**Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02), Vol. 3**
Lipo Wang, Jagath C. Rajapakse, Kunihiko Fukushima, Soo-Young Lee, and Xin Yao (Editors)

# INCREMENTAL SUPPORT VECTOR MACHINE FOR UNLABELED DATA CLASSIFICATION

*JinHyuk Hong and Sung-Bae Cho*
*hjinh@candy.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr*

*Department of Computer Science, Yonsei University*
*134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea*

## ABSTRACT

Due to the wide proliferation of the internet and telecommunication, huge amount of information has been produced as digital data format. It is impossible to classify this information with one's own hand one by one in many realistic problems, so that the research on automatic text classification has been grown. Machine learning technologies have applied in text classification. However, the traditional statistic machine learning technologies require large number of labeled training examples to learn accurately. To obtain enough training examples, we have to label on these huge training examples by hand. This paper presents a supervised learning algorithm based on support vector machine (SVM) to classify text documents more accurately by using unlabeled documents to augment available labeled training examples. Experimental results indicate that the classification with unlabeled examples using SVM is superior to the conventional classification with labeled examples.

## 1. INTRODUCTION

Research about automatic text classification was begun as one field of information retrieval in 1960's. But until late 1980's, the research had stayed at theoretical level and the real application had been implemented with the rules of expert. Coming up to 1990's, however, with the popularization of the computer and the development of the internet, the data as digital form have suddenly increased and have appeared as the overabundance of information. Under these circumstances, the research of automatic text classification that processes and classifies large amount of information automatically has been embossed and the importance of that work has been recognized widely. Various theories and methods have been studied thoroughly [1, 2, 3].

The recent research on automatic text classification usually is based on the various statistical machine learning techniques. It is because that information is produced abundantly and new information comes up unceasingly. Statistics can handle that information reasonably and effectively and machine learning techniques are useful methods to handle the large amount of information. The expert system or knowledge-based system uses the domain knowledge which needs so many hands and efforts to produce, leading to a limit to manage the excessive information. There is a research of automatic text classification that is independent of domain knowledge and handles lots of information [1].

The machine learning technique that is actively studied in the field of artificial intelligent applies to the domain of automatic text classification. Text classification with the statistical machine learning technique first learns the classifier with enough pre-classified documents using statistical machine learning algorithm. That classifier classifies new documents into predefined classes. K-nearest neighbor, decision tree, support vector machine, neural network, and naïve bayes are famous techniques in statistical machine learning technology. Recently automatic text classification algorithms are adapted to many practical problems in real world like as classifying news articles or web pages automatically, or learning user's taste. Automatic e-mail classification, homepage search, and book recommendation are the applications of those algorithms [1, 3, 4].

To get high performance of those machine learning algorithms, it is necessary to obtain large number of pre-classified examples. In most cases, however, it is impossible to classify all data by user's own hands and to label them, because it costs too much to classify samples one by one. Otherwise, many times in the problem of automatic text classification, it is very easy to gather a lot of unlabeled data.

This paper proposes a way to improve the performance of classifier using a little labeled documents and lots of unlabeled documents. Using support vector machine (SVM) that shows good performance in the text classification, we classify documents as supplement unlabeled docuemnts incrementally to the learning process. In the early stage the SVM is learned with a little labeled documents, and using that SVM unlabeled documents are classified. Among the unlabeled documents, we take a sample under some threshold and label on it. After that, the classifier is trained again with new set of labeled documents and labeled documents of previous iteration with other labeled documents generated by SVM. We show that the repetition of this process improves the performance of automatic text classification.

## 2. RELATED WORKS

### 2.1 Document Classification

The document classification is to categorize a new document into one of the predefined classes automatically. This can be also viewed as searching for the classification function $f: DXC=\{0, 1\}$ close to the optimal classification function $g: DXC=\{0, 1\}$, where $D = \{d_1, d_2, d_3, \ldots, d_i\}$ is a set of documents and $C = \{c_1, c_2, c_3, \ldots, c_j\}$ is a set of predefined classes. The optimal classification function $g$ is membership function between a class $c_i$ ($1 \leq i \leq n$) and a document $d_j$ ($1 \leq j \leq n$). If the value of $f(c_i, d_j)$ is 1, that $d_j$ belongs to $c_i$, and if the value of $f(c_i, d_j)$ is 0, that $d_j$ does not belong to $c_i$. We have to make classification function $f$ as close to the optimal classification function $g$ as possible. This closeness to the optimal classification function measures the performance of classifier [1].

### 2.2 Learning with Labeled/Unlabeled Data

To get the best performance of classifier using machine learning technique, it is prerequisite to prepare enough labeled data, but it costs a great deal to gain the labeled data. Therefore, the research about text classification using unlabeled data to get a classifier of high performance has been actively investigated. According to the Nigam's proof, the improvement of using unlabeled data is due to the fact that unlabeled data gives the information of joint probability between words in documents. The likelihood maximization is one way of the research to combine the labeled and unlabeled data in learning classifier. In that method, text classification seems to be one type of mixture models. Setting the parameters using EM-algorithm, it is finally used in text classification. The other way is selective sampling, integrating unlabeled data to the supervised learning [5, 6, 7, 8].

## 3. INCREMENTAL SUPERVISED LEARNING WITH UNLABELED DOCUMENTS

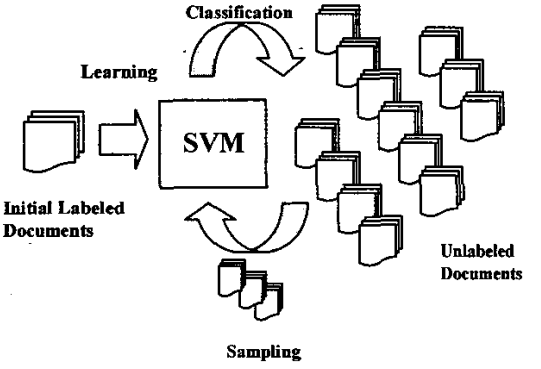


**Figure 1. Automatic text classification using unlabeled documents**

### 3.1 Support Vector Machine

As one of machine learning algorithms, SVM is a good method to get a high performance from real world problems. In the part of automatic text classification, it gets high marks. SVM is suggested by Vapnik and Chervonenkis in 1960s~1970s. Coming to 1990s, there appear real applications using SVM. SVM is a machine learning algorithm for pattern recognition and regression problem [9].

One distinctive feature of SVM against other learning algorithms is that SVM tries to find the optimal hyperplane making expected errors minimized to the unknown test samples. Optimal hyperplane can be expressed as a linear function. The function is expression of data for learning and is one of functions that have minimum VC dimension. According to the structural risk minimization inductive principle, the function is regardless of the dimension of input space. Based on this principle, SVM tries to find a linear function having minimum VC dimension [9, 10].

Most cases, however, the classes are not distinct as linear. So it is necessary to translate from non-linear distribution of data to linear form. SVM offers non-linear function mapping from input vectors to high dimensional feature space where the linear hyperplane can be made on. But it is not always true that the linear hyperplane exists in high dimensional feature space all the time, while it is generally known that it is possible to build the linear SVM in projection space. Figure 2 shows the decision boundary of SVM [9].

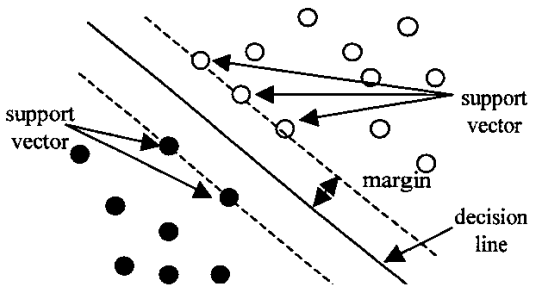


**Figure 2. Decision boundary of SVM in two dimensional spaces**

### 3.2 Proposed Algorithm

Using a special characteristic of SVM-classifier, this paper proposes the algorithm that improves the performance of text classification by incremental addition of unlabeled data to the training process of classifier. Figure 1 shows the structure of classifier that we propose. As described above, the main idea of SVM is finding an optimal decision boundary in vector space by means of the minimization of structural risk. Because of its characteristic, SVM is usually used as a method for binary classification problems. By seeking the proper decision boundary to data set, SVM executes classification.

If unlabeled data are added to the classification, the performance rises up by changing the decision boundary of classifier fitting to them. It is possible to improve the performance of classification, if we select unlabeled data useful to move the optimal decision boundary already learned by labeled data. The process of incremental supervised learning starts from learning SVM with small number of labeled data. Then the SVM puts a label on unlabeled data.

Not all unlabeled data are useful to change the optimal decision surface, so sampling is prior step to labeling on unlabeled data. When the system executes sampling, it samples data whose threshold is over the critical point among unlabeled data. Sampled data are labeled by SVM and included into the set of labeled documents. After that, SVM is retrained with reconstructed labeled data set consisted of pre-labeled data and post-labeled data. We repeat these processes until it satisfies the terminate-condition. To extract reliable information among unlabeled data, we use the semi-parametric model of Zhang and Oles. They gave a point that they have to use active learning to maximize Fisher information of unlabeled data in semi-parametric model. The criterion, to select full informative and unlabeled data, chooses data whose confidence is low by estimated parameter and that data must not be duplicated [ 7].

The traits of SVM are used for sampling unlabeled data helpful to classification. After learning process, SVM gets a decision boundary $w^*x-b=0$ dividing data into two classes. Let us suppose that $x$ is a data to be classified. The further $x$ is from the decision boundary, the higher the probability of correct classification is, and vice versa. Above all, we have to find the decision boundary using SVM learned by labeled data. Then we compute the distance between unlabeled data and decision boundary in hyper space, sampling unlabeled data whose distance is over zero and below two times of margin. Figure 3 shows

Given labeled example set $L=\{(\mathbf{x}_1, y_1) ..., (\mathbf{x}_l, y_l)\}$
and unlabeled example set $U=\{\mathbf{x}_1, ... , \mathbf{x}_u\}$
and test example set $T=\{(x_1, y_1), ..., (x_s, y_s)\}$;
Initialize SVM classifier $f_0$ by training with $L_0=L$, $U_0=U$
Do
1. Set $margin_t$ as margin of $f_t$
2. Set $\tau = margin_t * 2$
3. Set dist(x) = distance between
4. Set $U_{add} = \{(x_1,y_1),...,(x_t,y_t) | x \in U_t, y=f_t(x), dist(\mathbf{x}) \leq \tau \}$
5. Set $L_{t+1}= L_t+U_{add}$
6. Set $U_{t+1}=U_t-U_{add}$
7. Train $f_{t+1}$ with $L_{t+1}$
8.Classify test set T with $f_{t+1}$
9.Set $PR_t$ as classification precision of T
10. Set $t = t+1$
While($|U_{add}|>0$ & $PR_t>PR_{t-1}$)
Store the final SVM classifier

**Figure 3. Pseudo-code of proposed algorithm**

incremental supervised learning with unlabeled data represented as algorithm.

## 4. EXPERIMENTS

To demonstrate the performance of SVMs using unlabeled documents, we have used Reuter-21578 corpus as data set. Reuter-21578 corpus is the set of news that were appeared in Reuters newswire in 1987. We use ModApte topic and particularly the most frequently occurred 7 categories among 135 categories became the targets of experiments. For each category, 20 positive samples and 20 negative samples are selected and used as labeled documents. And we set linear kernel as default kernel function of SVMs. To analyze the performance of classification, we adopt precision/recall, and F1 measure. Four cases are considered as the result of classifier to the documents. Table 1 shows the four cases.

**Table 1. Cases of the classification for one category**

| Class C | | Result of classifier | |
|---|---|---|---|
| | | Belong | Not belong |
| Real classification | Belong | A | B |
| | Not belong | C | D |

Precision means the rate of documents classified correctly among the result of classifier and recall signifies the rate of correct classified documents among them to be classified correctly.

$$precision = \frac{A}{A+C}$$

$$recall = \frac{A}{A+B}$$

F-measure is the index that is combination of precision and recall. Pr means precision and Re means recall. β is a parameter to control the rate of precision and recall. Mostly F1-measure that β is 1 is used.

$$F_\beta = \frac{(\beta^2+1)\cdot Pr \cdot Re}{\beta^2 \cdot Pr + Re}$$

Table 2 is the result of the experiments. First one is the result when SVMs used initial labeled data. And the latter is the result when SVMs learned with unlabeled data incrementally added. In average, precision rises over two times from 27.64% to 59.21%. Recall falls a little. So the value of F1 is also rises one and a half times from 38.67 to 59.22.

**Table 2. The result of each category using proposed algorithm**

| | Result of only using Labeled data | | | Result of incrementally supervised learning | | |
|---|---|---|---|---|---|---|
| | precision | recall | F1 | precision | recall | F1 |
| Grain | 33.60 | 61.94 | 43.56 | 64.35 | 55.22 | 59.44 |
| Earn | 36.21 | 99.41 | 53.08 | 93.91 | 88.09 | 90.91 |
| Acq | 49.96 | 89.32 | 64.08 | 51.08 | 91.49 | 65.56 |
| Money-fx | 19.63 | 89.54 | 32.20 | 22.33 | 88.89 | 35.69 |
| Corn | 11.27 | 48.00 | 18.25 | 51.92 | 54 | 52.94 |
| Crude | 28.87 | 50.60 | 36.76 | 69.32 | 36.75 | 48.03 |
| Wheat | 13.94 | 62.50 | 22.80 | 61.54 | 62.50 | 62.02 |
| Average | 27.64 | 71.62 | 38.67 | 59.21 | 68.13 | 59.22 |

Shown as the vision of iteration in suggested algorithm, at first the performance starts very low. But after each iteration, adding unlabeled documents, total performance of classifier increases continuously until reaching the end of the algorithm. At the last position, the performance declines a little because of the incorrect classification. The stop position of the algorithm is another problem to be considered. In spite of a little loss, whole performance of classifier is improved after some iterations as shown in Figure 4. It chows the result for "corn" news category.

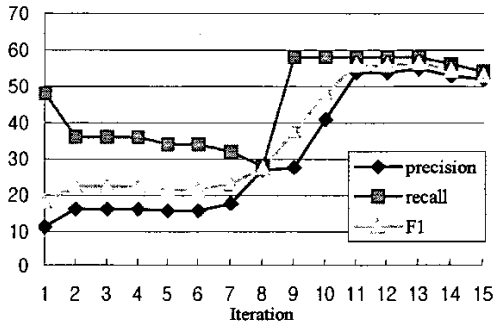


**Figure 4. The result of classification to "corn" news category**

As shown in Figure 5, in every iteration the performance of classifier comes close to that of classifier using all labeled data. S-F1 is the result of the system that uses incremental supervised learning with unlabeled data, and R-F1 is the result of the system that uses incremental supervised learning with labeled data. And All-F1 is the result of the system that uses all data labeled.

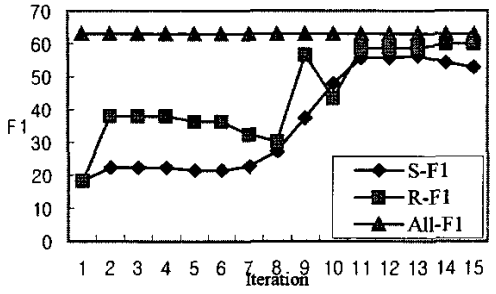


**Figure 5. The comparison of performance between suggested algorithm and learning with labeled data only**

## 5. CONCLUDING REMARKS

This paper suggests incremental supervised learning algorithm with unlabeled documents to promote the performance of text classification and to reduce human efforts. Because statistic machine learning methods need enough data, it cost too much to be applied. But unlabeled data is easy to get from world.

This suggested algorithm is to start from learning with a little number of labeled documents to learning with incrementally added unlabeled documents. This algorithm is designed from using the features of SVMs and the characteristics of unlabeled data in semi-parametric model. The experiment is practiced to the most frequently occurred classes in the Reuter-21578. Comparing the result by using unlabeled data incrementally with the result by only using labeled data, in all the classes shows that the performance of classification is improved. In average, 100% of improvement is occurred in precision, and the value of F-measure rises up to 150%.

As can be seen, incremental supervised learning with unlabeled documents marks better performance than just learning with labeled documents does.

## 6. ACKNOWLEDGEMENTS

This paper was supported by Brain Science and Engineering Research program sponsored by Korean Ministry of Science and Technology.

## 7. REFERENCE

[1] F. Sebastiani, "Machine learning in automated text categorisation," Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999.

[2] V. Gudivada, et.al, "Information retrieval on the world wide web," *IEEE Internet Computing*, Vol. 1, no. 5, September/October, 1997.

[3] Y. Yang, "An evaluation of statistical approaches to text categorization," *Journal of Information Retrieval*, vol 1, no. 1/2, pp. 67-88, 1999.

[4] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to construct knowledge bases from the world wide web," *Artificial Intelligence, 118(1-2)*, pp. 69-113, 2000.

[5] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, "Learning to classify text from labeled and unlabeled documents," *In Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, pp. 792-799, 1998.

[6] Z. Ghahramani and M. Jordan, "Supervised learning from incomplete data via an EM approach." *Advances in Neural Information Processing Systems 6*, pp.120-127, 1994.

[7] T. Zhang and F. Oles, "A probability analysis on the value of unlabeled data for classification problems," *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 1191-1198, 2000.

[8] D. Cohn, L. Atlas and R. Landner, "Improving generalization with active learning," *Machine Learning, 15*(2), pp. 201-221, 1994.

[9] V. Vanpik, "The Nature of Statistical Learning Theory," Springer-Verlag, 1995.

[10] T. Joachims, "Transductive inference for text classification using support vector machines," *Proceedings of ICML-99, 16th International Conference on Machine Learning*, 1999.