DOCUMENT FILTERING BOOSTED BY UNLABELED DATA

Seong-Bae Park and *Byoung-Tak Zhang*

Artificial Intelligence Lab (SCAI)
School of Computer Science and Engineering
Seoul National University, Korea

ABSTRACT

This paper describes three learning methods for document filtering that use unlabeled data. The proposed methods are based on a committee of the classifiers which are trained on a small set of labeled data and then augmented by a large number of unlabeled data. By taking advantage of unlabeled data, the effective number of labeled data needed is significantly reduced and the filtering accuracy is increased. The use of unlabeled data is important because obtaining labeled data is difficult and time-consuming, while unlabeled data are abundant. For all proposed methods, the experimental results show that the accuracy is improved up to 9.2% with only two-thirds as many labeled data as the method which does not use unlabeled data.

1. INTRODUCTION

Due to the massive volume of online text documents available, it is of great importance to classify or filter the documents. For the most machine learning algorithms applied to this task, plenty of labeled documents must be supplied [6,12]. However, it is very expensive and time-consuming to come by the labeled documents because labeling must be done by human experts. On the other hand, the unlabeled documents are significantly easier to obtain than the labeled ones, especially in the Web environment.

Some of previous research show that it is useful to use the unlabeled data in text processing [1,8,9,13]. In the word sense disambiguation field, Yarowsky used a learning algorithm based on the local context under the assumption that all instances of a word have the same intended meaning within any fixed document and achieved good results with only a few labeled examples and many unlabeled ones [13]. Park et al. also showed that a model learning a small set of labeled data can achieve high accuracy by utilizing additional unlabeled data whose labels are estimated by the current model [9]. Blum and Mitchell tried to classify Web pages, in which the description of each example can be partitioned into distinct views such as the words occurring on that page and the words occurring in hyperlinks [1]. By using both views together, they augmented a small set of labeled examples with a lot of unlabeled examples.

Given unlabeled data set $D = \{x_1, \ldots, x_T\}$,
and labeled data set L,
Resample L_j from L for each classifier C_j,
where $|L_j| = |L|$ as done in Bagging.
Train base classifier C_j $(1 \le j \le M)$ with L_j.
Do

- E-Step:
  Determine the label, y_t of each $x_t \in D$ by majority voting of C_j.
- M-Step:
  Re-learn C_j with $L_j + \{(x_1, y_1), \ldots, (x_T, y_T)\}$.

Until there is no change in labels of D

Output the final classifier:

$$y(x) = \arg\max_{y \in \{-1,+1\}} \sum_{j=1}^{M} I(C_j(x) = y).$$

Figure 1: The algorithm of committee-based EM-like method.

The usage of unlabeled examples in text classification provides information about the joint probability distribution over words though they may mislead the classifier. As combining or integrating the outputs of several classifiers leads to improved performance, the possibility misled by unlabeled examples can be reduced by the committee of classifiers.

This paper describes methods which learn how to filter documents using both labeled and unlabeled data. Basically, the committee of classifiers is learned with labeled data at first, and is augmented by unlabeled data. We show experimentally that the proposed methods effectively perform document filtering with only small number of labeled examples and adapts their own ability through a series of unlabeled data.

0-7803-7090-2/01/$10.00 © 2001 IEEE.

Given unlabeled data set $D = \{x_1, \ldots, x_T\}$,
and labeled data set L,
Initialize $W_1(j) = 1/M$, where M is the number of classifiers in the committee.
Resample $L_j^{(1)}$ from L for each classifier C_j,
where $|L_j^{(1)}| = |L|$ as done in Bagging.
Train base classifier C_j ($1 \le j \le M$) with $L_j^{(1)}$.
For $t = 1, \ldots, T$:

1. Each C_j determines the output $y_j \in \{-1, +1\}$ for $x_t \in D$.
$$Y = \langle y_1, \ldots, y_M \rangle$$
2. Find the most likely output y_t from Y using distribution W:
$$y_t = \arg\max_{y \in \{-1,+1\}} \sum_{j: C_j(x_t)=y} W_T(j).$$
3. Set $\alpha_t = (1 - \varepsilon_t) / \varepsilon_t$, where
$$\varepsilon_t = \frac{\text{No. of classifiers whose output is not } y_t}{M}.$$
4. **If** α_t is larger than a *certainty* threshold θ, **then** update W_t:
$$W_{t+1}(j) = \frac{W_t(j)}{Z_t} \times \begin{cases} \alpha_t & \text{if } y_j = y_t \\ 1 & \text{otherwise,} \end{cases}$$
where Z_t is a normalization constant.
5. **Otherwise**, every classifier C_j is restructured from new training set $L_j^{(t+1)}$:
$$L_j^{(t+1)} = L_j^{(t)} + \{(x_t, y_t)\}.$$

Output the final classifier:
$$y(x) = \arg\max_{y \in \{-1,+1\}} \sum_{j: C_j(x)=y} W_T(j).$$

Figure 2: The algorithm of committee-based active sampling method.

The rest of this paper is organized as follows. Section 2 explains three methods to filter documents using both labeled and unlabeled data. Section 3 presents the data set of WebKB and the experimental settings. Section 4 shows the experimental results. Section 5 draws conclusions.

2. COMMITTEE-BASED METHODS USING UNLABELED DATA

In general, the committee of classifiers leads to improved generalization. In consequence, the possibility to misestimate the labels of unlabeled data can be reduced by using a committee instead of a single classifier. All methods described below are based on a committee of classifiers.

2.1 EM-LIKE METHOD

The basic method to handle unlabeled data is to use the EM algorithm [3], because the unlabeled data can be considered to be incomplete. Nigam et al. showed that the accuracy of text classification could be increased with the EM algorithm and a naïve Bayes classifier [8].

The method using the EM algorithm is presented in Figure 1. In the *Expectation* step, the label of each unlabeled data is determined by majority voting of current classifiers. And then, in the *Maximization* step, the classifiers are retrained using both labeled and unlabeled data where the labels of unlabeled data are given in the E-step. Because the committee model proposed is not formulated with likelihood, it is not guaranteed that the number of errors is reduced monotonically in determining labels of unlabeled data through iteration. Thus, we heuristically executed the iteration until there is no change in labels of unlabeled data.

2.2 ACTIVE SAMPLING METHOD

The number of examples needed can also be reduced by doing as the active learning algorithms do. Liere and Tadepalli experimentally showed that active learning with committees achieves accuracy as good as a passive single learner, but uses only 2.9% as many training examples as the single learner [7].

Park et al. showed that a committee machine which estimates the label of unlabeled data could increase the accuracy in word sense disambiguation [9]. In this method, the committee is first trained with given small set of labeled data, and then enhanced by a stream of unlabeled data, as shown in Figure 2. The committee consists of simple base classifiers which are trained with examples drawn by bootstrap replicates [2]. While training unlabeled data, every unlabeled example is given to all committee members and each member determines its label. When the members agree on the label, they need not learn it because they already have a talent for correct labeling of it. In this case, instead, the importance weight of each member is adjusted according to its output and the agreed label. When the members do not agree, they learn the unlabeled example assuming that the label of the example is the one determined by majority voting of committee members. After the training, the label of unseen data is determined by a kind of weighted majority voting.

The main drawback of this method is that it requires a number of unlabeled data when the sufficient labeled data are not given.

2.3 ADABOOST-LIKE METHOD

Figure 3 is a variant of **AdaBoost.M1** proposed by Freund and Schapire [4]. The only difference of the method from AdaBoost.M1 is that the ε_t is applied not to an individual classifier but to a committee. In AdaBoost the training data are reused, so that the limitation of active sampling mentioned above which requires a large training data can be overcome.

Given unlabeled data set $D = \{x_1, \ldots, x_T\}$,
and labeled data set L,
Resample L_j from L for each classifier C_j,
where $|L_j| = |L|$ as done in Bagging.
Train base classifier C_j $(1 \le j \le M)$ with L_j.
Initialize $D_1(i) = 1/T$ for all i.
For $t = 1, \ldots, N$:

1. Resample $U^{(t)}$ from U with the distribution D, where the label of an unlabeled example is determined by majority voting
2. Compose the committee CM_t of C_j trained with $L_j + U^{(t)}$.
3. Calculate the error of CM_t:

$$\varepsilon_t = \frac{\sum_{i=1}^{|U^{(t)}|} \varepsilon_t(i)}{|U^{(t)}|}$$

and

$$\varepsilon_t(i) = \frac{M_t(i)}{M}$$

where Mt(i) is the number of C_j whose output is not $CM_t(x_i)$.

4. Update distribution D_t:

$$D_{t+1}(i) = \frac{D_t(i) \times \varepsilon_t}{Z_t},$$

where Z_t is a normalization constant.

Output the final classifier:

$$y(x) = \arg\max_{y \in \{-1,+1\}} \sum_{j: C_j(x) = y} W_T(j).$$

Figure 3: The algorithm of committee-based AdaBoost-like method.

The label of unlabeled data is estimated by majority voting of base classifiers as done in the EM-like method. In this method, the distribution D over the training data is maintained only on unlabeled data. The base classifiers of the committee at very iteration are trained with both labeled data and label-estimated unlabeled data resampled with distribution D. Since ε_t is determined by the number of classifiers which do not agree on the label of an unlabeled example, the examples that are easily agreed on their label by the committee get lower value, and the examples whose label is difficult to be agreed on get a higher weight value. Thus, this method focuses on the examples which the committee members do not agree on their label.

The final classifier is a weighted vote of the committees. The weight of the committee CM_t is defined to be $\log(1 / \varepsilon_t)$, so that a great weight is given to the committee which agrees on as many examples as possible.

3. TEXT FILTERING

3.1 DATA SET

The data set[1] for the experiments is the one used in [1], which is a subset of "The 4 Universities Data Set" from "World Wide Knowledge Base Project" of CMU text learning group. It consists of 1,051 web pages collected from computer science departments of four universities: Cornell, University of Washington, University of Wisconsin, and University of Texas. The 1,051 pages are manually classified into *course* or *non-course* category. The categories are shown in Table 1 with the number of web pages in each university. The baseline performance implies the accuracy achieved by answering non-course for all examples.

Since the web pages of each university have their own idiosyncrasies, it is recommended by the group that training and testing on pages from the same university should not be conducted. Instead, it recommends training on three of the universities and testing on the pages from a fourth, held-out university, so that the experiments are performed four times varying training and test sets.

In our experiments, the web pages are expressed by binary features, where the value of each feature signifies whether a word presented by the feature occurs in the page or not. The value +1 implies the occurrence of the word and the value 0 represents the absence of it. The document label has +1 or −1 which respectively presents that the document is relevant or irrelevant.

Data Set	Course	Non-Course	Baseline
Cornell	40	203	83.5%
Texas	38	216	85.0%
Washington	74	220	71.1%
Wisconsin	78	220	73.8%
Total	230	821	78.1%

Table 1: The number of web pages used in the experiments.

3.2 TEXT FILTERING BOOSTED BY UNLABELED DATA

We use Quilan's C4.5 release 8 [10] as a base classifier. The merits of decision trees that are distinguished from other learning algorithms are:

- Decision trees are strong classifiers. The stronger the classifiers, the faster the committee converges. This is because the possibility being

[1] The data set is available at http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-51/www/co-training/data/.

Data Set	Using Partially Labeled Data	Using All Labeled Data	Co-Training	Baseline
Cornell	**94.2%**	93.4%	NA	83.5%
Texas	**97.2%**	96.5%	NA	85.0%
Washington	**91.4%**	89.8%	NA	71.1%
Wisconsin	**94.3%**	91.3%	NA	73.8%
Average	**94.3%**	92.8%	93.8%	78.1%

Table 2: The accuracy of web page filtering in active sampling method. The accuracy is measured when it performs best varying the number of labeled data.

Ratio (%)	Using Partially Labeled Data	Using All Labeled Data	Difference (%)
5	21.7%	21.7%	0.0
10	21.7%	21.7%	0.0
15	21.7%	21.7%	0.0
20	21.7%	21.7%	0.0
25	26.9%	29.3%	-2.4
30	39.6%	40.9%	-1.3
35	54.7%	50.0%	4.7
40	67.9%	61.6%	6.3
45	75.8%	72.4%	3.4
50	78.5%	75.6%	2.9
55	90.7%	81.5%	9.2
60	91.9%	85.2%	6.7
65	93.6%	89.7%	3.9
70	92.9%	89.7%	3.2
75	93.6%	90.4%	3.2
80	92.7%	91.6%	1.1
85	92.7%	91.9%	0.8
90	92.7%	92.3%	0.4
95	92.7%	92.4%	0.3
100	92.6%	92.6%	0.0

Table 3: The accuracy difference caused by using unlabeled data. The learning method for the experiment is the active sampling method.

misled by unlabeled data is reduced as the classifiers get stronger.

- There is a fast restructuring algorithm for decision trees. Adding an unlabeled example with a predicted label to the existing set of training examples makes the classifiers restructured. Because the restructuring of classifiers is time-consuming, the proposed methods are of little practical use without an efficient way to restructure. Utgoff et al. [11] presented two kinds of efficient algorithms for restructuring decision trees and showed experimentally that their methods perform well with only small restructuring cost.

For the experiments, 21 decision trees are used to form a committee. According to Breiman [2], the number of classifiers needed to be large when the examples are numerical and more are required with an increasing number of classes. Since all attributes of an example in our experiments are discrete and the number of classes is only two, 21 seems to be reasonable. If there is a tie in predicting classes, the class with the lowest order is chosen as in [2]. For each experiment, 90% of the data are used for training and the remaining 10% are used for testing.

In active sampling method, we set the threshold θ of certainty factor α_t to be 4.0. In AdaBoost-like method, we set the iteration number N to be 10, and use *roulette wheel selection* [5] to resample with distribution D.

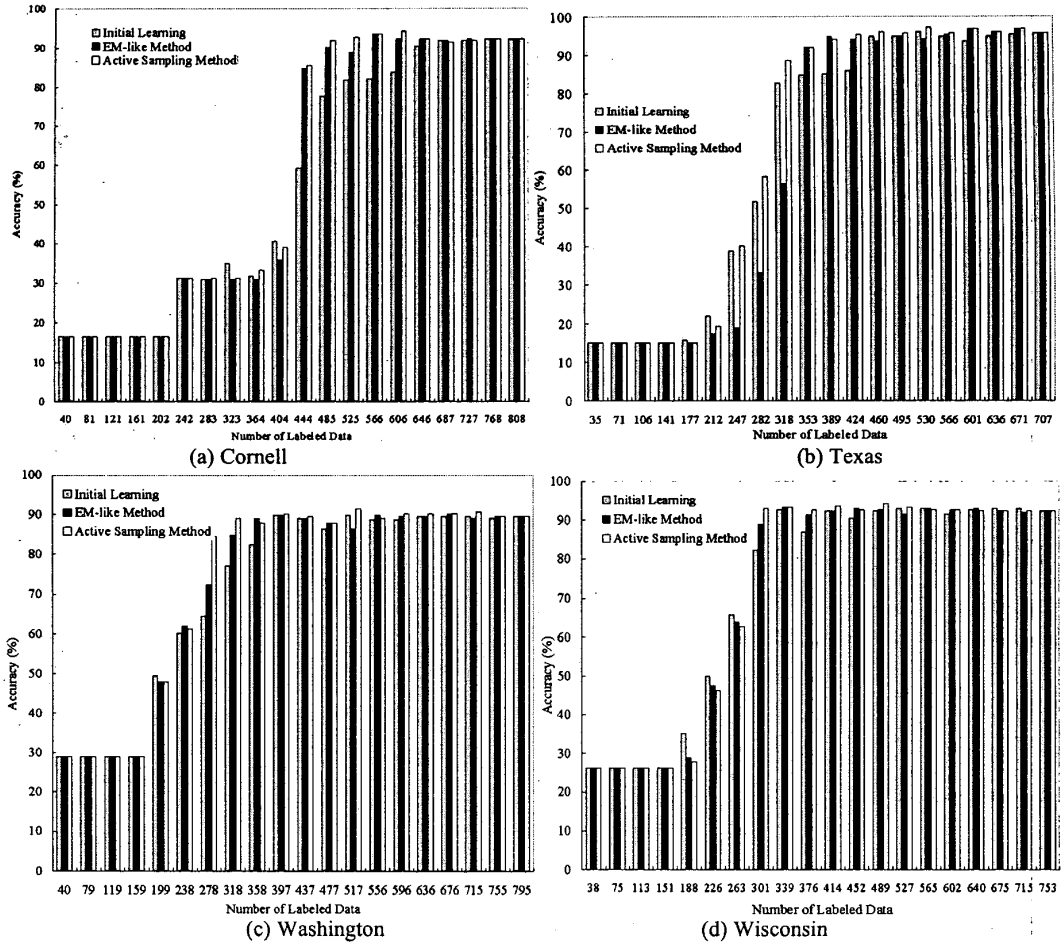

Figure 4: The accuracy improvement by using unlabeled data.

4. RESULTS

Table 2 shows the experimental result of Web page filtering on WebKB data set, where the accuracy is measured when the accuracy is in its best for various ratios of the number of labeled data. The result implies that it improves the filtering performance by using unlabeled data for all data sets. The accuracy is increased by 0.8% for Cornell data set, 0.7% for Texas, 1.6% for Washington and 3.0% for Wisconsin, which is 1.5% improvement on the average. This is implies 16.2% of higher accuracy than the baseline on the average. In addition, the active sampling method achieves 0.5% improvement over Co-Training [1].

Figure 4 shows the usefulness of unlabeled data. It shows experimental results of three methods: *Initial Learning*, *EM-like method*, and *Active Sampling method*. The initial learning in the figures means that the classifier is trained with labeled data, but is not augmented by unlabeled data. According to the experimental results, both methods which use unlabeled data outperform the initial learning. In addition, the active sampling shows slightly higher accuracy than the EM-like method.

The main drawback of the active sampling method is that it requires a lot of unlabeled data when the labeled data are not sufficiently supplied. On the other hand, the AdaBoost-like method overcomes this drawback by resampling difficult data several times. Figure 5 shows

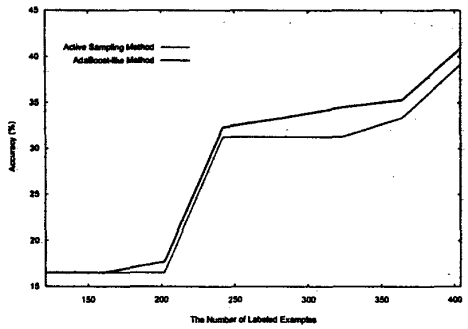

Figure 5: Accuracy comparison of the active sampling method and AdaBoost-like method in the initial state. This experiment was performed on 'Cornell' data set.

experimentally that AdaBoost-like method is better than the active sampling method in the initial state.

Finally, it is interesting to note that the accuracy converges fast by using unlabeled data. In Figure 4, the accuracy of both methods using unlabeled data rises up faster than the initial learning. Especially, Table 3 shows the accuracy gain obtained by unlabeled data. When we use only 55% of labeled data, we achieve, on the average, 9.2% of accuracy improvement.

5. CONCLUSIONS

In this paper, we presented three methods for filtering documents that use unlabeled data to supplement the limited number of labeled data, and compared the performance among them. The small training set of labeled data and the training set is augmented by a large number of unlabeled data, so that the methods overcome the knowledge acquisition bottleneck.

We also showed empirically that unlabeled data enhance the learning methods for document filtering. All the proposed methods outperform the method which does not use unlabeled data by up to 9.2% of accuracy. The active sampling method performs slightly better than the EM-like, and the AdaBoost-like method shows higher accuracy than active sampling method in initial state of learning. In addition, the active sampling method achieves over 90% of accuracy with only 55% of labeled data, while the method without unlabeled data needs 75% of labeled data to achieve same accuracy.

Future study includes understanding why the proposed methods show low accuracy in the initial stage.

ACKNOWLEDGEMENTS

This result was supported in part by the Korean Ministry of Education under the BK21-IT Program and by the Korean Ministry of Information and Communication through IITA under grant 00-023.

REFERENCES

[1] A. Blum and T. Mitchell, Combining Labeled and Unlabeled Data with Co-Training, In *Proc. of COLT 98*, pp. 209-214, 1998.

[2] L. Breiman, Bagging Predictor, *Machine Learning*, Vol. 24, pp. 123-140, 1996.

[3] A. Dempster, N. Laird and D. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Series B, Vol. 39, No. 1, pp. 1-38, 1977.

[4] Y. Freund and R. Schapire, Experiments with a New Boosting Algorithm, In *Proc. of ICML 96*, pp. 148-156, 1996.

[5] D. Goldberg, *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.

[6] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, In *Proc. of ICML 98*, pp. 137-142, 1998.

[7] R. Liere and P. Tadepalli, Active Learning with Committees for Text Categorization, In *Proc. of AAAI 97*, pp. 591-596, 1997.

[8] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, *Learning to Classify Text from Labeled and Unlabeled Documents*, Machine Learning, Vol. 39, pp. 1-32, 2000.

[9] S.-B. Park, B.-T. Zhang and Y.-T. Kim, Word Sense Disambiguation by Learning from Unlabeled Data, In *Proc. of ACL 2000*, pp. 547-554, 2000.

[10] R. Quinlan, *C4.5: Programs For Machine Learning*, The MIT Press, 1993.

[11] P. Utgoff, N. Berkman and J. Clouse, Decision Tree Induction Based on Efficient Tree Restructuring, *Machine Learning*, Vol. 29, pp. 5-44, 1997.

[12] Y. Yang and J. Pederson, Feature Selection in Statistical Learning of Text Categorization, In *Proc. of ICML 97*, pp. 412-420, 1997.

[13] D. Yarowsky, Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, In *Proc. of ACL 95*, pp. 189-196, 1995.

[14] C. Lanquillon, Partially Supervised Text Classification: Combining Labeled and Unlabeled Documents Using an EM-like Scheme, In *Proc. of ECML 2000*, pp. 229-237, 2000.