Knowledge-Based Systems xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data

Q1 Li Yijing[a,b], Guo Haixiang[a,b,c,d,*], Liu Xiao[a,b], Li Yanan[a,b], Li Jinling[a,d]

[a] College of Economics and Management, China University of Geosciences, Wuhan 430074, China
[b] Research Center for Digital Business Management, China University of Geosciences, Wuhan 430074, China
[c] Mineral Resource Strategy and Policy Research Center of China University of Geosciences, Wuhan 430074, China
[d] The Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 14 May 2015
Revised 13 November 2015
Accepted 16 November 2015
Available online xxx

Keywords:
Imbalanced data
Multiple classifier system
Adaptive learning
Oil reservoir

ABSTRACT

Learning from imbalanced data, where the number of observations in one class is significantly rarer than in other classes, has gained considerable attention in the data mining community. Most existing literature focuses on binary imbalanced case while multi-class imbalanced learning is barely mentioned. What's more, most proposed algorithms treated all imbalanced data consistently and aimed to handle all imbalanced data with a versatile algorithm. In fact, the imbalanced data varies in their imbalanced ratio, dimension and the number of classes, the performances of classifiers for learning from different types of datasets are different. In this paper we propose an adaptive multiple classifier system named of AMCS to cope with multi-class imbalanced learning, which makes a distinction among different kinds of imbalanced data. The AMCS includes three components, which are, feature selection, resampling and ensemble learning. Each component of AMCS is selected discriminatively for different types of imbalanced data. We consider two feature selection methods, three resampling mechanisms, five base classifiers and five ensemble rules to construct a selection pool, the adapting criterion of choosing each component from the selection pool to frame AMCS is analyzed through empirical study. In order to verify the effectiveness of AMCS, we compare AMCS with several state-of-the-art algorithms, the results show that AMCS can outperform or be comparable with the others. At last, AMCS is applied in oil-bearing reservoir recognition. The results indicate that AMCS makes no mistake in recognizing characters of layers for oilsk81-oilsk85 well logging data which is collected in Jianghan oilfield of China.

1. Introduction

Classification is one of the crucial issues in the field of machine learning. Classical classifiers such as Decision Tree, Naïve Bayes, Artificial Neural Network (ANN), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) operate under the assumption that data sample contains a faithful representation of the population of interest, which means a balanced sample distribution is required [1]. When facing skewed class distribution, the traditional classifiers often come up to a disappointed performance [2–4]. Imbalanced data refers to such a dataset in which one or some of the classes contain much more samples in comparison to the others. The most prevalent class is called the majority class, while the rarest class is called minority class. Imbalanced situation often occurs in real word applications like fraud detection, disease diagnoses, financial risk analysis, etc. [5,6]. When addressing imbalanced data problems, people tend to care more about the minority class, for the reason that the cost of misclassifying minority samples are much higher than the others [2,6,7]. Taking cancer diagnoses for example, the number of cancer patients is much less than healthy people, if cancer patients are diagnosed as healthy people, they will exceed the best therapy time, which may cause a serious medical incidence [6]. So does oil-bearing recognition that is studied in this paper. Oil-bearing recognition refers to recognize the characters of each layer in the well [8,9], the class distribution of logging data is skewed and cost of misclassifying oil layer is much higher than other misclassification situations. Therefore, oil-bearing recognition is a typical imbalanced data classification problem.

Imbalanced learning is a well-studied problem, dozens of sampling methods [10,11], cost sensitive algorithms [17,18], one-class classifiers [53,54,57] have been proposed in literature. More recently, ensemble learning becomes a popular solution of addressing

* Corresponding author at: College of Economics and Management, China University of Geosciences, Wuhan 430074, China. Tel.: +86 15927389298; fax: +86 027 67883201.
Q2 *E-mail addresses:* faterdumk0732@sina.com (G. Haixiang), liyijing024@hotmail.com (L. Jinling).

http://dx.doi.org/10.1016/j.knosys.2015.11.013
0950-7051/

imbalanced data. A common way for constructing ensemble learning model for imbalanced data is based on sampling methods, that is, employing sampling methods as a pre-process to generate several balanced datasets and training different base classifiers independently. The main idea of constructing ensemble learning model is to learn from data by multiple classifiers, thus a designed ensemble learning model can also be viewed as Multiple Classifier System(MCS) [47].Since ensemble learning has been proved to be the most efficient way to tackle imbalanced learning problems [1,15,16,47,12], we aim to focus on constructing ensemble model in this paper.

Though various MCSs have been proposed, most of them model different types of data consistently and train a universal ensemble classifier to address all imbalanced data. In fact, using a specific ensemble classifier to tackle all kinds of imbalanced data is inefficient. The learning quality of a model can be affected by the choices of sampling methods, base classifiers, and final ensemble rules. For example, when the samples of minority class are extremely rare (saying we just have 1 or 2 minority samples), under-sampling methods may not be valid since we need to abandon tremendous number of majority samples to construct a balanced training set. The same concern should be highlighted when deciding which base classifier to use. In many previous work, the authors tested several base classifiers such as SVM, Naïve Bayes, CART in their model, but just the overall performances of different classifiers have been pointed out [15,25,40,47]. However, performances of different classifiers may vary in characteristics of datasets. For example, CART may perform well in datasets with high Imbalance Ratio(IR), but come up to a disappointed performance when classifying low dimension datasets. More specifically, IR, the number of features, the number of classes are all crucial factors that have to be considered when applying base classifier into the ensemble model. Therefore, in this paper, we divide imbalanced data into eight types based on their IR, dimension (the number of features) and the number of classes. We attempt to conduct an adaptive ensemble algorithm that is able to learn from different types of imbalanced data by different yet most efficiency algorithms constructed from a union ensemble paradigm.

While most MCSs take sampling methods as pre-processing, few literature has considered another common pre-processing technique, that is, feature selection. Feature selection is often separated as another issue for imbalanced learning, as is discussed [5,49] and [50]. These studies focus on developing novel feature selection algorithms, while the contribution of feature selection for imbalanced data classification is not clearly discussed. It is obvious that removing irrelevant and redundant features reduces the noise in the training space and decrease the time complexity [20,21]. For imbalanced case, samples of minority class are more easily to be ignored as noise, if we remove the irrelevant features in the feature space, the risk of treating minority samples as noise may also be reduced. [47,14] employed feature selection algorithm as a pre-processing procedure before carrying out classification, which gained good results. This motivates us to employ both feature selection and sampling method as pre-processes before training MCS.

Multi-class classification has been pointed out as a hard task for classification [40,19], due to that multi-class classification might achieve a lower performance than binary classification as the boundaries among the classes may overlap. This issue may become more complex when facing imbalanced data. In [40] the authors studied two ways of extending binary classification algorithms into multi-class case: One-versus-one approach (OVO) and One-versus-all approach (OVA). The conclusion, as they suggested, is OVO approaches gain better accuracy than OVA approaches. However, when considering computational complexity, OVO approaches may sacrifice too much on time cost when the number of classes increases. In their empirical study, OVA approaches also outperformed OVO approaches in some cases, which implies that there is no dogmatic approach that suit for all kinds of imbalanced data. It should be noted that the training of the OVA approach is inherently imbalanced, as the set of all data points from all other classes is likely to outnumber the representatives of the target class, for each sub-classifier [19]. Taking this into account, OVA approach may not suitable for high IR datasets. The third option of addressing multi-class imbalanced data is standard ad-hoc learning algorithms (algorithms that are natural for addressing multiple class learning problems), such as KNN, Naïve Bayes based ensemble algorithms. In our study, we specifically focus on multi-class imbalanced data. In order to build adaptive ensemble algorithm for different kinds of imbalanced data, OVO, OVA approaches and ad-hoc approaches will all be considered and we expect to find criteria to select the best approach for each type of data.

We argue that the above mentioned concerns are crucial issues that need to be clarified. Therefore, in our study, we attempt to build an adaptive ensemble learning algorithm for multi-class imbalanced data, which is called Adaptive Multiple Classifier System(AMCS). For adaptive learning, Three widely-accepted ensemble frameworks are considered, that are, Adaboost.M1 [46], Under-Sampling Balanced Ensemble(USBE) [15,47] and Over-Sampling Balanced Ensemble(OSBE) [16]. For the later two frameworks, five different ensemble rules (such as Max, Min, Product etc. shown in Table 2) to fuse sub-classifiers are optional. Moreover, as feature selection might avail to reduce the risk of treating minority samples as noise, in all the ensemble frameworks feature selection is employed as a pre-processing, for which both wrapper and filter feature selection techniques are considered. In empirical study, we first test the three ensemble frameworks with different ensemble rules and base classifiers, then conclude the adaptive criteria for different types of imbalanced data. Finally, we apply AMCS to oil-bearing reservoir recognition adaptively base on the characteristic of Jianghan well-logging data. Four significant contributions of our study are as follows:

(1) We present a comprehensive categorization of several recent works related to imbalanced data classification and highlight the need for an adaptive algorithm to solve different kinds of imbalanced data. To do so, we categorize imbalanced data into eight types based on their IR, dimension and the number of classes. For each type of data, the order of choosing feature selection algorithm, ensemble framework, base classifier and ensemble rule can be viewed as a *route* of framing a MCS, our algorithm can choose the best route for different types of data.
(2) The proposed ensemble method AMCS employs both feature selection and sampling method as pre-processes, in which feature selection may be an option when there is no irrelevant or redundant feature exits.
(3) We focus only on multi-class imbalanced data problems, which may be ignored by many previous studies. Since most classical performance metrics such as AUC are binary metrics, we enable a novel multi-class AUC metric called AUCarea to evaluate models by setting probabilistic outputs for both base classifiers and ensemble classifiers.
(4) The goodness of this novel adaptive methodology and the criteria of choosing routes to form AMCS are studied by means of thorough experimental analyses. Each node of the route is selected in a selection pool. The selection pool contains two feature selection methods, three ensemble frameworks, five base classifiers and five ensemble rules. In empirical study, all the possible routes for eight types of data are tested, the best routes for each type of data is selected as AMCS. The superior of AMCS compared with several existing methods is tested using various benchmarks and a real-world case of oil reservoir recognition.

The remainder of this paper is organized as follows. Literature related to imbalanced data are categorized in Section 2. The main framework of AMCS is described in Section 3, where the two feature selection methods, three ensemble frameworks and five ensemble

Please cite this article as: L. Yijing et al., Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data, Knowledge-Based Systems (2015), http://dx.doi.org/10.1016/j.knosys.2015.11.013