Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Probability model selection and parameter evolutionary estimation for clustering imbalanced data without sampling

Jiancong Fan [a,b,c,*], Zhonghan Niu [b], Yongquan Liang [a,b], Zhongying Zhao [b]

[a] Provincial Key Lab. for Information Technology of Wisdom Mining of Shandong Province, Shandong University of Science and Technology, Qingdao, 266590, China
[b] College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[c] State Key Lab. for Novel Software Technology, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form
15 October 2015
Accepted 26 October 2015

Keywords:
Imbalanced data
Data mining
Clustering
Model selection

ABSTRACT

Data imbalance problems arisen from the accumulated amount of data, especially from big data, have become a challenging issue in recent years. In imbalanced data, those minor data sets probably imply much important patterns. Although there are some approaches for discovering class patterns, an emerging issue is that few of them have been applied to cluster minor patterns. In common, the minor samples are submerged in big data, and they are often ignored and misclassified into major patterns without supervision of training set. Since clustering minorities is an uncertain process, in this paper, we employ model selection and evolutionary computation to solve the uncertainty and concealment of the minor data in imbalanced data clustering. Given data set, model selection is to select a model from a set of candidate models. We select probability models as candidate models because they can solve uncertainty effectively and thereby are well-suited to data imbalance. Considering the difficulty of estimating the models' parameters, we employ evolutionary process to adjust and estimate the optimal parameters. Experimental results show that our proposed approach for clustering imbalanced data has the ability of searching and discovering minor patterns, and can also obtain better performances than many other relevant clustering algorithms in several performance indices.

1. Introduction

Class imbalance problem is a challenging issue in data mining community because most of the data mining algorithms always pay more attention to learn from the major samples. Some minor samples are often ignored and misclassified into major sample sets, which leads to high errors or low precisions of the discovered patterns. In class imbalance problem one or several classes only include much fewer data instances than others, while these fewer instances probably play a key role in some data mining tasks. For example, a great many of texts are published in the Internet every day, and we want to analyze these data and try to discovery the relations among hot topics or events. These topics or events are minor among all data in the Internet but they are important. Similar situations in real-world applications include medical diagnose, fraud detection, telecommunication analysis, Web mining, etc.

Most of the recent efforts focus on two-class imbalanced datasets, that is, one is minority class (also called positive class), and the other is majority class (also called negative class). However, multi-class imbalance problems also occur frequently. For example, some local events on the Web emerge during different periods. These local data instances are minority compared with other global event data but they are identically important with the global majority for considering different goals. Therefore, in this paper, we not only consider two-class imbalanced data clustering problem, but also attempt to solve multi-class problem.

There are two kinds of methods available for learning from imbalanced data. One is preprocessing approach such as sampling and feature selection, the other is algorithmic approach, such as ensembles of classifiers and modifications of current algorithms, which deals with imbalanced data by designing algorithms to construct learners. Up to now, most of the above approaches are fit for binary-class imbalance problem. In recent years, multi-class imbalanced data has attracted an increasing number of minds. However, most of the existed approaches are applicable for supervised classification area. There are only few discussions focusing on imbalanced data clustering problem.

In this paper, we contribute to study imbalanced data clustering algorithm via selection of probability models and parameter evolutionary estimation for these models. The reason why we

* Corresponding author at: Provincial Key Laboratory for Information Technology of Wisdom Mining of Shandong Province, Qingdao, 266590, China.
E-mail address: fanjiancong@sdust.edu.cn (J. Fan).

http://dx.doi.org/10.1016/j.neucom.2015.10.140
0925-2312/

apply probabilistic approach to class imbalance problem will be analyzed in the next section. In addition, the next section recalls the related work and their characteristics. Section 3 contains preliminaries and background including model selection, and parameter evolutionary estimation. In Section 4, we detail the proposed approach based on probability model selection without over- and under-sampling. Section 5 introduces the experimental setup including the data sets, selected probability models with their corresponding parameters and the statistical testing, and then presents the experimental results and analysis over the most significant algorithms for imbalanced data. Finally, in Section 6, we present our concluding remarks.

2. Related work

In this section, we first introduce related work about the class imbalance problem in classification and clustering. Then, we introduce the hot topics on imbalanced data clustering.

2.1. Related work about learning from class imbalanced datasets

As stated in Section 1, a dataset is said to be imbalanced when the quantity of data instances of one or several classes is much smaller than others. Furthermore, the positive class (smaller class) is equal with or more important than the negative one. We define an imbalance problem as multi-class imbalance when there is more than one positive class in a dataset. On the contrary, if there is only one positive class and one negative class in a dataset, this problem is called two-class imbalance.

Many studies and applications for classification of two-class imbalanced datasets have been existed in a great many of fields, such as medical diagnosis [1–5], fraud detection [6–8], credit assessment [9], malcode detection [10,11], and network traffic [12]. In addition to the application study, there are many empirical and theoretical studies for two-class imbalance problem [13–22].

In addition to the study and application for two-class imbalance problems, multi-class problems are attracts a growing number of interests. Although Zhou et al. [22] suggested that multi-class imbalance problem was more difficult than two-class task, and almost all the approaches were effective on two-class task, while most were ineffective and even might cause negative effect on multi-class task, there are an increasing number of contributions and endeavors to solve this problem [23–30]. Most of the literatures study this problem by some standard classification algorithms with over- or under-sampling, neural network, or only empirical methods. So far probability-based approaches have not been considered as the means of mining minority classes in skewed data distribution.

The above literatures mainly focus on the classification problem. However, in many real-world applications, unlabeled data are usually more common but the amount of labeled examples is often limited. Especially for the imbalanced data, the minority are probably not selected to label as training examples because of the small ratio in large dataset. In this situation, clustering imbalanced data is a practically applicable scenario and worthwhile to study.

2.2. The existed methods for clustering imbalanced data

It is well known that there are a large number of clustering methods that have been proposed and examined in many kinds of areas, both theoretically and empirically. However, there are only few methods for clustering imbalanced data [31–34]. A differential evolution clustering hybrid resampling algorithm was proposed and used for over-sampling process to enlarge the ratio of positive samples, which utilized the similar mutation and crossover operators of Differential Evolution (DE) to cluster the oversampled training dataset [31]. In [32], the K-means was used as the base learner of ensemble to research the clustering problem based on class imbalanced data. The cDNA microarray time-series imbalanced data was studied with PAM clustering algorithm in [33] but it did not investigate the clustering method how to influence the results. All of the above three imbalanced data clustering approaches used sampling process to balance the distribution of classes. A spectral clustering approach for imbalanced data without sampling was proposed in [34], which proposed a graph partitioning framework by parameterizing a family of graphs by adaptively modulating node degrees in a k-NN graph. Although the spectral approach did not utilize sampling but there existed two limitations: (1) main assumption was that prior knowledge of the smallest cluster size needed to be obtained in advance; (2) labeled samples for semi-supervised learning used in spectral clustering needed to be chosen with at least one sample from each class.

In multi-class imbalance problem, nevertheless, to the best of our knowledge, systematic approach of adapting clustering process to possibly multi-class imbalanced data has not appeared. Although multi-class imbalance problem has been paid more and more attentions, the research emphasis is put on supervised classification.

3. Preliminaries and proposed concepts

In this section, we will briefly introduce several basic concepts employed in our proposed approach such as model selection and probability-based model selection, evolutionary computation and evolutionary computation-based parameter estimation. We also specify the advantages of applying probability model selection approach to class imbalance problem.

3.1. Model selection and probability model selection

Model selection is the process of identifying the best approximating model for problems to be modeled and solved. The goal of model selection is to select approximately true predictive models to best fit the observed data. There are many approaches used by model selection such as maximum likelihood (ML), hypothesis testing (HT), Akaike's information criterion (AIC), Bayesian information criterion (BIC), and cross validation (CV), et al.

The main reason of employing ML is that ML is principally a method of parameter estimation which is one of the important steps in model selection. Thus ML extends straightforwardly to model selection. HT is a classical methodology of statistics which can be applied to many problems in model selection. HT is able to be applied to a situation as follows. Let x be a variable (also called parameter) ranging over a data set. The hypothesis θ=0 specifies a probabilistic density $f(x;\theta=0)$. Let $\hat{\theta}$ be the ML estimate for θ, and $\hat{\theta}$ is a function of x whose probability distribution is determined by θ=0 which is written by $g(\hat{\theta};\theta=0)$. If θ=0 is chosen as the imprecise hypothesis $\bar{H}$, we may set up a 5% critical region, or rejection set, such that if $\hat{\theta}$ is not equal to 0, $\bar{H}$ is not rejected.

The AIC's goal is to minimize the Kullback-Leibler (K-L) distance of the selected probabilistic density from the true density. In clustering, however, the true density is unknown due to the lack of training set and other priori information about data set. So AIC cannot be used in clustering approaches. BIC uses Bayes method to select models, which needs the prior probabilities of all models and then derives an asymptotic expression for the likelihood of each model. But the prior probabilities are difficult to be estimated because of the same reason as AIC. BIC is also not adopted in our proposed approach.

Please cite this article as: J. Fan, et al., Probability model selection and parameter evolutionary estimation for clustering imbalanced data without sampling, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.10.140