Neurocomputing 119 (2013) 350–358

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Certainty-based active learning for sampling imbalanced datasets ☆

JuiHsi Fu *, SingLing Lee

Department of Computer Science and Information Engineering, National Chung Cheng University, 168 University Road, Minhsiung Township, 62162 Chiayi, Taiwan, ROC

ARTICLE INFO

Article history:
Received 30 April 2012
Received in revised form
13 March 2013
Accepted 18 March 2013
Communicated by M. Wang
Available online 24 May 2013

Keywords:
Active learning
Imbalanced data classification
Neighborhood exploration
Certainty-based neighborhood
Local classification behavior

ABSTRACT

Active learning is to learn an accurate classifier within as few queried labels as possible. For practical applications, we propose a Certainty-Based Active Learning (CBAL) algorithm to solve the imbalanced data classification problem in active learning. Without being affected by irrelevant samples which might overwhelm the minority class, the importance of each unlabeled sample is carefully measured within an explored neighborhood. For handling the agnostic case, IWAL-ERM is integrated into our approach without costs. Thus our CBAL is designed to determine the query probability within an explored neighborhood for each unlabeled sample. The potential neighborhood is incrementally explored, and there is no need to define the neighborhood size in advance. In our theoretical analysis, it is presented that CBAL has a polynomial label query improvement over passive learning. And the experimental results on synthetic and real-world datasets show that, CBAL has the ability of identifying informative samples and dealing with the imbalanced data classification problem in active learning.

1. Introduction

Since labeling data by human labors is a time-consuming job, many active learning approaches [1–6] have been proposed to decrease the requirement of labeled samples. Recent applications are involved in image/video annotation [7] and video concept detection for video indexing [8]. The framework of active learning is to identify informative unlabeled data and then query their true labels from human experts or oracles. The objective is to learn an accurate classifier using as few queried labels as possible.

In general active learning approaches [1,9–12], labels of unlabeled samples are queried to adjust potential classifiers (also called *hypotheses* in the process of active learning) for improving the prediction ability. In other words, each queried sample is usually selected according to the classification performance of existing hypotheses. However, while the underlying data distribution is skewed, the hypotheses would suffer from the problem of *imbalanced data classification* [13,14] in active learning. Since the majority class overwhelms the minority class, the majority samples would be frequently queried. Then the queried dataset also has a skewed data distribution, and the biased hypothesis, of which prediction is dominated by majority samples, is returned as the resulting classifier.

In this paper, we propose an approach to deal with the imbalanced data classification problem in active learning. We focus on specific local classification behaviors, rather than the whole global one, in dealing with the imbalanced learning problem. Within the specific area (neighborhood) the importance of an unlabeled sample is measured so that it is not affected by irrelevant samples which might overwhelm the minority class. Notably each neighborhood is incrementally explored, and there is no need to pre-define its size. It should also be concerned that, when noise exists in the labeled set, called the agnostic case, data is not perfectly separable. Then the hypothesis that performs badly on some samples might be the optimal hypothesis. For handling the agnostic case, the strategy in [12] guarantees there is only small difference between the optimal hypothesis and the resulting one. Through theoretical analysis, the hypothesis difference is able to interpret the sample importance as a query probability. This strategy to handle the agnostic case is integrated into our approach without any cost. Thus our approach is proposed to determine the query probability within an explored neighborhood for each unlabeled sample. It is presented that our approach has a polynomial label complexity improvement over passive learning on the number of queried samples. And through experimental results on the synthetic and real-world datasets, our approach has the ability of identifying informative samples and dealing with the imbalanced data classification problem in active learning. It is concluded that, the proposed approach is a suitable solution for practical applications. Specifically, our main methods in this paper include:

- *Utilization of the local classification behavior*: Rather than the entire imbalanced dataset, only labeled samples in the neighborhood are used for generating hypotheses to present local

☆This work is supported by NSC, Taiwan, ROC under Grant no. NSC 99-2221-E-194-023.

* Corresponding author.
E-mail addresses: fjh95p@cs.ccu.edu.tw (J. Fu), singling@cs.ccu.edu.tw (S. Lee).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.03.023

behaviors. In other words, those hypotheses are specifically trained without being affected by irrelevant data in imbalanced datasets. In active learning, while hypotheses are generated under the skewed distribution of queried samples, the problem of imbalanced data classification would be effectively handled by our proposed approach.

- *Exploration of the specific neighborhood*: Previous research works [15–17] that utilize local behavior for learning problems had to define the neighborhood size in advance. In this paper, we explore the neighborhood for each sample within which the potential query probability is obtained as large/small as possible. There is no need to define the neighborhood size in our approach. And the neighborhood is incrementally explored using exponentially-increased sizes and the binary-search strategy. So that, the requirement of computing time could be significantly decreased.

The rest of our paper is organized as follows. Related work is reviewed in Section 2. In Section 3, our proposed approach is explicitly explained to utilize local behavior within the neighborhood for determining probabilities of querying samples. Experimental results are presented in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

2.1. Active learning

Active learning approaches are designed to query a small number of unlabeled samples and expected to perform as well as passive learners. An intuitive sample selection strategy is to minimize the expected errors of the outputted classifier trained by the queried (training) data. Several results [2,18,19] have selected informative samples by estimating the reduction of expected errors. Besides, samples are usually queried while they could eliminate inconsistent hypotheses [3,20–22] or their assigned labels are predicted in low confidence [1,23]. In order to deal with the agnostic case, authors in [24] assign penalties values to experts/oracles to improve the quality of labeling results. And in several algorithms [9,10,12,22,25,26], the abilities of dealing with noisy queried samples have been presented through theoretical analysis. Those results fall into two main categories in terms of the existence of input samples. The first category is that unlabeled samples are collected in a data pool. Active learners should select informative samples from the pool and query their true labels to build a labeled dataset. Before identifying useful unlabeled samples, A^2 [9,10] maintains a set of potential hypotheses without being filtered out while they perform badly on the agnostic case. Those hypotheses should have less than the expected prediction mistakes. Then the uncertain unlabeled samples, of which predicted labels are disagreed by any two potential hypotheses, are queried because they are useful to eliminate inconsistent hypotheses. At last the best hypothesis in the candidate set is returned until the criteria for a low error rate is met. It is theoretically proven in [9] that A^2 has polynomial improvement over passive learning on the required labeled samples.

While the input samples come as a stream, only one unlabeled sample comes at each time and should be determined whether being queried or not. After that, the determination is never changed. IWAL-ERM [12] presents an approach to handle the agnostic case without maintaining a set of hypotheses. Each queried sample is selected depending on the two specific hypotheses, the best hypothesis and the alternative hypothesis. Then the query probability is determined according to the difference between those two hypotheses. Let X be the input space, $Y=\{+1,-1\}$ be the labels, and L be the labeled dataset with pairs of $x\in X$ and x's true label $y\in Y$. H is a set of hypotheses mapping from X to Y, and $err(h,L)$ is the empirical error of the hypothesis $h\in H$ on L. Given the kth coming unlabeled sample $x_k\in X$, $h_k=\text{argmin}\,\{err(h,L):h\in H\}$ is the existing best hypothesis and $h'_k=\text{argmin}\,\{err(h,L):h\in H\wedge h(x_k)\neq h_k(x_k)\}$ is the alternative hypothesis. Let G_k be the difference between errors of h_k and h'_k. The query probability P_k is assigned to x_k depending on Eq. (1)

$$P_k=p(G_k,k)=\begin{cases}1 & \text{if } G_k\le\sqrt{\frac{C_0\log k}{k-1}}+\frac{C_0\log k}{k-1}\\ s & \text{otherwise.}\end{cases} \tag{1}$$

where $s\in(0,1)$ is the positive solution to the equation

$$G_k=\left(\frac{c_1}{\sqrt{s}}-c_1+1\right)\cdot\sqrt{\frac{C_0\log k}{k-1}}+\left(\frac{c_2}{s}-c_2+1\right)\cdot\frac{C_0\log k}{k-1}$$

where C_0, c_1, and c_2 are constant values. A low query probability P_k is obtained based on the large difference G_k, and that also means it is insignificant to query x_k for discriminating h_k and h'_k.

Let h^* be the optimal hypothesis and $err(\cdot)$ be the function of estimated errors of the hypothesis. In [12], it is proven that IWAL-ERM satisfies $P_n\ge 1/n^n$. And using other derived lemmas, Theorem 1 guarantees the bound of the difference between $err(h_n)$ and $err(h^*)$. That is, only small difference between the resulting classifier and the optimal classifier. It is also presented that the expected number of label queries has polynomial improvement over passive learning.

Theorem 1. *In IWAL-ERM, for any $n\ge 1$, the following holds with probability at least $1-\delta$:*

$$err(h_n)\le err(h^*)+\sqrt{\frac{2C_0\log n}{n-1}}+\frac{2C_0\log n}{n-1}.$$

And with probability at least $1-\delta$, the expected number of labels queried by IWAL-ERM after n iterations is at most

$$1+\theta\cdot 2err(h^*)\cdot(n-1)+O(\theta\cdot\sqrt{C_0 n\log n}+\theta\cdot\log^3 n),$$

where θ is the disagreement coefficient.

For handling the agnostic case, IWAL-ERM is integrated into our approach without any cost. Our analysis shows that the proposed approach still follows IWAL-ERM on the expected number of queried labels and quality of the resulting classifier.

2.2. Imbalanced data classification

While classifiers are built by the imbalanced datasets, the minority class is usually overwhelmed by the majority class [13]. Then predicted results of those classifiers are dominated by the majority class. Cost-sensitive approaches [27,28] give penalties to misclassification in different classes, but in practical it is difficult to predefine the proper penalty for each class. The one-class support vector machines (SVM) [29] is to describe the distribution of the majority class. Then the samples of which distribution is different from that of the majority class are identified as minority samples. Nevertheless only one class is modeled, rather than generating a separating hyperplane for majority and minority classes. The difficulty is to precisely define the separating boundary for different class samples. And ambiguous/uncertain samples are likely to be misclassified. Furthermore, some sampling approaches are proposed to address this imbalanced problem. In order to generate a balanced dataset from the original imbalanced one, the *undersampling* approach [30] is to filter out irrelevant samples, and relatively the *oversampling* one [31] synthesizes new samples. Although the class sizes are almost equal, the real class distribution is altered. Then the prediction model trained by the new balanced dataset might not recognize labels of future data accurately. Recently several studies