Research on Query-by-Committee Method of Active Learning and Application

Yue Zhao, Ciwen Xu, and Yongcun Cao

School of Mathematics and Computer Science,
Central University for Nationalities, 100081 Beijing, China
zhaoyueso@sina.com

What is the given cost function? How does one select it?

Abstract. Active learning aims at reducing the number of training examples to be labeled by automatically processing the unlabeled examples, then selecting the most informative ones with respect to a given cost function for a human to label. The major problem is to find the best selection strategy function to quickly reach high classification accuracy. Query-by-Committee (QBC) method of active learning is less computation than other active learning approaches, but its classification accuracy can not achieve the same high as passive learning. In this paper, a new selection strategy for the QBC method is presented by combining Vote Entropy with Kullback-Leibler divergence. Experimental results show that the proposed algorithm is better than previous QBC approach in classification accuracy. It can reach the same accuracy as passive learning with few labeled training examples.

1 Introduction

Obtaining labeled training examples for some classification tasks is often expensive, such as text classification, mail filtering, credit classification and et al., while gathering large quantities of unlabeled examples is usually very cheap. For example, the available cases with actual classes are not enough for building credit classification model in practice, especially for the newly established system in which old customers' data do not exist. In this situation, one solution is the manual classification. The credit customers are evaluated by experts and classified into different risk levels. It is time-consuming and costly. Thus, active learning can reduce annotation cost by sample selection. QBC is a kind of sample selection method of active learning. It is less computation than the one based on Error Reduction Sampling, but existing QBC approaches do not reach the same classification accuracy as passive learning [1-3]. In this paper, the present selection strategy for QBC attempts to provide a solution for this problem.

We consider two selection functions for measuring the disagreement among committee members in QBC. One uses Kullback-Leibler divergence (KL-d) to the mean for capturing the information [4]. One disadvantage of KL divergence is that it misses some examples on which committee members disagree, but these examples are exactly needed by QBC. The other selection function measures the disagreement by Vote Entropy (VE) [5]. The disadvantage of Vote Entropy is that it does not consider the committee members' class distributions, $P_m(C \mid e_i)$. Each committee member

X. Li, O.R. Zaiane, and Z. Li (Eds.): ADMA 2006, LNAI 4093, pp. 985 – 991, 2006.

m produces a posterior class distribution, $P_m(C \mid e_i)$, where C is a random variable over classes and e_i is an input unlabeled example. So, Vote Entropy also misses some informative unlabeled examples to label. Because both of them above do not select enough useful examples, they can not achieve the same classification accuracy as passive learning.

We propose a new select strategy for QBC by combining the Vote Entropy with Kullback-Leibler divergence to improve the classification accuracy with much fewer training data than passive learning.

2 The Query-by-Committee Method of Active Learning

The Query-by-Committee method of active learning examines unlabeled examples and selects only those that are most informative for labeling. This avoids redundant labeling examples that contribute little new information. Our research follows theoretical work on sample selection in the Query-by-Committee paradigm [5]. In this committee-based selection scheme, the learning receives a stream of unlabeled examples as input and decides for each of them whether to ask for its label or not. To that end, the learner constructs a 'committee' of (two or more) classifiers based on the statistics of the current training set. Each committee member then classifies the candidate example and the learner measures the degree of disagreement among the committee members. The example is selected for labeling depending on this degree of disagreement according to some selection protocol. Its algorithm is available in [5].

There are two selection functions for QBC to measure the disagreement. One uses Kullback-Leibler divergence to the mean for capturing the information of disagreement. KL-d measures the strength of the certainty of disagreement by calculating differences in the committee members' class distributions. KL-d to the mean is an average of the KL divergence between each distribution and the mean of all distribution:

$$\frac{1}{K}\sum_{m=1}^{K} D(P_m(C \mid e_i) \| P_{avg}(C \mid e_i)), \tag{1}$$

where $P_{avg}(C \mid e_i)$ is the class distribution mean over all committee members, m:

$$P_{avg}(C \mid e_i) = (\sum_m P_m(C \mid e_i))/K. \tag{2}$$

KL divergence, $D(\bullet \| \bullet)$, is an information-theoretic measure of the difference between tow distributions. It is:

$$D(P_1(C) \| P_2(C)) = \sum_{j=1}^{|C|} P_1(c_j) \log(\frac{P_1(c_j)}{P_2(c_j)}). \tag{3}$$

The other selection function measures disagreement by the entropy of the distribution of classification 'voted for' by the committee members. This Vote Entropy is

natural measure for quantifying the uniformity of classes assigned to an example by the different committee member. It is:

$$-\frac{1}{\log \min(K, |C|)} \sum_{C} \frac{V(c, e_i)}{K} \log \frac{V(c, e_i)}{K}, \tag{4}$$

where $V(c, e_i)$ denotes the number of committee members assigning a class c for e_i and K is the number of committee members.

One disadvantage of KL-d is that it misses some examples on which committee members disagree, but these examples are exactly needed by QBC. An illustrative experiment of learning the binary classification task is presented in Table 1 and 2. Supposed the 2-member (tow classification model) committee is generated and 3 examples need to be decided whether to ask for its label or not. If we compare the Vote Entropy (VE) of e_1 with e_2 in Table 2, we see that they are both selected for labeling (when the degree of disagreement is most, VE is 1 for binary classification. If committee is unanimous for an example, VE is zero.). But their KL-divergence is quite different and only e_2 is selected. It shows that KL-d misses the examples which are very informative.

Table 1. The results of committee members' class vote for unlabeled examples

Model	e_1	e_2	e_3
1	0.52(c_1)	0.72(c_2)	0.60(c_2)
2	0.58(c_2)	0.60(c_1)	0.70(c_2)

Table 2. The Kullback-Leibler divergence and Vote Entropy of examples

Example	VE	KL-d
e_1	1	0.005(miss)
e_2	1	0.052
e_3	0	0.006

However, one disadvantage of Vote Entropy is that it does not consider the committee members' classifications distributions, $P_m(C \mid e_i)$, according the formula (4). It can also miss informative examples.

Because both of them above do not select enough useful examples, they can not achieve the same classification accuracy as passive learning.

3 A New Algorithm for QBC

Through discussing the disadvantages of exist selection functions of QBC in section 2, we consider to combine the Vote Entropy with Kullback-Leibler divergence by selecting some examples which committee agree on and have high uncertainty for a member of committee. The degree of disagreement among the committee members is

measured by Vote Entropy and each example's uncertainty of classification is measured by KL-d. So we redefine KL-d. Let it measure the difference between $P_m(C \mid e_i)$ and $P_{m\_avg}(C \mid e_i)$ which are class probability distributions in the most uncertainty. The minimum of KL-d among committee members for each unlabeled example is taken as an example's the most uncertainty from committee members, denoted by KL-d_{min}. If KL-d_{min} of an example satisfies the term of some threshold α (α is near to zero.), this example is selected for labeling and being added into training data set. These selected examples are informative and can contribute to improving classification accuracy [6]. KL-d_{min} is:

$$KL-d_{\min}(e_i) = \underset{m=1}{\overset{K}{MIN}}\left(\sum_{j=1}^{|c|} P_m(c_j \mid e_i)\log\left(\frac{P_m(c_j \mid e_i)}{P_{m\_avg}(C \mid e_i)}\right)\right), \tag{5}$$

where

$$P_{m\_avg}(C \mid e_i) = \left(\sum_{j=1}^{|c|} P_m(c_j \mid e_i)\right)/l, \tag{6}$$

$$l = \begin{cases} 2, \max(P_m(c_j \mid e_i)) = 1 \\ (|c| - num\_zero), \max(P_m(c_j \mid e_i)) \neq 1 \end{cases}, \tag{7}$$

num_zero is the number of the class probabilities which are zeros.

The other experiment is presented in Table 3 and 4. It shows that the example of e_1 is missed by VE, but it has the same high classification uncertainty as e_2 and e_3 by KL-d_{min}. It is helpful for building classification model by QBC.

Table 3. The results of committee members' class vote for unlabeled examples

Model	e_1	e_2	e_3	e_4
1	0.55(c_1)	0.55(c_2)	0.52(c_1)	0.80(c_2)
2	0.55(c_1)	0.55(c_2)	0.75(c_2)	0.90(c_2)
3	0.60(c_1)	0.55(c_1)	0.85(c_2)	0.75(c_2)
4	0.60(c_1)	0.55(c_1)	0.95(c_2)	0.85(c_2)

We combine the Vote Entropy with KL-d_{min} to propose our algorithm for QBC. When an example is agreed by committee members, it is measured by KL-d_{min} further. If its KL-d_{min} satisfies some threshold α, it is selected for labeling and added into training data set. The new algorithm is described as follows.

Table 4. The KL-d_{min} and Vote Entropy of examples

Example	VE	KL-d
e_1	0.0(miss)	0.005
e_2	1.0	0.005
e_3	0.81	0
e_4	0.0	0.1308

A new algorithm for QBC

Input: Classification algorithm: A
The number of committee members: K
Few labeled examples: L
Unlabeled examples: UL
The condition of stopping: ζ
The threshold of Vote Entropy: θ
The threshold of KL-d_{min} : α
1. Learn K classifiers $\{M_h\}$ from L using A;
2. While not ζ
{ 1) $\forall e_i \in UL$, for h=1,…, K:
Classify e_i using M_h to get class label C_h;
2) using formula (4) to Compute $VE(e_i)$;
3) If $VE(e_i) > \theta$,
Select e_i from UL, get true label and add e_i to L, learn K classifiers from L using A again;
Else
Compute $KL - d_{\min}(e_i)$;
If $KL - d_{\min}(e_i)$ satisfies α, select e_i from UL, get true label and add e_i to L, learn K classifiers from L using A again;
End
4) Check the condition of stopping ζ;}
3. Learn classifier M from L using A;
Output: Classifier M.

4 Experimental Results

We now discuss the results of our experiments on Nursery database and Tic-Tac-Toe Endgame database from the UCI Machine Learning Repository.

We randomly sample 4171 instances from Nursery database. The data set is randomly partitioned into labeled examples set, unlabeled examples set and independent test set. Each instance contains 9 attributes.

Tic-Tac-Toe Endgame database consist of 958 instances. Each instance contains 10 attributes. The data set is also randomly partitioned into labeled examples set, unlabeled examples set and independent test set. The class attribute has 2 values. About 65.3% are positive of class distribution.

We choose the TAN classifier [7] as classification algorithm. For QBC, we use a committee size of two. The condition of stopping ζ is that classification accuracy reaches the expected value or unlabeled data set is empty.

Fig.1 plots the learning curves obtained from the 3 learning methods—VE of QBC、VE&KL-d_{min} of QBC and passive learning—on Nursery database. It is clearly

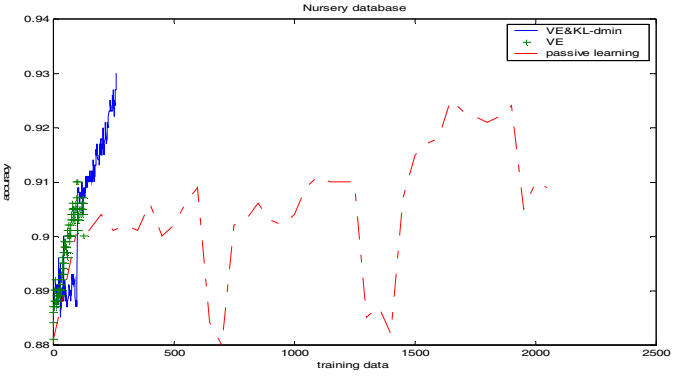

Fig. 1. The classification accuracy comparison of VE、VE&KL-d_{min} and passive learning on Nursery database

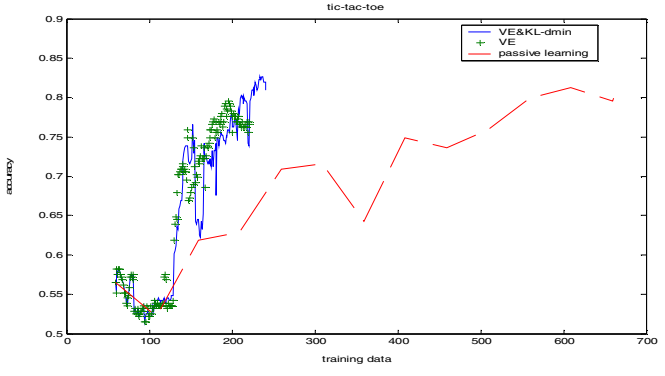

Fig. 2. The classification accuracy comparison of VE、VE&KL-d_{min} and passive learning on Tic-Tac-Toe Endgame database

seen from these graphs that the VE of QBC achieves the accuracy of 91% after selecting 128 unlabeled examples, VE&KL-d_{min} achieves 93% after 264 unlabeled examples and passive learning achieves 92.51% after all unlabeled data.

Fig.2 plots the learning curves on Tic-Tac-Toe Endgame database. It is clearly seen from these graphs that the VE achieves the accuracy of 79% after selecting 138 unlabeled examples, VE&KL-d_{min} achieves 83% after 176 unlabeled examples and passive learning achieves 81% after all unlabeled data.

A Chinese credit rating data set for telecom clients was collected from Jan. to May, 2001. It consists of 33512 instances. Each instance contains 15 attributes and one credit class attribute which has 4 values of credit risk levels. Fig.3 plots the learning curves on the credit scoring data set. It is clearly seen from these graphs that the VE achieves the accuracy of 81% after selecting 278 unlabeled examples, VE&KL-d_{min} achieves 84% after 921 unlabeled examples and passive learning achieves 84% after all unlabeled data.

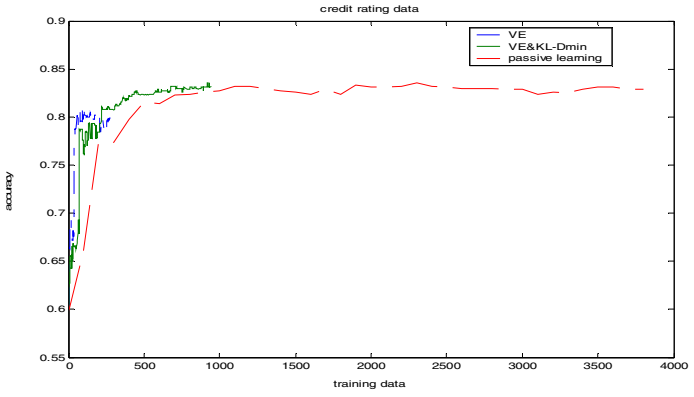

Fig. 3. The classification accuracy comparison of VE、VE&KL-d_{min} and passive learning on Chinese credit scoring data

The results show that the VE&KL-d_{min} is better than previous VE of QBC. It can reach the same accuracy as passive learning with few labeled examples. The selected newly examples contribute to improve classifier's accuracy.

5 Summary

In this paper, we present a new selection strategy for QBC to improve the accuracy using Vote Entropy and Kullback-Leibler divergence. The experimental results indicate that the proposed algorithm is better than previous QBC in classification accuracy with much fewer labeled examples than passive learning.

References

1. Freund, Y., Seung, H.S., Samir, E., Tishby, N.: Selective Sampling Using the Query by Committee Algorithm. Machine Learning, 28(1997)133-168
2. Gong, X.J., Shun, J.P., Shi, Z.Z.: An Active Bayesian Network Classifier. Computer research and development, 39 (2002)574-579
3. Riccardi, G., Hakkani-Tür, D.: Active Learning: Theory and Applications to Automatic Speech Recognition. IEEE Transaction on Speech and Audio Processing, 13 (2005)504-511
4. McCallum, A. K., Nigam, K.: Employing EM and Pool-based Active Learning for Text Classification. In: Proceeding of the 15th International Conference on Machine Learning, Morgan Kaufmann, San Francisco Madison (1998) 350—358
5. Argamon-Engleson, S., Dagan, I.: Committee-based Sample Selection for Probabilistic Classifers. Journal of Artificial Intelligence Research, 11 (1999)335-460
6. Lewis, D.D., Gale, W.A.: A Sequential Algorithm for Training Text Classifiers. In: Proceedings of {SIGIR}-94, 17th {ACM} International Conference on Research and Development in Information Retrieva, Springer-Verlag, Berlin Heidelberg Dublin (1994)3-12
7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning, 29(1997)131-161