Automatic Webpage Classification Enhanced by Unlabeled Data

Seong-Bae Park and Byoung-Tak Zhang

School of Computer Science and Engineering
Seoul National University
Seoul 151-744, Korea
sbpark@bi.snu.ac.kr

Abstract. This paper describes a novel method for webpage classification that uses unlabeled data. The proposed method is based on a sequential learning of the classifiers which are trained on a small number 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 classification accuracy is increased. The use of unlabeled data is important because obtaining labeled data, especially in Web environment, is difficult and time-consuming. The experiments on two standard datasets show substantial improvements over the method which does not use unlabeled data.

1 Introduction

Due to the massive volume of online text documents available on the Web, it is important to classify or filter the documents. Especially, automatic webpage classification is of great importance to provide a Web portal service. For the most machine learning algorithms applied to this task, plenty of labeled webpages must be supplied [2]. However, it is very expensive and time-consuming to come by the labeled webpages because labeling must be done by human experts. On the other hand, the unlabeled webpages are ubiquitous and significantly easier to obtain than the labeled ones. Thus, in learning webpage classification, it is natural to utilize the unlabeled data in addition to the data labeled by the oracle.

This paper describes a novel method to classify webpages using both labeled and unlabeled data. We assume that the webpages are represented in the *bag-of-words*. In this scheme, the webpages are represented by the vectors in a space whose dimension equals the size of the *vocabulary*. For simplicity, we shall consider only binary classification problems. The label for a given document vector $\mathbf{x}$ is denoted by $y \in \{-1, +1\}$, where $+1$ represent that the document is relevant and -1 being irrelevant.

According to Cramér-Rao inequality, the mean squared error of unbiased estimator $T(\mathbf{x})$ of the parameter θ is bounded by the reciprocal of the Fisher information. That is, $\mathrm{var}(T) \geq \frac{1}{I(\theta)}$. Since the larger Fisher information produces the smaller variance and the expected error of the estimator is proportional to the variance [3], the larger Fisher information gets, the smaller the expected

J. Liu et al. (Eds.): IDEAL 2003, LNCS 2690, pp. 821–825, 2003.

```
Given unlabeled data set $U = \{\mathbf{x}_1, \ldots, \mathbf{x}_u\}$
    and labeled data set $L = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$,

Train a classifier $f_0$ with $L_0 = L$.
Set $t = 0$ and $\tau_{-1} = 1$.
Do

  1. Calculate $\tau_t = \tau_{t-1} \times \tau$, where $\tau$ is the probability given to the
     negative example in $L_t$ with the highest probability.
  2. Sort data in $L_t$ according to $f_t(\mathbf{x} \in L_t)$.
  3. Sort data in $U_t$ according to $f_t(\mathbf{x} \in U_t)$.
  4. Delete data in $L_t$ and $U_t$ such that $f_t(\mathbf{x}) > \tau_t$.
  5. Set $s = |L_t|$.
  6. Set $U_{add}$ such that $U_{add} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l-s}, y_{l-s}) \mid \mathbf{x} \in U_t, y =$
     $f_t(\mathbf{x})\}$.
  7. Set $L_{t+1} = L_t + U_{add}$.
  8. Train $f_{t+1}$ with $L_{t+1}$.
  9. Set $t = t + 1$.

While $(|U_{add}| > 0$ and $\tau_t > 0.5)$
Output the final classifier: $f^*(\mathbf{x}) = \left(\prod_{i=1}^{k_{\mathbf{x}}-1} \tau_i\right) f_{k_{\mathbf{x}}}(\mathbf{x})$.
```

Fig. 1. The text classification algorithm using unlabeled data. $k_{\mathbf{x}}$ in the final classifier is an index to the best classifier of $\mathbf{x}$.

error of the estimator becomes. Shahshahani and Landgrebe [3] showed that the Fisher information using both labeled and unlabeled data, $I_{labeled+unlabeled}$, is $I_{labeled} + I_{unlabeled}$. Since the Fisher information and variance are reciprocal, using unlabeled data increases the accuracy of the estimator.

However, Zhang and Oles argued that $I_{unlabeled} = 0$ in the semi-parametric models [4]. They argued that active learning is helpful to maximize Fisher information of unlabeled data for the semi-parametric models, and proposed two principles to select informative unlabeled data points:

- Choose an unlabeled data of low confidence with the estimated parameter.
- Choose an unlabeled data that shall not be redundant with other choices.

2 Labeling Unlabeled Text by Sequential Learning

In order to measure the confidence of the unlabeled data, we adopt an idea from SEQUEL (SEQUEnce Learner) proposed by Asker and Maclin [1]. Assume that a classifier f_t produces a probability estimate. That is, $f_t(\mathbf{x})$ gives the probability that a document $\mathbf{x}$ is relevant. And, it has a threshold τ_t which is the probability given to the negative example with the highest probability. For a given data, if the output of the classifier is above the threshold, the classifier is considered *competent* to make a prediction. The confidence of a classifier is set by the number of positive examples such that $\mathbf{x} : f_t(\mathbf{x}) \geq \tau_t$, divided by the total number of examples above the threshold.

In SEQUEL, a set of classifiers for the ensemble is created by varying the set of features, since it considers only labeled data. In training SEQUEL, the first classifier labels some part of data as sure data. Such data have been given a probability of at least τ_t. To determine the next classifier in the sequence, all of the sure examples labeled by the first classifier are removed and the classifier with the highest confidence score for the remaining examples is chosen as the most confident classifier. This process is repeated until the best classifier's confidence score is less than a predetermined threshold.

Figure 1 shows a version of SEQUEL modified to utilize the unlabeled data. It takes two sets of examples as input. A set L is the one with labeled data and U is the one with unlabeled data, where $\mathbf{x}$ is a document and y in L represents the relevancy of $\mathbf{x}$. First of all, the first classifier f_0 is trained with L. After a classifier f_t is trained on labeled data, the threshold τ for the labeled data is calculated and the confidence of f_t is updated by $\tau_t = \tau_{t-1} \times \tau$. The data in both L_t and U_t are sorted according to their margin. The examples in L_t and U_t whose probability is larger than τ_t are removed from the sets, since they give no information in guiding the hyperplane. This coincides with the second principle. The remaining labeled data are augmented by some informative unlabeled data with the predicted labels, so that the number of labeled data for $(t+1)$th step is maintained to be that for tth step. With this new labeled data, a new classifier f_{t+1} is trained.

This process is repeated until the unlabeled examples are exhausted or τ_t gets lower than 0.5, the predefined threshold. For a given unknown document $\mathbf{x}$, the probability of $\mathbf{x}$ being produced by the best classifier, which is indexed by $k_{\mathbf{x}}$, is multiplied by the thresholds of all previous classifiers: $f^*(\mathbf{x}) = \left(\prod_{i=1}^{k_{\mathbf{x}}-1} \tau_i\right) f_{k_{\mathbf{x}}}(\mathbf{x})$.

3 Experiments

3.1 Data Sets

The first data set is the one used in “Using Unlabeled Data for Supervised Learning” workshop of NIPS 2000. This dataset has two kinds of webpage data for the competition of learning unlabeled data. The first problem (P2) is to distinguish the homepages of “http://www.micro soft.com” and those of “http://www.linux.org”, and the second problem (P6) is to distinguish the homepages of “http://www.mit.edu” from those of “http:// www.uoguelph.ca”. Table 1-(a) describes the NIPS 2000 workshop data set. Each document in the dataset is represented in $tf \cdot idf$.

The second data set for the experiments 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 webpages 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-(b) with the number of webpages in each university. The baseline performance shows the accuracy achieved by answering non-course for all examples.

Table 1. The statistics for the dataset.

Data Set	P2	P6
# Labeled Data	500	50
# Unlabeled Data	5,481	3,952
# Test Data	1,000	100
# Terms	200	1000

(a) NIPS 2000 Dataset

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%

(b) WebKB Dataset

Table 2. The effect of removing the confident examples.

	Accuracy	Elapsed Time
Removing Confident Examples	99.5%	992 sec
Not Removing Confident Examples	99.4%	13,337 sec

3.2 Experimental Results

We use the multi-layer perceptron (MLP) as a classifier, since it can provide the framework of probabilistic learning model and shows reasonably high accuracy by itself. The proposed method is implemented on a PC with Pentium III 500MHz and 256MB RAM running Linux. Table 2 gives the advantage obtained by removing the confident examples. The accuracy of the final classifier on P2 of NIPS 2000 workshop data set is little changed though the informative examples are removed, whereas the training time is far reduced. In Web environment, since the input dimension is generally very large, it takes long time to train large number of training data. Thus, it is of great importance to reduce number of training data for practical use.

Table 3-(a) shows the experimental results on NIPS 2000 workshop data set. The result implies that the proposed method improves the classification accuracy for both problems in NIPS 2000 workshop data set by additionally using unlabeled data. The accuracy increase obtained by using unlabeled data is 0.7% for P2 and 15.0% for P6.

In the experiments on WebKB data set, the proposed method achieves the higher accuracy than using all labeled data, where the accuracy of the proposed method is measured when the accuracy is in its best for various ratios of the number labeled data (Table 3-(b)). The accuracy is increased by 0.8% for Cornell data set, 0.6% for Texas, 1.6% for Washington and 2.9% for Wisconsin, which is 1.5% improvement on the average. This also implies the better accuracy of 16.2% than the baseline on the average.

4 Conclusions

In this paper, we presented a novel method for classifying webpages that uses unlabeled data to supplement the limited number of labeled data. The proposed

Table 3. Accuracy of the proposed method.

Labeled Only		Labeled + Unlabeled	
P2	P6	P2	P6
98.8%	60.0%	99.5%	75.0%

(a) NIPS 2000 Dataset

Data Set	Labeled + Unlabeled	Only Labeled	Baseline
Cornell	**94.2%**	93.4%	83.5%
Texas	**97.1%**	96.5%	85.0%
Washington	**91.4%**	89.8%	71.1%
Wisconsin	**94.2%**	91.3%	73.8%
Average	**94.2%**	92.8%	78.1%

(b) WebKB Dataset

learning method first trains a classifier with a small training set of labeled data and the confidence of the classifier is determined. After that, a series of classifiers is constructed in sequence with unlabeled data. The labeled data and some informative unlabeled examples with their predicted labels are used to train the next classifier in the sequence, so that the method ovecomes the knowledge acquisition bottleneck.

We also showed empirically that unlabeled data enhace the learning method for webpage classification. The proposed method outperforms the method which does not use unlabeled data. The classification accuracy is improved by 7.9% for NIPS 2000 data set and up to 9.2% for WebKB data set.

Acknowledgements. This research was supported by the Korean Ministry of Education under the BK21-IT Program, by BrainTech programs sponsored by the Korean Ministry of Science and Technology.

References

1. L. Asker and R. Maclin. Ensembles as a Sequence of Classifiers. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pp. 860–865, 1997.
2. M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell and K. Nigam. Learning to Construct Knowledge Bases from the World Wide Web. In *Proceedings of the Fifteenth National Conference on Artifical Intelligence*, pp. 509–516, 1998.
3. B. Shahshahani and D. Landgrebe. The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigrating the Hughes Pheonomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5), pp. 1087–1095, 1994.
4. T. Zhang and F. Oles. A Probability Analysis on the Value of Unlabeled Data for Classification Problems. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 1191–1198, 2000.