Neurocomputing 158 (2015) 62–72

Contents lists available at ScienceDirect

Neurocomputing

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

Actively constructing an effective training set by expected gain maximization criterion

Weining Wu [a,*], Shaobin Huang [a], Maozu Guo [b]

[a] College of Computer Science and Technology, Harbin Engineering University, No. 145, Nantong Street, Harbin, Heilongjiang, China
[b] School of Computer Science and Technology, Harbin Institute of Technology, No. 92, West Dazhi Street, Harbin, Heilongjiang, China

ARTICLE INFO

Article history:
Received 24 August 2014
Received in revised form
24 January 2015
Accepted 31 January 2015
Communicated by M. Wang
Available online 10 February 2015

Keywords:
Object categorization
Active learning
Sampling strategy
Experimental design
Different distribution

ABSTRACT

We address the problem of active learning which aims to select the most representative points. Out of many existing active learning techniques, close-to-boundary criterion has received considerable attention recently. The goal of typically uncertain-based sampling is to minimize the distance between the examples and the classification boundary. However, these methods only adapt to the independent identically distributed (i.i.d) data, while the non-i.i.d is ignored. In this paper, we propose an active scheme which takes into account of the situation that unlabeled data and training data may come from different distribution, and we approximately obtain the information of every example by expected gain maximization. Given the parameters of a classification model and a pool of unlabeled data, the real risk on the unknown distribution is estimated. The expected gain of every candidate example is obtained, and then the most representative examples of all are thus defined as those who can maximize the expected gain of the classification model. At last, a sampling sequence is acquired by solving the optimization problem. Moreover, we also give an active scheme of two-stage sampling strategy based on the criterion of expected gain maximization in order to obtain an effective training set. The experimental results on MIRFLICKR and Caltech-256 datasets have demonstrated the effectiveness of our proposed methods.

1. Introduction

It is a fundamental problem to estimate the relation between inputs and outputs. In many real tasks, there are always a huge number of unlabeled examples while there are a few training examples with precise annotations, because it is an expensive and time-consuming process to collect precise labels. In order to address this practical problem in applications, both semi-supervised learning [1] and active learning [2,3] are main technology which directly explores potential information contained in unlabeled data. In either method, active learning algorithms reduce total labeling costs by assuming a learner can freely control input distribution. In a standard process of active learning, some unlabeled examples, which are considered as the most informative ones of all, are submitted to human experts and queried for labels. And then these selected examples are added into the training set and used for estimating the statistical structure behind the whole unlabeled data. Clearly, the most remarkable property of active learning is that input examples are sampled according to a training distribution, and it is important that how to design an effective sampling strategy. Now, considering the following three situations:

One reason is that an effective sampling strategy should be designed to have an ability to solve the problem of extreme class imbalance. It is hard to make use of existing sampling strategies of active learning to obtain a classifier with low error, especially when there are a lot of negative examples but a few positive examples in the unlabeled pool. It needs to design the training distribution from an unbalanced dataset in order to reduce labeling costs for negative examples which are not necessary for the predictive model.

Another reason is that most existing sampling strategies from active learning mechanism are often assumed to be applied when observed data and unlabeled data are independent identically distributed [4]. Obviously, the assumption is too rigorous to be satisfied in most real tasks, particularly there are much more unlabeled data than observed data or unlabeled data have to be collected from the external environment that dramatically changes, e.g. online environment. Without using the assumption of independent identical distribution, there is few existing sampling strategies can improve accuracy of an active learner effectively and efficiently.

The last reason is that sampling strategies based on distribution optimization of input data need to estimate the real risk of a

* Corresponding author.
E-mail address: wuweining@hrbeu.edu.cn (W. Wu).

http://dx.doi.org/10.1016/j.neucom.2015.01.065
0925-2312/

classification model on unlabeled data in order to iteratively select the most representative and informative examples of all in a sequential process [5,6]. Since it is considered as an effective method to reduce total labeling costs and successfully avoid the impact from noisy or mispecified data, the real risk on unlabeled data should be taken into account when a novel sampling strategy is developed. More importantly, the real risk should be directly estimated by using unsupervised technology in order to decrease the deviation between real values and approximate values.

Based on above reasons, this work is motivated by real problems of active sampling strategies. After given a large number of unlabeled examples which may come from the biased or unknown distribution, the objective is to do the best to select the most representative examples of all for annotations, i.e. predicting whether an unlabeled example is the most helpful of all for training a classifier, with minimal labeling workload for human experts. We propose an active sampling strategy of error reduction based on the real risk in order to iteratively obtain an effective training set on the condition of non-independent identified distribution. Our contributions can be summarized as follows:

(1) We construct an effective training set on a group of carefully selected examples which are most representative and helpful of all to the current classifier. And then we can obtain a relatively balanced number of positive and negative examples for learning a classifier.
(2) We evaluate the information contained in every unlabeled example without using the assumption of independent identified distribution, and then iteratively select the example which can extremely reduce expected error of risk in every active round.
(3) We develop a novel algorithm for exploiting potential correlation between the unknown distribution of unlabeled data and the observed distribution of labeled data. In other words, the active sampling is designed to have an agnostic ability of overcoming noisy or mispecified data in order to reduce the classification error.

The rest of this paper is organized as follows: in Section 2, we summarize the related work; in Section 3, we describe and explain the proposed algorithm in detail; we give the experimental results on MIRFLICKR and Caltech-256 datasets in Section 4; finally, we offer conclusions and discussions in Section 5.

2. Related work

Until now, active learning algorithms have been widely used in order to help human experts from the workload in providing annotations. Here, one goal of active learning is to do human experts a favor to find the most helpful examples of all at an early stage of training a classifier. The active learning algorithms have play an irreplaceable role which has been proved by reducing the labeling costs of training set in obtaining a highly accurate classifier [2,7,8]. In an universal paradigm of active learning systems, the problem is addressed by designing an effective and efficient sampling strategy which is used to evaluate the information of a group of pre-collected examples and then select the most representative examples of all for querying labeling. However, in order to obtain the "representative" examples from a raw dataset, it is a challenging task of comparing the impact of every unlabeled example after it is labeled. The general evaluation of sampling methods is not technically viable because it depends on both characteristics of the unlabeled data and the classification model.

When the classification model is realistic or the unlabeled data can be specified or unbiased, the most representative examples of all can be obtained by two main approaches, i.e. uncertain-based sampling [7,10–12] and expected error or variance reduction [4,8,11]. The former method equally treats all unlabeled examples and prefers the most confused example of the current classification model, while the latter method weights individual examples by utilizing density estimation or predicted outputs as prior knowledge. Now, we give some comparisons and analysis between the two methods as follows:

In uncertain-based sampling methods, the pool of unlabeled data is regarded as identical with labeled data, and then the predicted outputs of the classification model are given for evaluating every unlabeled example. For example, Tong and Koller [10] propose a close-to-boundary criterion in which the examples are selected according to their distances of classification boundary and the nearest example of all is labeled. Seung et al. [9] propose the Query-by-Committee (QBC) method to select the most confused example of all by generating numerously viable classifiers on observed labeled examples, and also submit the examples confused by all these classifiers. Although the uncertain-based sampling is widely used and its practical value has been demonstrated with real tasks of many fields, it is still faced the adverse impact coming from noisy data. Additionally, the effectiveness of uncertain-based sampling methods ultimately decreases as the size of the training set increases [12].

In order to overcome the above shortcomings, the sampling methods of risk or expected error reduction have been developed recently [8,13,15]. In these methods, if binary labels are respectively added for every example, then the expected error or variance of the classification model is estimated. For example, Hoi and Lyu [13] propose a risk reduction criterion with the graph-based semi-supervised learning algorithm [13]. Yan et al [14] extend the sampling criterion of risk reduction to multi-label tasks. Roy and McCallum [8] propose a method of expected error reduction with Monte Carlo sampling. Although the sampling methods of risk or expected error reduction have gotten around that negation fairly easily to some extent, it is usually computationally intensive because of traversing all unlabeled examples and updating the classification model. Obviously, it is not suitable to apply in the large-scale tasks.

On the other hand, in practice, the distribution of unlabeled data and training data is usually different and unknown. So that some researchers think that an effective training set should be constructed by considering the distribution difference between observed data and unlabeled data. When the classification model is non-realistic or unlabeled data is mispecified or biased, the optimized training data can be achieved in two approaches, i.e. PAC-guarantee active learning algorithms (agnostic active learning) [15–17] and importance weighted algorithms [5,6]. Based on the intentions that some classification hypothesis with bad performance on unlabeled data may be the optimized ones with respect to training data [17], the main idea of PAC-guarantee methods is to obtain an uncertain field by sampling some unlabeled points. Since rigorous sampling or label complexity bounds are used in the sampling process, some researchers [6] indicate that the generalization bounds of the classification model are over loose, and then labeling costs are more than necessary.

JOLT3609The other active sampling methods used in the non-realistic situation are importance weight active learning [5,6]. Similar to the importance sampling, the basic idea is to make the classification model obtained on the training set be applied on unknown distribution by adding the density ratio, and then acquire an asymptotically unbiased classifier [18–20]. According to the optimal experiment designs [23], an consistent estimator of parameters is obtained on the training distribution which is directed by the active sampling, and other inconsistent estimators of parameters can be eliminated regardless of the training