Information Sciences 445–446 (2018) 22–37

Contents lists available at ScienceDirect

Information Sciences

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

Dynamic ensemble selection for multi-class imbalanced datasets

Salvador García[a,*], Zhong-Liang Zhang[b], Abdulrahman Altalhi[c], Saleh Alshomrani[d], Francisco Herrera[a,c]

[a] *Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain*
[b] *School of Management, Hangzhou Dianzi University, Hangzhou 310018, China*
[c] *Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia*
[d] *Faculty of Computing and Information Technology, University of Jeddah, Jeddah 21589, Saudi Arabia*

ARTICLE INFO

Article history:
Received 10 October 2017
Revised 1 March 2018
Accepted 3 March 2018

Keywords:
Multi-class classification
Imbalanced datasets
Multiple classifier system
Ensemble learning
Binary decomposition
Resampling techniques

ABSTRACT

Many real-world classification tasks suffer from the class imbalanced problem, in which some classes are highly underrepresented as compared to other classes. In this paper, we focus on multi-class imbalance problems which are considerably more difficult to address than two-class imbalanced problems. On this account, we develop a novel and effective procedure, called dynamic ensemble selection for multi-class imbalanced datasets (DES-MI), in which the competence of the candidate classifiers are assessed with weighted instances in the neighborhood. The proposed DES-MI consists of two key components: the generation of balanced training datasets and the selection of appropriate classifiers. To do so, we develop a preprocessing procedure to balance the training dataset which relies on random balance. To select the most appropriate classifiers in the scenario of multi-class imbalance problems, we propose a weighting mechanism to highlight the competence of classifiers that are more powerful in classifying examples in the region of underrepresented competence. We develop a thorough experimental study in order to verify the benefits of DES-MI in handling multi-class imbalanced datasets. The obtained results, supported by the proper statistical analysis, indicate that DES-MI is able to improve the classification performance for multi-class imbalanced datasets.

1. Introduction

In the research community of machine learning, the class imbalance learning problem is considered to be one of the most important challenges [22]. Imbalanced data refers to such a dataset in which several classes are underrepresented (minority classes) when compared with others (majority classes) [37]. This skewed distribution makes many classic classification algorithms, all of which designed with the premise of a balanced training dataset, such as Bayes classifier, decision tree, *k*-nearest neighbor (*k*NN), artificial neural network (ANN), support vector machine (SVM), less effective [47]. On the other hand, many real-world classification tasks involve the class imbalance learning problems, for example, social media analysis [25], action recognition [24], water quality prediction [26], and text categorization [43].

[*] Corresponding author.
E-mail addresses: salvagl@decsai.ugr.es (S. García), zlzhang@hdu.edu.cn (Z.-L. Zhang), ahaltalhi@kau.edu.sa (A. Altalhi), sshomrani@kau.edu.sa (S. Alshomrani), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.ins.2018.03.002
0020-0255/© 2018 Elsevier Inc. All rights reserved.

In fact, many efforts have been devoted to imbalanced learning problems in the research community, such as resampling methods, cost sensitive approaches, ensemble learning algorithms, kernel-based methods and active learning methods [15]. However, most of them so far have been focused on two-class imbalanced datasets. Obviously, multi-class imbalance learning problems are much more difficult to address than the binary scenario, since the decision boundary involves distinguishing between more classes. Unfortunately, directly applying the methods proposed for dealing with two-class imbalanced problems to the multi-class problems may be invalid [9].

Three types of difficulties related to a multi-class imbalanced dataset exist: one majority and many minority classes, one minority and many majority classes, and many minority and many majority classes [31]. In addition, the skewed distribution of instances among classes is not the only source of difficulties for classification algorithms to deal with multi-class imbalanced datasets. Difficulties embedded in the structure of data are also always present, such as class overlapping, small disjuncts (minority class can consist of several sub-concepts) and small sample size (lack of representative minority instances).

As explained above, in multi-class imbalance learning tasks, different characteristics are represented in different regions, which are associated with different classification difficulties. Therefore, it is reasonable for us to adopt a dynamic scheme to select different classifiers for different query examples. Dynamic classifier selection (DCS) and dynamic ensemble selection (DES) are the most famous techniques based on dynamic selection [49]. The former tends to select the most appropriate single classifier for the query instance, while the latter aims to dynamically acquire a classifier system consisting of several competent classifiers. The latest research results [4] indicate that DES usually outperforms the DCS method, since the former approach distributes the risk of this over-generalization by selecting a group of classifiers instead of one individual classifier for a query example.

Motivated by the above analysis, a simple and yet effective method named DES-MI, which stands for dynamic ensemble selection for multi-class imbalanced datasets, is proposed in this paper to deal with multi-class imbalanced datasets, paying attention to the following tasks:

- The diversity of classifiers in the candidate pool is achieved based on the random balance framework, which hybridly uses the techniques of random under-sampling (RUS), random over-sampling (ROS) and synthetic minority oversampling technique (SMOTE) [5].
- Then, the competence of candidate classifiers is evaluated by using the weighted instances in the neighborhood surrounding the query example x_t. That is, we consider higher competence for the classifier that is more powerful in classifying minority classes within the local region.
- Finally, every selected classifier submits a vote on the query example x_t. The votes received by each class are counted and the class with the largest number of votes is considered as the final output class.

In order to show the validity of our approach, we carry out a thorough experimental study on 20 multi-class imbalanced datasets selected from the KEEL dataset repository [36]. We measure the performance of the methods based on both marco average arithmetic (MAvA) [32] and mean F-measure (MFM) [12] metric. The effectiveness of dynamic selection in the scenario of multi-class imbalanced datasets is analyzed and DES-MI is compared with the state-of-the-art methods which are also developed for solving the multi-class imbalanced problem. The significance of the results is studied by the use of the proper statistical tests as suggested in [7].

The main contributions of this paper can be summarized as follows:

- In this paper, we focus on multi-class imbalance learning tasks, which are considerably more complicated than the binary scenario. A comprehensive analysis of difficulties in multi-class imbalanced datasets is presented, which provides a deep insight into the nature of multi-class imbalanced datasets.
- We propose a novel method to deal with multi-class imbalance learning problems by using a dynamic scheme, in which a group of appropriate classifiers are selected for each query example.
- Since classic DES approaches are developed for balanced datasets, we in this paper present a new DES model for imbalanced multi-class datasets, which generates the classifier pool by using a novel data preprocessing method. Then, it associates the voting weights of instances in the neighborhood of the query example which are adaptively adjusted according to the data distribution.

The remainder of this paper is organized as follows. In Section 2, we first describe the multi-class imbalance learning problem, including the notations used throughout this paper and solutions for multi-class imbalance learning tasks. Next, in Section 3 we review related works on DES. Then, Section 4 presents the proposed method, DES-MI. Section 5 describes the experimental framework, whereas Section 6 shows the results and discussion. Finally, the conclusions are given in Section 7.

2. Multi-class imbalance learning problem: notations and solutions throughout the literature

In order to clearly present our findings, we first establish the following notations used in this paper:

- $D = \{x_i, y_i\}_{i=1}^{n}$ is a dataset containing n examples, where $y_i \in \{\omega_1, \ldots, \omega_m\}$ is the corresponding class label for the sample x_i and m is the number of calsses.
- x_t is a test example with an unknown class label.