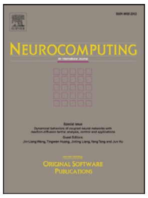
Neurocomputing 285 (2018) 176–187

Contents lists available at ScienceDirect

Neurocomputing

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



DRCW-ASEG: One-versus-One distance-based relative competence weighting with adaptive synthetic example generation for multi-class imbalanced datasets

Zhong-Liang Zhang[a,b,c], Xing-Gang Luo[a,b,*], Sergio González[c], Salvador García[c], Francisco Herrera[c,d]

[a] School of Management, Hangzhou Dianzi University, Hangzhou 310018, China
[b] School of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[c] Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain
[d] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 5 November 2017
Revised 4 January 2018
Accepted 16 January 2018
Available online 1 February 2018

Communicated by Dr. Ma Lifeng Ma

Keywords:
Multi-class problems
Imbalanced datasets
Ensemble learning
Binary decomposition
Synthetic samples generation

ABSTRACT

Multi-class imbalance learning problems suffering from the different distribution of classes occur in many real-world applications. One-versus-One (OVO) decomposition strategy is a common and useful technique used to address multi-class classification problems, which consists in dividing the original multi-class problem into all binary class sub-problems. The effort to reduce the effect of non-competent classifiers has proven to be a useful way of improving the performance in the OVO scheme. However, these approaches might not be effective for imbalance scenarios, since they are based on standard biased learning procedures. On this account, we propose a novel approach named Distance-based Relative Competence Weighting with Adaptive Synthetic Example Generation (DRCW-ASEG), which properly addresses the synergy between imbalance learning and dynamic classifier weighting in OVO scheme. This new proposed algorithm aims to dynamically produce synthetic examples of minority classes in the stage of dynamic weighting process. We develop a thorough experimental study in order to verify the benefits of the proposed algorithm considering different base binary classifiers.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

The class imbalance learning problem, also known as imbalanced dataset, refers to the classification task, where one or more classes are under represented in the dataset [1,2]. Learning from datasets with this under-representation results in classifiers biased towards classes with more representatives and loss of prediction of the minority classes [3].

The class imbalance problem has drawn the attention of practitioners over the years, due to its relevance in many real-world classification tasks, such as human activity recognition [4], medical diagnosis [5], data stream analysis [6], and facial expression recognition [7]. In these problems, the goal or the objective class tends to be an under-represented class. However, this attention has been focused mostly on binary class imbalance problems.

Multi-class imbalance problems are considerably more difficult to address than two-class scenarios, since the decision boundary involves distinguishing among more classes. And the vast production for binary imbalance problems unfortunately cannot be directly applied to multi-class imbalance problems [8].

However, multi-class classification problems can be addressed by the usage of binarization decomposition strategies, which aims to divide the complex problem into several simpler two-class problems [9]. Several alternatives following this divide-and-conquer strategy can be found in the specialized literature [10]. Among them, One-versus-One (OVO) [11–13] and One-versus-All (OVA) [14] are the most popular techniques. In this study, we only focus on dealing with multi-class imbalanced datasets using the OVO strategy, as it has proven to perform better than the OVA scheme in this scenario [8,9].

OVO decomposition strategy constructs a system with multiple binary classifiers to discriminate between each pair of classes. For a given test pattern, all the binary classifiers will be triggered. Although, as they are only trained with two-class instances, some of them do not have sufficient knowledge to predict correctly.

[*] Corresponding author at: School of Management, Hangzhou Dianzi University, 310018 Hangzhou, China.

E-mail addresses: zzl19860210@126.com, zlzhang@hdu.edu.cn (Z.-L. Zhang), xgluo@mail.neu.edu.cn (X.-G. Luo), sergiogvz@decsai.ugr.es (S. González), salvagl@decsai.ugr.es (S. García), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.neucom.2018.01.039
0925-2312/© 2018 Elsevier B.V. All rights reserved.

This problem is called as the "non-competent classifiers problem" [15–17].

Recently, some techniques have been proposed to manage this problem for the standard multi-class classification task. Among them, Distance-based Relative Competence Weighting for OVO strategy (DRCW-OVO) [17] has better and more robust performance. In DRCW-OVO, the confidence degrees of the classifiers are weighted depending on their competence associated with the average distances between the classes and the query instance. These average distances are computed with the nearest neighbors of the unlabeled example for each class. The classes closer to the instance to be classified get higher weights in the aggregation method.

However, it may be useless to directly employ DRCW-OVO to handle the class imbalance problems. In the multi-class imbalanced dataset, the minority classes might be much sparser than those of the majority classes within the neighborhood of a given instance to be classified. Therefore, the average distances of the former would be larger, deviating the decision from the minority classes. In this sense, the essential of the drawback for DRCW-OVO to deal with the non-competent classifier problem is the skewed distribution of the local region surrounding the tested example. Therefore, it seems to be reasonable for us to develop an approach to interpolate synthetic examples with the aim of balancing the local region when managing the non-competent classifiers existing in the OVO system for handling multi-class imbalanced datasets.

The goal of this paper is to introduce a new decomposition framework named DRCW-ASEG that is applicable to multi-class imbalanced datasets and empowers OVO decomposition with the DRCW strategy. At the same time, DRCW-ASEG deals with multi-class imbalance problems and with the "non-competent classifiers problem". This novel approach alleviates the bias towards the majority classes of the post-processing DRCW method, due to the use of the Adaptive Synthetic Example Generation procedure (ASEG) in the stage of dynamic weighting process. We should emphasize that whereas the approaches of adaptively generating minority examples according to their distributions, like ADASYN [18], have been already considered for imbalanced datasets, they have not been considered to manage the non-competent classifiers in OVO scheme. Notice that in this paper, the synthetic examples are generated to balance the local region, and they are not used to establish binary classifiers.

Concretely in DRCW-ASEG, the original multi-class imbalance problem is divided into several two-class sub-problems according to the OVO strategy. And, these are addressed by the binary class imbalanced methods. The non-competent classifier problem is addressed with DRCW but applied to a balanced local region of the sample needed to be classified generated by ASEG. This balanced region is obtained by analyzing the neighborhood of the query example. Then, new synthetic examples are locally generated by the interpolation of those minority classes in the targeted neighborhood.

In order to show the validity of our method, we carry out a thorough experimental study on twenty multi-class imbalanced datasets selected from the KEEL dataset repository [19,20]. Regarding the imbalance learning method in our scenario, six well-known class imbalance learning methods are selected to handle the two-class imbalanced datasets derived from the OVO decomposition, including three basic resampling techniques (Random Under Sampling (RUS), Random Over Sampling (ROS) and SMOTE [21,22]) and three ensemble learning methods (SMOTEBagging [23], RUSBagging [24] and SMOTEBoost [25]).

Marco Average Arithmetic (MAvA) metric [26] is employed as the performance measure and the proper statistical tests suggested in [27] are used to study the significance of the results.

The main contributions of this paper with respect to previous works can be summarized as follows:

- In this paper, we focus on dealing with considerably complicated multi-class imbalance learning tasks using the OVO decomposition technique. We provide the detailed analysis showing that the classic approaches for the management of non-competent classifiers are inefficient in the scenario of multi-class imbalanced datasets.
- In order to alleviate the negative effect of non-competent classifiers using OVO method to handle multi-class imbalanced problems, we develop a new method named DRCW-ASEG which generates synthetic examples by dynamically learning the local information from the data distribution.
- We carry out extensive experiments on real-world datasets to evaluate the performance of the proposed methods. We also study the internal mechanism of the proposed method by analyzing the classification performance on per class.

The rest of this paper is organized as follows. The previous research related to this study, namely the imbalanced classification problem, solutions for multi-class imbalanced datasets, the OVO decomposition scheme and the non-competent classifier problem are introduced in Section 2. Then, in Section 3 we present our methodology for integrating an adaptive synthetic example generation into the OVO scheme, the DRCW-ASEG method. In Section 4, the experimental framework is given, including the datasets, base learner, binary imbalance learning algorithms and measure metric. And then, the complete empirical study and its analysis is carried out in Section 5. Finally, the conclusions and discussions are summarized in Section 6.

2. Background

In this section, we present two scenarios for class imbalanced datasets according to the number of classes. Section 2.1 is devoted to presenting the traditional paradigm of the imbalanced classification problem as a binary scenario. Then, in Section 2.2, we introduce the multi-class imbalanced datasets, its difficulties and possible solutions. Next, we descripe OVO decomposition strategy for multi-class classification problems in Section 2.3. Finally, the non-competent classifier problem in OVO scheme is discussed in Section 2.4.

2.1. The imbalanced classification problem

The classification task on a dataset is considered a class imbalance problem, when several classes (minority classes) are underrepresented with respect to the other classes (majority classes), i.e., the number of minority class examples is much less than the number of representatives of the majority classes.

Traditionally practitioners have addressed the problem as a binary scenario dilemma, where the relationship between classes and the goal is well-defined: to reduce the bias towards the majority class and meanwhile, balance the performance on both classes. With this objective, a multitude of approaches have been designed according to four different lines of research: data level methods, algorithmic approaches, cost-sensitive learning solutions, and ensemble learning approaches.

Data level solutions address the origin of the class imbalance problem, i.e. the skewed distribution in the dataset. This is done by sampling the data space to re-balance and reduce the impact of class imbalance [2]. One of the advantages of such a solution is its independence from the classifier used. The two basic ideas of resampling approaches are Random Under/Over- sampling, that randomly remove or duplicate examples from the majority/minority class. Dozens of sophisticated resampling techniques have been proposed over the years. Among them, the Synthetic Minority Oversampling Technique (SMOTE) [21] proposed by Chawla

Download English Version:

https://daneshyari.com/en/article/6864459

Download Persian Version:

https://daneshyari.com/article/6864459

Daneshyari.com