Engineering Applications of Artificial Intelligence 49 (2016) 176–193

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification

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

[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] School of Business, Central South University, Changsha, Hunan 410083, China

ARTICLE INFO

Article history:
Received 8 April 2015
Received in revised form
14 August 2015
Accepted 28 September 2015
Available online 29 October 2015

Keywords:
Imbalanced data
Ensemble
Feature selection
Classification
Oil reservoir

ABSTRACT

This paper proposes an ensemble algorithm named of BPSO-Adaboost-KNN to cope with multi-class imbalanced data classification. The main idea of this algorithm is to integrate feature selection and boosting into ensemble. What's more, we utilize a novel evaluation metric called AUCarea which is especially for multi-class classification. In our model BPSO is employed as the feature selection algorithm in which AUCarea is chosen as the fitness. For classification, we generate a boosting classifier in which KNN is selected as the basic classifier. In order to verify the effectiveness of our method, 19 benchmarks are used in our experiments. The results show that the proposed algorithm improves both the stability and the accuracy of boosting after carrying out feature selection, and the performance of our algorithm is comparable with other state-of-the-art algorithms. In statistical analyses, we apply Bland–Altman analysis to show the consistencies between AUCarea and other popular metrics like average G-mean, average *F*-value etc. Besides, we use linear regression to find deeper correlation between AUCarea and other metrics in order to show why AUCarea works well in this issue. We also put out a series of statistical studies in order to analyze if there exist significant improvements after feature selection and boosting are employed. At last, the proposed algorithm is applied in oil-bearing of reservoir recognition. The classification precision is up to 99% in oilsk81-oilsk85 well logging data in Jianghan oilfield of China, which is 20% higher than KNN classifier. Particularly, the proposed algorithm has significant superiority when distinguishing the oil layer from other layers.

1. Introduction

Classification is one of the crucial issues in the field of machine learning. Though many mature classification methods have been proposed since the last century such as Decision Tree, Naïve Bayes, Artificial Neural Network (ANN), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), most of those approaches are based on the assumption that the sample distribution among various classes are balanced. When facing imbalanced distribution, the traditional classifiers often come up to a disappointed performance (Galar et al., 2013). Therefore, traditional classifiers remain to be modified in addressing the problem of imbalanced data classification.

Imbalanced data refers to such a dataset in which one or some of the classes have much more samples in comparison to the others. The class that has far more samples is called the majority class, while a class that contains relative small number of samples is called minority class (earle, 1987). When addressing imbalanced data problems, people tend to care more about the minority class, and the cost of misclassifying those samples belong to minority class are much higher than the others (Menardi and Torelli, 2014; López et al., 2013; Maldonado et al., 2014). Taking cancer diagnosis for example, the number of cancer patient is much less than healthy people, if cancer patients are diagnosed as healthy people, they will miss the best treatment time, which may cause a disaster (Menardi and Torelli, 2014). So does oil-bearing recognition (oil-bearing recognition means to recognize the characters of each layer in the well (Feng et al., 1999; Guo et al., 2011)) that is studied in this paper, the class distribution of logging data is imbalanced and cost of misclassifying oil layer into other layers is much higher than other misclassification situations. Therefore, oil-bearing recognition is a typical imbalanced data classification problem.

The approaches to deal with imbalanced data recognition can be summarized into the following three categories:

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

http://dx.doi.org/10.1016/j.engappai.2015.09.011
0952-1976/© 2015 Elsevier Ltd. All rights reserved.

- Data level approaches: data level approaches focus on resizing the training datasets in order to balance all kinds of classes. Two main ideas of resizing are over-sampling and under-sampling, which are all based on sampling technologies. Dozens of over-sampling and under-sampling algorithms have been proposed before, such as Chawla put forward an algorithm called SMOTE (Chawla et al., 2002) to over-sample the instances of minority class and Dehmeshki proposed the data filtering technology (Vorraboot et al., 2015) to under-sample the samples of majority class.
- Algorithmic level approaches: algorithmic level approaches focus on carring out modification on existing algorithms/classifiers to strengthen their ability of learning from minority class (Galar et al., 2012; López et al., 2013). In this family there are various commonly used methods such as cost sensitive and one-class learning. (Galar et al., 2012; Cao et al., 2013). Cost sensitive algorithms attempt to increase the learning ability of classifiers by assigning larger misclassified cost for minority class samples. López et.al. (2015) proposed Chi-FRBCS-BigDataCS algorithm to deal with large-scale imbalanced data using a fuzzy rule and cost-sensitive learning techniques. Krawczyk et.al. (2014) constructed a fusion algorithm based on cost-sensitive decision tree ensembles, in which choice of cost matrix is estimated by ROC analysis. Nguyen et.al introduced two empirical cost-sensitive algorithms, one combined sampling, cost-sensitive and SVM and the other treated the cost ratio as a hyperparameter which needs to be optimized before training the final model (Thai-Nghe et al., 2010). One-class classifier only models single class by defining a boundary to classify binary classes datasets. According to different approaches of computing the boundary, one-class learning can be divided into the method based on density, neural network, clustering and SVM (Maldonado et al., 2014; Chawla et al., 2004; Naimul et al., 2014). However, the most popular one-class classifier that is used for imbalanced data classification is One-Class SVM(OSVM). Krawczyk et al. (2014) proposed a weighted OSVM to model the minority class. Tian et al. (2011) employed a SMOTE-OSVM algorithm to detect noises of minority samples, utilized the oversampled minority samples and a partial of majority samples to train SVM classifier. Krawczyk and Woźniak (2015) introduced an ensemble classifier, in which OSVM is used.
- New evaluation metrics: traditional evaluation metrics such as classification accuracy may ignore the minority class or treat them as noise (Galar et al., 2013; Menardi and Torelli, 2014). Performance metrics adapted into imbalanced data problems, such as Receiver Operating Characteristics (ROC) (Richard and Jonathan, 2006), G-mean, and *F*-value (López et al., 2013), are less likely to suffer from imbalanced distributions because they have less bias toward majority class. In fact most of algorithms cope with class imbalanced problems have abandoned using accuracy as performance metric, while ROC, G-mean and *F*-value are most popular used (López et al., 2013, 2015; Sun et al., 2015; Krawczyk et al., 2014). Recently, many modified metrics are proposed to evaluate the performance of imbalanced data classification problems. Gao et al. (2014) built several neuro-fuzzy models for imbalanced data classification, aimed to minimize the leave-one-out(LOO) mean square error(MSE). All of the neurofuzzy models used modified AUC and *F*-value as performance metrics called LOO-AUC and LOOFM, since the basic AUC and *F*-value are no longer applicable if MSE is chosen as the optimization goal. Krawczyk and Schaefer (2013) proposed a multiple classifier system, in which double-fault diversity measure is considered to prune those base classifiers that output similar hypotheses. For multi-class imbalanced problems, basic imbalanced data metrics also need to be modified due to most of those metrics are designed for addressing binary-class problems.

However, in recent years, several ensemble algorithms combined with different stratergies are proposed to cope with imbalanced data classification. Most of the ensemble algorithms are based on two specific algorithms, which are, the Boosting algorithm proposed by Schapire (Freund, 1995) and Bagging proposed by Breiman (1996). Sun et.al., (2015) proposed a novel ensemble strategy for imbalanced data classification, which converts an imbalanced dataset into multiple balanced subset and gets final hypothesis using an ensemble classifier , similar strategy can be found in (Díez-Pastor et al., in press). Nitesh et al. (2003) puts forward an algorithm called SMOTEBoost, which integrate over-sampling method SMOTE and boosting algorithm. Krawczyk and Schaefer, (2013) created an ensemble algorithm called PUSBE that contains sampling, pruning and boosting technologies. Nanni et al. (2015) denoted that these approaches generally contain 3 processes, that are, resampling, ensemble building, and voting for the final classification. In this family, ensemble algorithms like EasyEnsemble (Liu et al., 2009), EUSboost (Galar et al., 2013) are all well-accepted state-of-the-art. In this paper we also utilize boosting algorithm as classifier due to its efficiency in dealing with imbalanced problems.

Instead of proposing new methods, Prati et al. (2015) focused on setting up a series of experiments to assess the performance of some proposed treatment methods like SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008) and MetaCost (Wang et al., 2013) for imbalanced data. These treatment methods are all based on sampling technologies and cost sensitive learning, which are considered as the powerful methods to increase the learning ability of classifier to classify the minority class. In their research, they defined a value called “performance loss” to figure out whether all the learning models are equally affected by class imbalanced. Besides, they also defined a metric called “performance recovery” to evaluate how much of the performance losses caused by imbalanced distribution can be recovered by the treatment methods. The results showed not all the treatment methods are suit for all basic algorithms. For example, SMOTE are considered as the most common sampling method for imbalance data, but it seems to harm the performance of SVM and Naïve Bayes. This idea inspired us to find new treatments for imbalanced data instead of sticking on sampling methods.

While most of literatures focused on binary problems, Alberto et al. (2013) reviewed plenty of novel multi-class imbalanced data classification algorithms. They pointed out that multiple class classification might achieve a lower performance than binary classification as the boundaries among the classes may overlap. They studied two ways of extending binary classification algorithms into multi-class case: One-versus-one approach (OVO) and One-versus-all approach (OVA). It shows in their experimental studies that OVO approaches gain better results than OVA approaches. However, when comparing pairwise learning with standard ad-hoc learning algorithms (algorithms that are natural for addressing multiple class learning problems), there is no significant evidence to illustrate that pairwise learning algorithms are superior to standard ad-hoc learning algorithms. Considering the computational cost of pairwise learning is expensive, we would like to utilize standard ad-hoc approaches instead of pairwise learning.

Feature selection is often separated as another topic for imbalanced data problems, as is discussed in several literatures like (Maldonado et al., 2014; Yin et al., 2013; Shanab et al., 2011). These literatures all focus on developing novel feature selection algorithms, while the contribution of feature selection for imbalanced data classification is not clearly discussed. Krawczyk and Schaefer (2013) employed FCBF feature selection algorithm as a