Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

A probabilistic model of active learning with multiple noisy oracles

Weining Wu, Yang Liu, Maozu Guo*, Chunyu Wang, Xiaoyan Liu

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China

ARTICLE INFO

Article history:
Received 17 May 2012
Received in revised form
20 February 2013
Accepted 24 February 2013
Communicated by M. Wang
Available online 6 April 2013

Keywords:
Active learning
Noisy oracle
Oracle selection
Accuracy estimation

ABSTRACT

In this paper, we focus on obtaining an accurate classifier in active learning, when there are multiple noisy oracles with different and unknown levels of expertise to provide labels for selected instances. We propose a probabilistic model of active learning with multiple noisy oracles (PMActive). Our goal is formulized as to select the most reliable oracle and estimate the actual label on training data. When an instance is selected in every round of active learning, we firstly model the accuracies of individual oracles based on observed noisy labels, and select the most reliable oracle of all to provide a label for the instance. After adding the new instance-label pair into the training set, the actual label of the instance is estimated and used for enhancing the performance of the current classifier. The experimental results indicate that the PMActive method can work with different noise levels of oracles. Compared with the baselines which are commonly used in this area of active learning, the PMActive method is superior in obtaining a more accurate classifier.

1. Introduction

In supervised learning methods, all labeled data are requested at beginning and the parameters of the classifier are set by using all training data given to them. By contrast, in active learning algorithms, the classifier is trained on the labeled data which are selected by the sampling strategy in an iterated process. The aim of standard active learning algorithms is to achieve high performance of the classifier by using as few labeled instances as possible, thereby minimizing the total labeling costs when the data are abundant but the labels are scarce or expensive [1]. Until now, active learning algorithms have been widely applied to collecting training data in real tasks, such as text classification [2–4], information extraction [5,6], image classification and retrieval [7,8], video annotation and retrieval [9–11], speech recognition [12], cancer diagnosis [13] and so on. In these applications, active learning algorithms provide an effective way of tackling hard tasks of obtaining labels by asking a single oracle, who is assumed to never make mistakes, to label some of the most informative instances.

In recent years, many labeling tasks are finished by online tools [14,15] in which there are multiple oracles to provide labels because their costs are less than asking a single expert without making mistakes. However, in these applications, it is impossible to get completely correct labels from individual oracles for training a classifier. The reasons can be briefly described as follows: (a) all the oracles are possible to make mistakes; (b) the levels of expertise of individual oracles are different and unknown; and (c) the returned labels by asking oracles contain noise.

With the assumption that all the oracles have identical level of expertise, their majority vote [16] is used as the actual label of the given instance. The classifier is obtained without considering reliabilities of individual oracles. But when the expertise of oracles is very different, it is hard to obtain an accurate classifier by the majority vote because the qualities of collected labels decrease as the number of poor-quality oracles increases. Thus, two main problems are needed to be solved as follows:

(a) How to choose a reliable oracle based on his/her labeling accuracy, which has been proved to be crucial to mitigate the effects of label noise [16–18]? Since it is hard to estimate the accuracies of individual oracles without using additionally labeled instances as the "golden standard".
(b) How to obtain the actual labels of the instances? Since the returned labels by querying oracles contain noise, it will receive adverse effects if the noisy labels are directly used for training the classifier. Hence, the actual labels are needed to be estimated from the observed noisy labels, and moreover, the actual labels are also needed for evaluating the accuracies of individual oracles in (a).

In this paper, we focus on obtaining an accurate classifier in active learning with multiple noisy oracles when the levels of their expertise are very different and unknown. In order to solve this problem, we propose a probabilistic model of active learning with multiple noisy oracles (PMActive). In the proposed probabilistic model, our goal is divided into the selection of the reliable oracle

* Corresponding author. Tel.: +86 136 54646103; fax: +86 451 86221048.
E-mail addresses: maozug@hotmail.com, maozuguo@hit.edu.cn (M. Guo).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.02.034

and the estimation of actual label. The two objectives can be summarized as follows:

(a) Selecting the reliable oracle: We model the accuracies of individual oracles based on their provided noisy labels. In every round of active learning, the accuracies of all oracles are estimated. And then, the most reliable oracle of all is selected for labeling the selected instance.
(b) Estimating the actual label: We estimate the actual label of the selected instance by considering the current accuracies of individual oracles. Then, the actual label is used for training the classifier and evaluating the oracles in next rounds.

The experimental results on twelve UCI datasets and 20-newsgroup corpus indicate that the PMActive method is capable to select the most reliable oracle of all, when the levels of expertise of multiple oracles are very different. The PMActive method can obtain a more accurate classifier than other commonly used methods of active learning.

The rest of the paper is organized as follows: in Section 2, we summarize the related work; in Section 3, we describe and explain the proposed PMActive method in detail; we give the experimental results in Section 4; finally, we offer the conclusions in Section 5.

2. Related work

Active learning is considered as a kind of machine learning technique that learns a model in an interactive way [10]. In comparison with passive learning, the model can freely select the most representative instance in every round and improve its performance in an iterative manner. The instances which can maximize the performance or reduce future error of the current model are usually defined as the most informative ones [25]. In a variety of scenarios, there have been widely explored a lot of sampling criteria which are used to evaluate the information of selected instances. For example, risk reduction, uncertainty, diversity, density, relevance and so on. Furthermore, these basic criteria are used to be combined in order to fit more complicated applications. In multi-media retrieval and annotation, Wang et al. [11] propose an effective criterion by combining the uncertainty, diversity and density in order to learn the complex semantic concepts. The model in [11] can achieve favorable performance with the same amount of human annotations.

In traditional active learning, one main assumption is that there is only a single oracle to provide a correct label for the selected instance in every round. But when there are multiple oracles with different and unknown levels of expertise, the assumption lacks reality, and the labels provided by the oracles for the selected instance contain noise. In this situation, the accuracy of the classifier strongly depends on annotation qualities in the training process, and the actual label of the selected instance is needed to be estimated from noisy labels.

Until now, there have existed some works which focus on enhancing the accuracy of the classifier by obtaining an actual label for the selected instance. For example, Sheng et al. [16] propose a repeated labeling method in which the selected instance is labeled by querying all oracles and their majority vote is used as the actual label of the given instance. Donmez et al. [18] propose the IEThresh (Interval Estimate Threshold) method to repeatedly label the selected instance by querying some oracles with relatively high quality, and also obtain the actual label by the majority vote. The above methods of majority vote are widely used in this area of active learning with multiple oracles, but their practical values vary greatly with different levels of skills and expertise of oracles. Since it is needed to have the selected instance labeled by multiple oracles at the same time, some arguments are provided that the majority vote methods are wasteful. Additionally, the effectiveness of majority vote methods ultimately decreases as the size of the training set increases [19,20].

In the above-mentioned methods, the expertise of being queried oracles is regarded as identical, and then gives them the same weight in estimating the actual label. However, in practice, the expertise of individual oracles is different and unknown, and then some researchers think that the accuracy of the classifier should be enhanced by considering both the actual label of the selected instance and the labeling qualities of individual oracles. For example, Snow et al. [15] propose a method to select oracles by modeling the reliabilities of individual labelers. In their work, some additional instances with ground-truth labels are exploited to judge the annotation qualities of the oracles and estimate the actual labels. But in realistic works, there are always no instances with correct labels to be used as the evaluation. Considering this point, Donmez et al. [21] select some of all the oracles to compose a committee by estimating their accuracies. The actual label is directly estimated from observed noisy labels of the selected committee. Yan et al. [22] select oracles based on the uncertainty of their labels. The most uncertain label is used for estimating the actual label. In all these works, the accuracies of individual oracles are supposed to be described by a specified function (e.g. truncated Gaussian distribution in [21] or logistic function in [22]). After firstly given labeling qualities of individual oracles in the current round, the actual label is estimated from current noisy labels. Whereas there are some works [19,23] which consider that the actual label and labeling qualities of individual oracles influence each other in every round, and they focus on jointly learning the actual label and the labeling qualities of individual oracles. But they do not use the labeling qualities of individual oracles as prior knowledge to actively select the oracles in the following process.

Compared with the existed works, the proposed probabilistic model in this paper goes beyond the majority vote methods by relaxing the assumption of identical level of expertise among the oracles. While considering the differently labeling qualities of individual oracles, a distinguishing factor from the previous works is that the oracle selection is based on prior knowledge of labeling qualities of individual oracles which is learned jointly with the actual label. By directly modeling the instances and noisy labels, the reliabilities of individual oracles and the actual label are both derived by maximizing likelihood of observed data. Moreover, in our model, the reliabilities of individual oracles are evaluated on different categories which are obtained by a Bayesian estimation based on previously observed labels.

3. PMActive

In this section, we show the PMActive method in detail. Before describing our model, some basic notions and necessary preliminaries are given firstly. Let $\mathbf{X}=\{x_1,x_2,\ldots,x_n\}$ be the input points and $\mathbf{Y}=\{y_1,y_2,\ldots,y_n\}$ be the actual labels which cannot be observed. Let $\mathbf{Z}=\{z_1^1,z_1^2,\ldots,z_j^i,\ldots,z_m^n\}$ be the observed labels which are provided by multiple oracles. The z_j^i is the observed label which is given by the j_{th} oracle for the instance x_i. Suppose that the random variable $X\in\mathbf{X}$ and $Z\in\mathbf{Z}$, respectively, denote the observed instance and the noisy labels, then the random variable $Y\in\mathbf{Y}$ denotes the hidden actual label, and then we have

$$p(Z|X)=\sum_{Y} p(Z|X,Y)\cdot p(Y|X) \tag{1}$$

For simplicity, we make an assumption that $p(Z|X,Y)=p(Z|Y)$. In practice, it is a reasonable assumption because the reliabilities of individual oracles should be consistent across different sub-groups of