Active Learning Based on diversity maximization

Yongcheng Wu
Computing school
Jingchu University of Technology
Jingmen, Hubei, 448000, China
wuyongcheng11@126.com

***Abstract*—In many practical data mining applications, unlabeled training examples are readily available but labeled ones are fairly expensive to obtain. Therefore, as one type of the paradigms for addressing the problem of combining labeled and unlabeled data to boost the performance, active learning has attracted much attention. In this paper, we propose a new active learning approach based on diversity maximization. Different from the well-known co-testing algorithm, our method does not require two different views. The comparative studies with other active learning methods demonstrate the effectiveness of the proposed approach.**

Keywords-machine learning; active learning; classification ; diversity

I. INTRODUCTION

Supervised learning algorithms require a large amount of labeled data in order to achieve high accuracy and this accuracy declines as the amount of available labeled data decreases. However, labeling data is often difficult, expensive, or time consuming, as it requires the efforts of experienced human annotators. In many practical data mining applications such as content-based image retrieval, computer-aided medical diagnosis [1], object detection and tracking, web page categorization [2], or e-mail classification [3], there is often an extremely large pool of data available.

In the machine learning literature, there are mainly three paradigms for addressing the problem of combining labeled and unlabeled data to boost the performance: semi-supervised learning, transductive learning and active learning. Semi-supervised learning (SSL) refers to methods that attempt to either exploit the unlabeled data for supervised learning where the unlabeled examples are different from the test examples or to exploit the labeled data for unsupervised learning. Transductive learning refers to methods which also attempt to exploit unlabeled examples but assuming that the unlabeled examples are exactly the test examples. Active learning [4] refers to methods which selects the most important unlabeled examples, and an oracle can be asked for labeling these instances, where the aim is to minimize data labeling. Sometimes it is called selective sampling or sample selection.

In this paper, we propose a new active learning approach which is a variation of co-testing [5, 6]. Different from co-testing, our method uses two different base learner rather than two different views. The proposed approach is based on the diversity maximization. The interesting feature of the proposed approach is that it measures the informativeness of an instance by its prediction confidence: the informativeness of an instance x is measured by its prediction confidence based on the labeled data. Experimental results show that by using diversity maximization, our method is more effective to apply in various data sets than its counterpart.

The remaining of this paper is organized as follows. Section Ⅱ reviews previous works on active learning. Section Ⅲ presents the proposed approach in details; experimental results are reported in Section Ⅳ. Finally, Section Ⅴ concludes the paper and discusses directions for future work.

II. RELATED WORK

A. Active learning query strategy

All active learning involve evaluating the informativeness of unlabeled instances, which can either be generated de novo or sampled from a given distribution. There have been many proposed ways of formulating such query strategies in the literature.

Uncertainty Sampling. Perhaps the simplest and most commonly used query framework is uncertainty sampling [7] (Lewis and Gale, 1994). In this framework, an active learner queries the instances about which it is least certain how to label. This approach is often straightforward for probabilistic learning models. For example, when using a probabilistic model for binary classification, uncertainty sampling simply queries the instance whose posterior probability of being positive is nearest 0.5 [7, 8] (Lewis and Gale, 1994; Lewis and Catlett, 1994).

Query-By-Committee. Another, more theoretically-motivated query selection framework is the query-by-committee (QBC) algorithm [9]. The QBC approach involves maintaining a committee C = ($h^{(1)}\dots h^{(C)}$)of models which are all trained on the current labeled set L, but represent competing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree.

Expected Model Change. Another general active learning framework uses a decision-theoretic approach, selecting the instance that would impart the greatest change to the current model if we knew its label. An example query strategy in this framework is the “expected gradient length”

Published by Atlantis Press, Paris, France.

(EGL) approach for discriminative probabilistic model classes. This strategy was introduced by Settles et al. [10] (2008b) for active learning in the multiple-instance setting , and has also been applied to probabilistic sequence models like CRFs [11] .

B. Co-testing algorithm

The first algorithm for active learning in multi-view setting is co-testing [5, 6]. It focuses on the set of contention points (i.e., unlabeled examples on which different views predict different labels) and asks the user to label some of them. This is somewhat related to Query-by-Committee since co-testing also uses more than one learners to identify the most informative unlabeled examples to query, but the typical Query-by-Committee works under a single-view setting while co-testing exploits the multi-views explicitly. It was reported that co-testing outperforms existing active learners on a variety of real-world domains such as wrapper induction, Web page classification, advertisement removal and discourse tree parsing.

Co-testing algorithm is shown in Figure 1.

```
Given:
    a set U of unlabeled examples(with two views)
    a set L of labeled training examples(with two views)
Process:
    Loop for K iterations:
    Use L to train a classifier $h^1$ that consider only the $x_1$ portion of x;
    Use L to train a classifier $h^2$ that consider only the $x_2$ portion fo x;
    Apply $h_i^1$ and $h_i^2$ to the unlabeled data set U and ¯find out the contention points set $Q_i$;
    Ask the user to label $m_{i+1}$ unlabeled examples drawn randomly from $Q_i$, then add them into L and delete them from U. Add these self-labeled examples to L;
Output:
    $h_{final}$ = combine($h_{s^1}$; $h_{s^2}$)
```

Figure 1. co-testing algorithm

The standard co-testing algorithms requires two views, that is, the attributes be naturally partitioned into two sets. However, in many domains there are not such natural feature splits.

III. OurActive learning Method

Instead of using two views, our method uses two different base learning algorithms. Using the same set L and two different base learning algorithms to train two classifiers h^1 and h^2. Then using h^1 and h^2 to predict the class of the unlabeled examples. The example selected to query is the most informative example which is: with very high estimation confidence, the example is predicted by h^1 and h^2 with different class.

For example, given two pairs of classifiers(A,B)and (C,D), if we know that all of them are with 100% accuracy on labeled training data, then there will be no difference for taking either the ensemble consists of (A,B)or the ensemble consists of (C,D); however, if we find that A and B make the same predictions on unlabeled data, while C and D make different predictions on some unlabeled data, then we will know that the ensemble consists of (C,D) would have good chance to be better.

Based on these ideas, our algorithm is shown in Figure 2.

```
Given:
    a set U of unlabeled examples(with single views)
    a set L of labeled training examples(with single views)
Process:
    Loop for K iterations:
    Use L to train a classifier $h^1$ on base learning algorithm 1;
    Use L to train a classifier $h^2$ on base learning algorithm 2;
    Apply $h_i^1$ and $h_i^2$to the unlabeled data set U and find out the contention points set $Q_i$ (with high estimation confidence, the example is predicted by $h^1$ and $h^2$ with different class);
    Ask the user to label $m_{i+1}$ unlabeled examples drawn randomly from $Q_i$, then add them into L and
delete them from U. Add these self-labeled examples to L
Output:
    $h_{final}$ = combine($h_{s^1}$; $h_{s^2}$)
```

Figure 2. our active learning algorithm

Two views are not required in our method, which is convenient to apply in various data sets. Meanwhile, our method requires two base learning algorithms which can undertake the class probability estimation.

Consider a classification issue with M classes C_i, where i=1, 2….M. For each test date x, the class probability estimation for each category (class) is defined as $P(c_i|x)$. And the data x is assigned the most probable target value label_c based on $\text{label}_c=\text{argmax}_{c_i \in x}P(c_i|x)$. For different classifiers, the calculation of $P(c_i|x)$ is based on diverse mechanisms.

The diversity from the different classifiers can help to choose the appropriate data for recovering labels, which shares similar property with disagreement-based co-testing.

IV. Experiments

Five UCI data sets [12] are used in the experiments as shown in Table 1. Each data set is randomly divided into two parts of equal size, with one part as the test data and the other part as the unlabeled data that is used for active learning. We assume that only a few labeled data is available at the very beginning of active learning.

We compare our algorithm with the following two baseline approaches: (1) RANDOM: randomly select query instances, (2) MARGIN: margin-based active learning [13], a representative approach which selects informative instances.

For MARGIN, instances are randomly selected when no classification model is available, which only takes place at the beginning. In each iteration, an unlabeled instance is first selected to solicit its class label and the classification model is then retrained using additional labeled instance. We evaluate the classification model by its performance on the holdout test data.

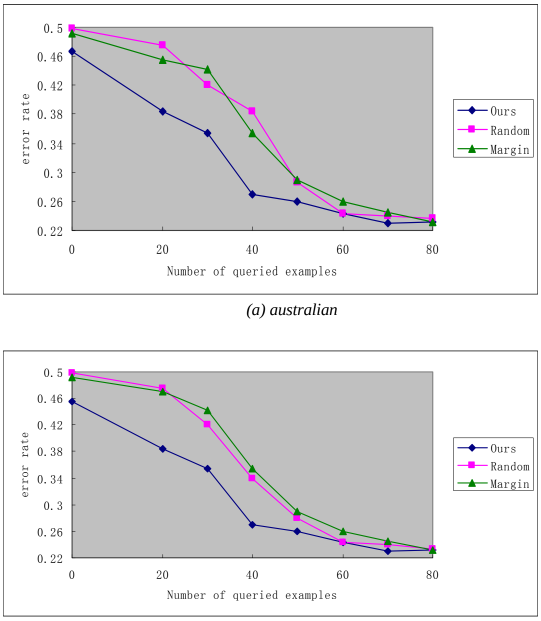
Classification accuracy is used for evaluation metrics. For every data set, we run the experiment for ten times, each with a random partition of the data set.

The base classifier used in our experiment is decision tree (DT) and Naive Bayes (NB), which are from the WEKA library [14].

TABLE I. EXPERIMENTAL DATA SETS

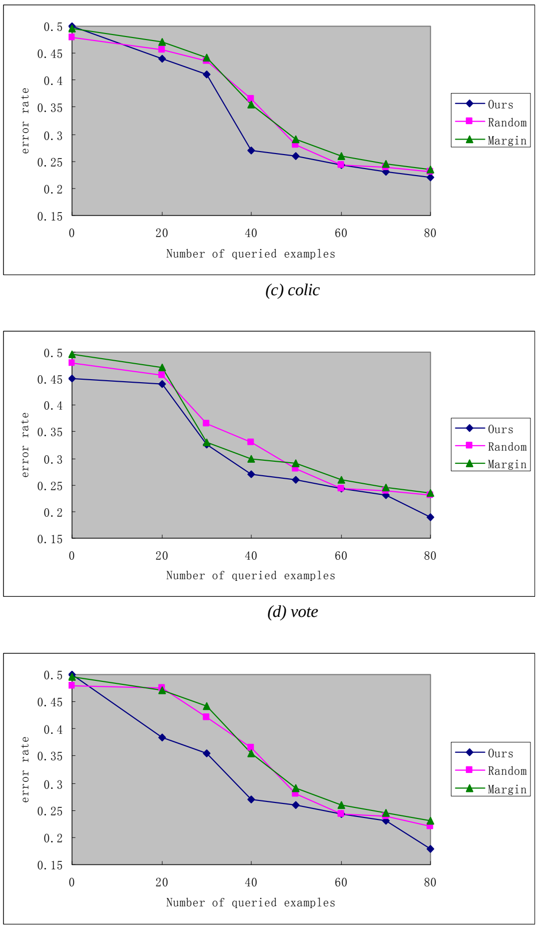
Data set	Attribute	Size	Class	Pos/Neg
australian	14	690	2	55.5%/44.5%
bupa	6	345	2	42.0%/58.0%
colic	22	368	2	63.0%/37.0%
vote	16	435	2	61.4%/38.6%
wdbc	30	569	2	37.3%/62.7%

Figure 3 shows the classification error rate of different active learning approaches with varied numbers of queries.(*Ours* refers to our active learning algorithm).



(a) australian

(b) bupa



(c) colic

(d) vote

(e) wdbc

Figure 3. Comparison on classification error rate

We observe that the RANDOM approach tends to yield decent performance when the number of queries is very small. However, as the number of queries increases, this simple approach loses its edge and often is not as effective as the other active learning approaches. MARGIN, the most commonly used approach for active learning, is not performing well at the beginning of the learning stage. As the number of queries increases, we observe that MARGIN catches up with the other approaches and yields decent performance. This phenomenon can be attributed to the fact that with only a few training examples, the learned decision boundary tends to be inaccurate, and as a result, the unlabeled instances closest to the decision boundary may not be the most informative ones.

We observe that for most cases, our method is able to outperform the baseline methods significantly, as indicated by Figure 3. We attribute the success of our method to the principle of choosing unlabeled instances that are informative based on the diversity maximization.

V. CONCLUSIONS

In this paper, we propose a new active learning approach based on diversity maximization. Different from co-testing, our method does not require two different views, which is more convenient and practical. The comparative studies with other active learning methods demonstrate the effectiveness of the proposed approach.

Strategies for dealing with highly uncertain answers from the oracle, and for preventing dramatic changes of data distribution when new examples are included in the training set are also interesting research issues to further improve the performance

Our current work is restricted to binary classification. In the future, we plan to extend this work to multi-class learning It is interesting to see whether our method works well with other base learners. It would be insightful to analyze why our method can achieve good performance theoretically.

REFERENCES

[1] Li M, Zhou ZH. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Trans. on Systems, Man and Cybernetics- Part A: Systems and Humans, 2007, 37(6): 1088-1098.

[2] Nigam K, McCallum AK, Thrun S, Mitchell T. Text classi_cation from labeled and unlabeled documents using em. Machine Learning, 2000, 39(2-3): 103-134.

[3] Kiritchenko S, Matwin S. Email classi_cation with co-training. In: Proc. of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research (CASCON'01). IBM Press, 2001.8-19.

[4] Z.-H. Zhou. Learning with unlabeled data and its application to image retrieval. In: Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06), Guilin, China, LNAI 4099, 2006, pp.5-10.

[5] Muslea, I., Minton, S., & Knoblock, C. A. Selective sampling with redundant views. Proceedings of the 17th National Conference on Artificial Intelligence . Austin, TX. 2000, 621-626

[6] Muslea, I., Minton, S., & Knoblock, C. A. Active learning with multiple views. Journal of Artificial Intelligence Research, 2006, 27, 203-233.

[7] Lewis and W. Gale. A sequential algorithm for training text classifiers. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12. ACM/Springer, 1994.

[8] Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In Proceedings of the International Conference on Machine Learning (ICML), pages 148–156. Morgan Kaufmann, 1994.

[9] Seung, M. Opper, and H. Sompolinsky. Query by committee. In Proceedings of the ACM Workshop on Computational Learning Theory, pages 287–294,1992.

[10] Settles, M. Craven, and S. Ray. Multiple-instance active learning. In Advances in Neural Information Processing Systems (NIPS), volume 20, pages 1289–1296. MIT Press, 2008b.

[11] Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1069–1078. ACL Press, 2008.

[12] C. Blake, E. Keogh, and C.J. Merz, “UCI repository of machine learningdatabases”[http://www.ics.uci.edu/»mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA, 1998.

[13] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In Proceedings of the 17th International Conference on Machine Learning, pages 999–1006, 2000.

[14] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, October 1999.

Published by Atlantis Press, Paris, France.