Knowledge-Based Systems 106 (2016) 251–263

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data

Zhongliang Zhang[a], Bartosz Krawczyk[b,*], Salvador Garcìa[c], Alejandro Rosales-Pérez[d], Francisco Herrera[c,e]

[a] *School of Information Science and Engineering, Key Laboratory of Integrated Automation of Process Industry, Northeastern University, Shenyang 110819, China*
[b] *Department of Systems and Computer Networks, Wrocław University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland*
[c] *Department of Computer Science and Artificial Intelligence, University of Granada, P.O. Box 18071, Granada, Spain*
[d] *Instituto Nacional de Astrofísica, Óptica y Electrónica, Computer Science Department, Luis E. Erro No. 1, Santa María Tonantzintla, Puebla, 72840, Mexico*
[e] *Faculty of Computing and Information Technology, King Abdulaziz University, 21589, Jeddah, Saudi Arabia*

ARTICLE INFO

Article history:
Received 3 November 2015
Revised 1 April 2016
Accepted 24 May 2016
Available online 25 May 2016

Keywords:
Multi-class classification
Imbalanced data
Ensemble learning
Binary decomposition
Classifier combination

ABSTRACT

Multi-class imbalance classification problems occur in many real-world applications, which suffer from the quite different distribution of classes. Decomposition strategies are well-known techniques to address the classification problems involving multiple classes. Among them binary approaches using one-vs-one and one-vs-all has gained a significant attention from the research community. They allow to divide multi-class problems into several easier-to-solve two-class sub-problems. In this study we develop an exhaustive empirical analysis to explore the possibility of empowering the one-vs-one scheme for multi-class imbalance classification problems with applying binary ensemble learning approaches. We examine several state-of-the-art ensemble learning methods proposed for addressing the imbalance problems to solve the pairwise tasks derived from the multi-class data set. Then the aggregation strategy is employed to combine the binary ensemble outputs to reconstruct the original multi-class task. We present a detailed experimental study of the proposed approach, supported by the statistical analysis. The results indicate the high effectiveness of ensemble learning with one-vs-one scheme in dealing with the multi-class imbalance classification problems.

1. Introduction

In machine learning and data mining, while one or more classes are underrepresented in the data set, it is called as class imbalance classification. Many real-world classification tasks suffer from the class imbalance problem, which is considered as one of the important challenges for the data mining community [18]. The main difficulty of these problems is that the skewed distribution makes conventional classification algorithms less effective, since standard learning algorithms consider a balanced training data set, which result in making it harder to predict minority class examples [50].

In recent years, many efforts have been focused on the binary class imbalance problems [31,41], which only contain two classes. However, multi-class imbalance classification, is widely applied in many areas, such as text categorization [47], human activity recognition [1] and medical diagnosis [35]. Unfortunately, it may be invalid to directly apply the solutions proposed for the two-class problems to the multi-class imbalance problems, and some algorithms cannot be used to solve the multi-class imbalance problems directly [18].

Fortunately, in the research community, decomposition strategies turn up to deal with multi-class classification problem. In this solution framework, the multi-class classification problems are transformed into binary class sub-problems, which are much easier to discriminate [53,61]. Such well-known approaches are the one versus one (OVO) [23,33] and one versus all (OVA) [7]. As OVA introduces an artificial class imbalance (e.g., for 10 class problem with roughly equally represented classes, the binary sub-problem will have an imbalance ratio 1:9), it is not advisable to use it for handling problems with initially skewed distributions [46].

In this paper, we focus on multi-class imbalance classification problems and develop a complete empirical study to explore the effectiveness of ensemble learning methods [62] in the multi-class

[*] Corresponding author.
E-mail addresses: zzl19860210@126.com (Z. Zhang), bartosz.krawczyk@pwr.edu.pl (B. Krawczyk), salvagl@decsai.ugr.es (S. Garcìa), arosales@inaoep.mx (A. Rosales-Pérez), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.knosys.2016.05.048
0950-7051/© 2016 Elsevier B.V. All rights reserved.

imbalanced datasets with OVO scheme, where binary-class classifiers are trained from the subset containing each pair of classes by ensemble learning approaches based on data preprocessing [22]. Our initial works in this domain showed that empowering binary decomposition with pairwise ensemble learning can significantly improve mining imbalanced multi-class problems [34].

Regarding ensemble learning methods, six state-of-the-art approaches are selected to carry out the experiment: UnderBagging [3], SMOTEBagging [15,56], RUSBoost [49], SMOTEBoost [10], SMOTE+AdaBoost [40], EasyEnsemble [40]. Additionally, to show the efficiency of ensemble learning with OVO scheme for addressing the multi-class imbalance problems, the original data preprocessing strategies, including random under-sampling (RUS) [4], random over-sampling (ROS) [4] and synthetic minority oversampling technique (SMOTE) [8], are also implemented in the OVO scheme for our comparative analysis.

Finally, we carry out a thorough experimental study that supports the effectiveness of our methodology. Concretely, 20 multi-class imbalanced data sets are selected from the UCI repository in our experiment. The average accuracy rate [20] is used as the performance measures in this study. In order to analyze the results obtained from the different solutions, statistical analysis suggested in [28] is given to support the significance of the results.

The main contributions of this paper with respect to previous studies are as following:

- We propose to enhance the OVO scheme for multi-class imbalanced data by using ensemble techniques for each sub-problem.
- We show, how to extend the area of applicability of binary imbalanced ensemble classifiers to handling far more challenging multi-class imbalanced scenarios.
- We develop a complete experimental study of comparison of the state-of-the-art ensemble learning techniques with conventional resampling methods with OVO strategy and state-of-the-art solutions for multi-class imbalance problems.
- In order to obtain the impacts of the base classifier used in our scenario, we choose three different algorithms, including Classification and Regression tree (CART) [6], Back Propagation Neural Network (BPNN) [17] and Support Vector Machine (SVM) [54].

The rest of this paper is organized as follows. The background of this study is introduced in Section 2, including multi-class imbalance classification problems and decomposition strategies. Next, in Section 3 we present the framework of our methodology of ensemble learning with OVO scheme for dealing with multi-class imbalance classification problems. In Section 4, the experimental framework is given, including the data sets, the base classifiers and the relative parameters setting, the performance measures and the statistical tests. The complete empirical study is presented in Section 5. Lessons learned from the paper are given in Section 6, while conclusions and potential directions for future works are to be found in the final Section.

2. Background

In this section, we first introduce the problem of multi-class imbalance classification. Then, we present the solutions for addressing the imbalance problems. Finally, we describe the decomposition strategy for dealing with multi-class classification problems.

2.1. Multi-class imbalanced data analysis

Multi-class imbalanced data sets, where there are much more instances of some classes (referred to as the majority classes) than others (referred to as the minority classes), is one of the most challenging problems with data quality that always reduces classification performance in machine learning and data mining [57]. The minority classes are usually the most important concepts to recognize, since they represent the rare cases [59]. Additionally, it is expensive or hard to select these examples [58].

However, standard classification algorithms are designed with the premise of a balanced training set [42]. With such a precondition, it is much more difficult for the classical classification algorithms to deal with class imbalance problems, especially for identifying the minority class instance [9]. Additionally, most of the methods however are specific to address the binary class imbalance problems. Obviously, multi-class imbalance problems are far more complex, since these issues are involved with large number of classes and the relationships among the classes are complicated. Furthermore, it is hard to distinguish between minority classes and noise examples and the minority classes can be ignored by the classifier as the noise examples.

2.2. Solutions for imbalanced classification problems

To overcome the dilemma of skewed class distribution, a large amount of techniques have been developed to deal with such problem. These proposals can be roughly categorized into four groups:

- Data level: the origin of the problem is the class distribution in the data sets, therefore, it is natural to consider of rebalancing by sampling the data space to reduce the impact of class imbalance, known as an external approach. One of the advantages of such solution is independent from the classifier used, so they are also considered as pre-sampling method [27,51].
- Algorithm level: these solutions try to adopt appropriate decision threshold to reinforce the learning towards the minority class instances. The proposed algorithms that take the class imbalance into consideration belong to such techniques. They are defined as internal approaches in some papers [11,52], since the effect depends on the problems and the classifier [13]. One of the most well-known solutions is the direct modification of the learning procedure for a selected algorithm [45].
- Cost-sensitive level: these approaches consider higher costs for misclassifying the minority classes with respect to the majority classes, that is, misclassification of minority class is much more expensive [44]. The learning process turns to minimize the cost errors instead of maximization of accuracy rate [63].
- Ensemble level: these solutions combine the efficient ensemble learning solutions [62] with one of the three previously mentioned strategies in order to create a balanced training sets for base classifier and at the same time introduce diversity into the pool of base learners. Special attention should be paid to recent combination of intelligent and directed data-level approaches with Bagging solution [5] or randomized oversampling [15], hybrid combination of algorithm-level methods [55] and cost-sensitive pruning for decision tree ensembles [36].

Due to the advantage of the data level solutions (as pointed out by a recent tutorial on data preprocessing [29]) we focus on such methods in this study.

RUS [51] is the basic under-sampling, which randomly removes the majority class instances to balance the class distribution. This approach is efficient for dealing with class imbalance problems, since most of the majority class instances are redundant. Additionally, RUS makes the training process become much faster, since the training set contains less instances than original data set. However, some potential useful information contained in the majority class instance may be neglected, since RUS randomly generates the subset without considering the relationship among the instances.