Neurocomputing 101 (2013) 309–318

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data

Hualong Yu [a,*], Jun Ni [b], Jing Zhao [c]

[a] School of Computer Science and Engineering, Jiangsu University of Science and Technology, Mengxi Road No.2, Zhenjiang 212003, China
[b] Department of Radiology, Carver College of Medicine, The University of Iowa, Iowa City, IA 52242, USA
[c] College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

ARTICLE INFO

Article history:
Received 25 December 2011
Received in revised form
25 August 2012
Accepted 26 August 2012
Communicated by T. Heskes
Available online 19 September 2012

Keywords:
DNA microarray
Ant colony optimization
Class imbalance
Undersampling
Support vector machine

ABSTRACT

In DNA microarray data, class imbalance problem occurs frequently, causing poor prediction performance for minority classes. Moreover, its other features, such as high-dimension, small sample, high noise etc., intensify this damage. In this study, we propose ACOSampling that is a novel undersampling method based on the idea of ant colony optimization (ACO) to address this problem. The algorithm starts with feature selection technology to eliminate noisy genes in data. Then we randomly and repeatedly divided the original training set into two groups: training set and validation set. In each division, one modified ACO algorithm as a variant of our previous work is conducted to filter less informative majority samples and search the corresponding optimal training sample subset. At last, the statistical results from all local optimal training sample subsets are given in the form of frequence list, where each frequence indicates the importance of the corresponding majority sample. We only extracted those high frequency ones and combined them with all minority samples to construct the final balanced training set. We evaluated the method on four benchmark skewed DNA microarray datasets by support vector machine (SVM) classifier, showing that the proposed method outperforms many other sampling approaches, which indicates its superiority.

1. Introduction

In the past decade, DNA microarray has been one of the most important molecular biology technologies in the post-genomic era. By this technology, biologists and medical experts are permitted to detect the activity of thousands of genes in a cell simultaneously. At present, DNA microarray has been widely applied to predict gene functions [1], investigate gene regulatory mechanisms [2,3], provide invaluable information for drug discovery [4], classify for cancer [5,6] and mining new subtypes of a specific tumor [7–9] etc. Among these applications, cancer classification has attracted more attentions. However, it is well-known that microarray data generally has some particular features, such as high-dimension, small sample, high noise and most importantly, imbalanced class distributions. Skewed class distributions will underestimate greatly the prediction performance for minority classes and provide inaccurate evaluation for classification performance, while the other features of microarray data will further intensify this damage [10]. Therefore, it is necessary to remedy this bias by some effective strategies.

In fact, class imbalance learning has drawn a significant amount of interest since 2000 from artificial intelligence, data mining and machine learning, which can be reflected by launch of several major workshops and special issues [11], including AAAI'00 [12], ICML'03 [13] and ACM SIGKDD Explorations'04 [14] etc. There are two major methods to solve class imbalance problem: sampling-based strategy and cost sensitive learning. Sampling, which includes oversampling and undersampling, deals with class imbalance by inserting samples for minority class or discarding samples of majority class [15–20]. While cost-sensitive learning treats class imbalance by incurring different costs for different classes [21–29]. Recently, some research also focused on ensemble learning built on multiple different sampling or weighting data sets with presenting excellent performance and generalization ability [30–35]. More details about class imbalance learning methods are presented in Section 2.

In this study, we introduce a novel undersampling method based on the idea of ant colony optimization (ACO), which is named ACOSampling, to classify for skewed DNA microarray data. In fact, this method is a modified version of our previous work [36], the difference between them is this work converts the information selection from feature space to sample space. First, the original training dataset is randomly and repeatedly divided into two groups: training dataset and validation dataset. Then for

* Corresponding author. Tel.: +86 511 88690470; fax.: +86 511 88690471.
E-mail address: yuhualong@just.edu.cn (H. Yu).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.018

each partition, ACOSampling is conducted to find the corresponding optimal majority class sample subset. Different from the traditional ACO algorithm, ACOSampling impels ants to leave from the nest, then to pass all majority class samples one by one, by either pathway 0 or pathway 1, at last to reach the food source, where pathway 0 indicates the corresponding sample is useless and should be filtered, while pathway 1 represents it is important and should be selected. Considering the particularity of the classification tasks in this study, the overall accuracy is not an excellent measure as the fitness function, thus we construct it by three weighted indicative metrics, namely *F-measure*, *G-mean* and *AUC*, respectively. After that, many local optimal majority class sample subsets can be generated by iterative partitions, so the significance of each majority sample may be estimated according to its selection frequence, i.e., the higher the selection frequence, the more information the corresponding sample can provide. Next, a global optimum balanced sample set can be created by combining the highly ranked samples of majority class with all examples of minority class. At last, we construct a SVM classifier upon the balanced training set for recognizing future unlabeled samples.

The remainder of this paper is organized as follows. Section 2 reviews some previous work related with class imbalance problem. In Section 3, the idea and procedure of ACOSampling method is described in detail. Experimental results and discussions are presented in Section 4. At last, we conclude this paper in Section 5.

2. Previous work

As mentioned in the Section 1, the existing class imbalance learning methods could be roughly categorized into two major groups: sampling strategy and cost sensitive learning. Here, we pay special attention to sampling strategy because it is more related with our study.

The sampling is actually a re-balancing process for the given imbalanced data set. It can be distinguished into oversampling and undersampling. Oversampling, as its name indicates, increases some samples belonging to minority class, while undersampling takes away some examples of majority class. The simplest sampling methods are Random Over Sampling (ROS) and Random Under Sampling (RUS) [15]. The former will make the learner to be overfitting by simply duplicating some samples of minority class, while the latter may lose some valuable classification information due to many majority examples are randomly removed [11]. To overcome their drawbacks, some complicated sampling methods were developed. Synthetic Minority Over-sampling TEchnique (SMOTE), proposed by Chawla et al. [16], can create artificial data based on the feature space similarities between existing minority examples. Specifically, randomly select one sample x_i in minority class, find its K-nearest neighbors belonging to the same class by Euclidian distance. To create a synthetic sample, randomly select one of the K-nearest neighbors, then multiply the corresponding feature vector difference with a random number between [0,1], and finally, add this vector to x_i. Han et al. [17] observed that most misclassified samples scatter around the borderline between two categories, then presented two improved versions of SMOTE, Borderline-SmOte1 (BSO1) and Borderline-SmOte2 (BSO2), respectively. For BSO1, SMOTE only runs on those minority class samples near borderline, while for BSO2, it generates synthetic minority class samples between each frontier minority example and one of its K-nearest neighbors belonging to majority class, thus mildly enlarges decision region of minority class. One Side Selection (OSS) has very similar idea with BSO2. It shrinks the decision area of majority class by cleaning noisy samples, redundant samples and boundary examples in majority category [18]. As another improved oversampling method, Adaptive Synthetic Sampling (ADA-SYN) uses a density distribution as criterion to automatically decide the number of synthetic samples that need to be generated for each minority example by adaptively changing the weights of different minority class examples to compensate for the skewed distributions [19]. Another sampling method using density distribution is Under-sampling based on clustering (SBC), presented recently by Yen and Lee [20]. SBC may automatically decide to remove how many majority class samples in each cluster, according to the corresponding density distribution. García *et al.* [37] have simply compared two kinds of sampling strategies and found oversampling generally produces better classification performance when the dataset is highly skewed, while undersampling is more effective when imbalance ratio is very low. All in all, sampling possess many advantages, such as simple, intuitive, low time complexity and low storage cost, thus it can be more convenient to apply in real-world imbalanced classification tasks. In Section 4, we would investigate the performance of the proposed ACOSampling method compared with original data without sampling (ORI) and several benchmark sampling strategies described above, such as ROS, RUS, SMOTE, BSO1, BSO2, OSS, ADA-SYN and SBC.

Cost sensitive learning methods consider the costs associated with misclassifying samples [21]. Instead of creating balanced data distributions through sampling, cost-sensitive learning assigns different costs for the samples belonging to different classes by creating a cost matrix. Based on the cost matrix, misclassifications on the minority class are more expensive than the majority class. Moreover, cost sensitive learning pursues to minimize the total cost but not error rate, thus the significance of minority class is highlighted. There are generally three kinds of cost sensitive learning methods. The first one is based on translation theorem [22]. It applies misclassification costs to the data set in terms of data-space weighting. The second class, building on metacost framework [23], uses cost-minimizing techniques to the combination schemes of ensemble methods. Some existing research has combined these two strategies, such as AdaCX series algorithms [24] and AdaCost [25]. The last class of cost sensitive learning methods directly designs appropriate cost functions for specific classifier, including cost-sensitive decision tree [26], cost-sensitive neural network [27] and cost-sensitive support vector machine [28] etc. In some application fields, it has been demonstrated that cost sensitive learning is superior to sampling approaches [29]. However, it is difficult to pre-design an appropriate cost function when class imbalance problem occurs [27].

In recent several years, ensemble learning has also become popular to be employed for solving class imbalance problems. Generally speaking, in this technology, ensemble learning framework is incorporated with sampling approach or weighting strategy to acquire better classification performance and generalization capability. Chawla et al. introduced SMOTE into Boosting ensemble learning framework to develop the SMOTEBoost learning method [30]. Unlike the base classifiers generation strategy in traditional Boosting, SMOTEBoost promotes weak classifiers through altering distributions for the samples of different classes by SMOTE. Liu et al. combined RUS and AdaBoost classifier to overcome deficiency of information loss of traditional RUS method and presented two ensemble strategies: EasyEnsemble and BalanceCascade [31]. In contrast with Boosting framework, Bagging seems to leave less room to be modified for class imbalance problem. However, there are still some improved versions about Bagging, including Asymmetric Bagging (asBagging) which has been used to retrieve image [32] and predict drug