algorithms aim to maximize the confidence, measured as a ratio of positive and negative covered examples. Rules that identify small areas, in which minority class examples are poorly represented and overlap with majority class examples, will be discarded in favor of more general rules whose consequent will be unequivocally associated with the majority class.

Among all types of rule systems, linguistic Fuzzy Rule Based Systems have shown good behavior in the context of classification with imbalanced datasets. Accordingly, we propose a feature weighting approach which aims at analyzing the significance of the problem's variables by weighting the membership degree within the inference process. This is done by applying a different degree of significance to the variables that represent the dataset, enabling to smooth the problem boundaries. These parameters are learnt by means of an optimization process in the framework of evolutionary fuzzy systems. Experimental results using a large number of benchmark problems with different degrees of imbalance and overlapping, show the goodness of our proposal.

1. Introduction

The significance of classification with imbalanced data arises when researchers realize that the datasets they are analyzing hold more instances or examples from one class than that of the remainder, and that they therefore obtain classification models below a desired accuracy threshold for that class. This scenario, known as the problem of classification with imbalanced datasets [41,28], is commonly addressed in a binary context where there is a single minority (positive) class, and a majority (negative) class.

The bias of standard classification algorithms towards the majority class examples [52,27], is the most straightforward consequence derived from the uneven class distribution. Those algorithms which obtain a good behavior in the framework of standard classification do not necessarily achieve the best performance for imbalanced datasets [20]. The imbalanced problem usually appears in combination with several additional data intrinsic characteristics [41]. This imposes further restrictions on the learning stage in terms of it being able to develop a classifier with a high accuracy for the positive and negative classes of the problem.

One of the main drawbacks for the correct identification of the positive class of the problem is related to overlapping between the classes [36,24,13]. Rules with a low confidence and/or coverage, because they are associated with an overlapped boundary area, will be discarded. Therefore, positive class examples belonging to this area are more likely to be misclassified.

[*] Corresponding author. Tel.: +34 953 213016; fax: +34 953 212472.
E-mail addresses: sshomrani@kau.edu.sa (S. Alshomrani), abawakid@kau.edu.sa (A. Bawakid), seongo@gmail.com (Seong-O Shim), alberto.fernandez@ujaen.es (A. Fernández), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.knosys.2014.09.002
0950-7051/© 2014 Elsevier B.V. All rights reserved.

The representation of a classification problem by means of its variables or features, will determine the way in which the classifier will discriminate between the examples of both classes. It is well known that a large number of features can degrade the discovery of the borderline areas of the problem [39], either because some of these variables might be redundant or because they do not show a good synergy among them. For this reason, some works on the topic have proposed the use of feature selection for imbalanced datasets in order to overcome this problem [54], and to diminish the effect of overlapping [13]. However, the application of feature selection might be too aggressive and therefore some potential features could be discarded. In most cases, every variable of the problem should make at least a small contribution in the learning stage, and the combination of all of them may help to achieve a better separability of the classes.

Linguistic FRBCSs have the advantage of achieving a good performance in the context of classification with imbalanced datasets [19,47]. The use of linguistic fuzzy sets allows the smoothing of the borderline areas in the inference process, which is also a desirable behavior in the scenario of overlapping.

In this paper, we propose the use of a feature weighting approach in the context of Linguistic Fuzzy Rule Based Classification Systems (FRBCSs) [34]. Basically, we propose the consideration of the feature weights as a part of the reasoning model. We modify the computation of the membership functions associated with the fuzzy labels in the antecedents of the rules, in order to take into account the significance of the problem's variables throughout the inference process.

The computation of the optimal parameters for setting the weight of each variable, will be carried out by means of Evolutionary Algorithms [15]. The hybridization of this approach with the previously introduced FRBCSs will lead to the development of an Evolutionary Fuzzy System (EFS) [11,17]. One of the main reasons for the success of this type of techniques is their ability to exploit the information accumulated about and initially unknown search space in order to bias subsequent searches into useful subspaces, i.e. their robustness [11]. For the fuzzy learning classifier, we have considered the use of a robust FRBCS, i.e. the Fuzzy Association Rule-based Classification for High-Dimensional problems (FARC-HD) [1]. The proposed algorithm using feature weighting will receive the acronym FARC-HD-FW, based on the previous name (FARC-HD) and the use of feature weighting.

In order to evaluate the goodness of the feature weighting proposal, we will contrast our results with the standard FARC-HD algorithm and FARC-HD with feature selection. Additionally, we will complement our comparison with the C4.5 decision tree [49] as a standard baseline algorithm, and several EFS approaches developed for both classical and imbalanced classification such as 2-tuples lateral tuning [18], the Hierarchical Genetic Programming-based learning of COmpact and ACcurate fuzzy rule-based classification systems for High-dimensional problems (GP-COACH-H) [40], and the Interval-Valued Fuzzy Decision Tree (IIVFDT) [50]. The validity of our approach in the scenario of imbalanced and overlapping datasets will be tested using a wide benchmark of 66 different problems commonly used in the topic of classification with imbalanced datasets [41].

This paper is organized as follows. Section 2 briefly introduces the problem of imbalanced data, its relationship with class overlapping and how to address and evaluate this problem. Then, Section 3 contains the central part of the manuscript, in which the proposed methodology for dealing with overlapping in imbalanced data with FRBCSs is described. Next, the details about the experimental framework selected for the validation of our approach are introduced in Section 4. The analysis and discussion of the experimental results is carried out in Section 5. Finally, Section 6 summarizes and concludes the work.

2. Imbalanced datasets in classification

In this section, we present some preliminary concepts regarding classification with imbalanced datasets. This section is divided into the following four parts:

- We will first introduce the problem of imbalanced datasets, describing its features and why is so difficult to learn in this classification scenario (Section 2.1).
- Then, we will focus on the presence of overlapping between the classes, which further complicates the correct identification of the positive instances (Section 2.2).
- In the next section, we will present how to address this problem, focusing on the preprocessing of instances for rebalancing the distribution between the positive and negative classes (Section 2.3).
- Finally, we will discuss how to evaluate the performance of the results in this situation (Section 2.4).

2.1. Basic concepts on classification with imbalanced datasets

The main property of this type of classification problem is that the examples of one class outnumber the examples of the other one [52]. The minority classes are usually the most important concepts to be learnt, since they might be associated with exceptional and significant cases [55] or because the data acquisition of these examples is costly [57]. Since most of the standard learning algorithms consider a balanced training set, this situation may cause suboptimal classification models to be obtained, i.e. a good coverage of the majority examples but a more frequent misclassification of the minority ones [27]. Traditionally, the Imbalance Ratio (IR), i.e. the ratio between the majority and minority class examples [45], is the main clue to identify a set of problems which need to be addressed in a special way.

We must stress the following reasons for this behavior [41]: the use of global performance measures for guiding the search process, such as standard accuracy rate, which may benefit the covering of the majority class examples, and the low coverage of the classification rules for the positive class, which are discarded in favor of more general rules, especially in the case of overlapping [36,13]; small clusters of minority class examples that can be treated as noise and wrongly ignored by the classifier [46,56]; few real noisy examples which may degrade the identification of the minority class, as it has fewer examples to begin with [51]; and dataset shift, i.e. different data distribution between training and test partitions [44]. For an in depth coverage of those data intrinsic characteristics which hinder the classification of imbalanced datasets, the reader may refer to a recent survey carried out in [41].

Finally, regarding the way to overcome the class imbalance problem, we may find a large number of proposed approaches, which can be categorized in three groups [42]:

1. Data level solutions: the objective consists of rebalancing the class distribution by sampling the data space to diminish the effect caused by class imbalance, acting as an external approach [21,25,38].
2. Algorithmic level solutions: these solutions try to adapt several classification algorithms to reinforce the learning towards the positive class. Therefore, they can be defined as internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration [4,58,61].
3. Cost-sensitive solutions: these types of solutions incorporate approaches at the data level, at the algorithmic level, or at both levels jointly. They consider higher costs for the mis-