http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (23): 7673-7676, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7673.7676

Progressive Similarity Transductive Support Vector Machine Algorithm for Small Sample Text Classification

[1]Jianbin Ma and [2]Ying Li
[1]College of Information Science and Technology,
[2]College of Economic and Trade, Agricultural University of Hebei, Hebei, 071001, Baoding, China

Abstract: Support Vector Machine (SVM) algorithm is applied to text classification widely. However, SVM's limitation is that it is difficult to label samples rightly if available training samples are small. So TSVM (Transductive Support Vector Machine) was introduced to minimize misclassification of test samples via., training on labeled and unlabeled samples. However, in the training process of TSVM, the parameter N (the number of positive samples) should be inputted artificially. The parameter N is difficult to estimate. In this study, PSTSVM (Progressive Similarity Transductive Support Vector Machine) was introduced which labeled most likely unlabeled samples pairwise by similarity computing and then retrained to readjust the hyperplane. The experimental results on Reuters dataset showed that PSTSVM algorithm was effective on a mixed training set of unlabeled samples and labeled samples.

Key words: PSTSVM, small sample, text classification, support vector machine

INTRODUCTION

Text classification is the task of automatically classifying a set of documents into categories from predefined samples which have widely applied to spam filtering (Drucker *et al*., 1999), authorship attribution (Ma *et al*., 2009, 2013), hierarchical catalogues of web resources (Wang *et al*., 2008) etc. Machine learning technique is the most used text classification method recently which form training model based on known properties learned from the training text samples to classify unknown documents automatically. There are several typical machine learning algorithm including Support Vector Machine (SVM), Decision Tree, Artificial Neural Network, Genetic Programming, K-nearest Neighbors etc. SVM algorithm was introduced by Vapnik (1998) based on the structural risk minimization principle of the computational theory. SVM algorithm could solve high-dimensional and non-linear classification application which was proved to be one effective algorithm for text classification (Joachims, 1998, 2001). To find optimal classifier, abundant training samples are required to apply SVM to text classification. However, samples are difficult to collect for many applications such as authorship attribution, information filtering. A problem of text classification using SVM is difficult to construct so many labeled samples. Either there is small available training set, or it is almost impossible to classify the unlabeled samples into labeled samples manually.

To solve small sample text classification problem, Transductive Support Vector Machine (TSVM) was introduced by Joachims (1999). TSVM is a kind of semi-supervised learning method which can optimize the trained classifier with poor representative labeled samples by mining the helpful information from great amount of unlabeled samples (Ren *et al*., 2010). So, it is suited well for small training set of text classification. However, in the training process of TSVM, the parameter N (the number of positive samples) should be inputted artificially before the algorithm is executed, but the value of N is very difficult to estimate accurately in the ordinary circumstances. Progressive Transductive Support Vector Machine (PTSVM) was introduced by Chen *et al*. (2003a, b) which labeled unlabeled samples pairwise and then retrained the support vector machine to readjust the hyperplane. Unlabeled samples with maximal decision function value are chosen and labeled. In general, samples having maximal decision function value are points in the class of boundary. These samples belong to the class with less confidence. This study aims to improve PTSVM and introduce one new algorithm named Progressive Similarity Transductive Support Vector Machine (PSTSVM). The basic idea of PSTSVM is labeling most likely unlabeled samples pairwise by similarity computing method and retaining to adjust the hyperplane iteratively to find the optimal hyperplane.

The rest of the study is organized as follows. Section 2 describes TSVM algorithm. Section 3 describes

Corresponding Author: Jianbin Ma, College of Information Science and Technology, Agricultural University of Hebei, Hebei, 071001, Baoding, China

our PSTSVM algorithm in detail. Section 4 is our experimental methodology and analysis the experimental results. And conclude the study in section 5.

TRANSDUCTIVE SUPPORT VECTOR MACHINE

The essential theory of TSVM is that it can learn from a mixed training set of unlabeled samples and labeled samples and tries to minimize misclassification of test samples.

Given a set of independent identically distributed labeled samples:

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \cdots (\vec{x}_n, y_n), \vec{x} \in X, y \in \{-1, +1\} \quad (1)$$

Given another set of independent identically distributed unlabeled sample S_{test}:

$$\vec{x}_1^*, \vec{x}_2^*, \cdots, \vec{x}_k^* \quad (2)$$

TSVM aims to select a function $h_L = L(S_{train}, S_{test})$ using S_{train} and S_{test} so that the expected erroneous predictions on the test samples is minimized:

$$R(L) = \int \frac{1}{k} \sum_{i=1}^{k} \theta(h_L(\vec{x}_i^*), y_i^*) dP(\vec{x}_1, y_1) \cdots dP(\vec{x}_k^*, y_k) \quad (3)$$

In a general linearly non-separable data case, the learning process of TSVM can be formulated as the following optimization problem.

Minimize over $(y_1^*, \cdots y_n^*, \vec{w}, b, \xi_1, \cdots \xi_n, \xi_1^*, \cdots \xi_k^*)$:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + C^*\sum_{i=1}^{n} \xi_j^* \quad (4)$$

Subject to:

$$\forall_{i=1}^{n} : y_i[\vec{w} \cdot \vec{x}_i + b] \geq 1 - \xi_i$$

$$\forall_{j=1}^{k} : y_j^*[\vec{w} \cdot \vec{x}_j^* + b] \geq 1 - \xi_j^*$$

$$\forall_{i=1}^{n} : \xi_i > 0$$

$$\forall_{j=1}^{k} : \xi_j^* > 0$$

In the training process of TSVM, the parameter N of the positive samples must be specified and usually the parameter N is difficult to estimate accurately. The simple method to estimate the parameter N is calculating the ratio of positive labeled samples to all the labeled samples. However, the number of positive labeled samples has little relative to unlabeled samples. So the estimated N has a big deviation from the actual number of the positive samples. Then, the performance of the classifier decrease significantly.

PROGRESSIVE SIMILARITY TRANSDUCTIVE SUPPORT VECTOR MACHINE (PSTSVM)

To solve estimation of parameter N, this study introduces an algorithm named Progressive Similarity Transductive Support Vector Machine (PSTSVM). The basic idea of the algorithm is labeling the samples dynamically according to a certain principle. Firstly, one initial classifier is trained on labeled samples. The unlabeled samples are classified into two classes by initial classifier. To adjust the current separating hyperplane to the correct orientation, one most likely unlabeled positive sample and negative sample are labeled first. Then retrain the Support Vector Machine to readjust the hyperplane. The algorithm iterate until all the unlabeled samples are labeled. The method is progressive labeling and dynamical adjusting to find optimal hyperplane. Specifically which sample should be labeled first? Apparently, sample with strongest confidence should be labeled first. In this study, similarity computing method was proposed to label samples with strongest confidence.

Similarity computing method: Figure 1, suppose write circles denote positive labeled samples, black circles denote negative labeled samples, grey circles denote unlabeled samples. The initial classifier training on labeled samples classifies the unlabeled samples into two classes. Seven unlabeled samples are classified into positive class. The degree that the fifth unlabeled sample belongs to the positive class is more than that of the third unlabeled

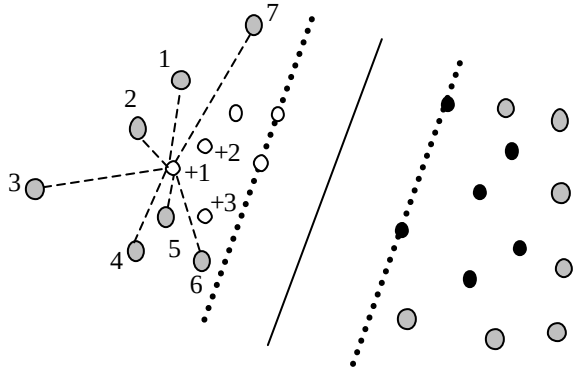

Fig. 1: Marked principle of unlabeled samples

sample. Because the fifth unlabeled sample is similar to the first labeled sample (+1) and the third labeled sample (+3).

To compute which unlabeled sample is most similar to one of the labeled samples, the similarity computing method is proposed. Equation 5 is the similarity computing formula of two vectors:

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2)(\sum_{k=1}^{M} W_{jk}^2)}} \quad (5)$$

where, d_i and d_j denote two samples which are represented as M dimension vectors. W_{ik} and W_{jk} denote the weight of kth feather in d_i and d_j vector.

The similarity weight of labeled sample to one unlabeled sample is computed by Eq. 6:

$$W_{ij} = \frac{sim(x_i, y_j)}{\sum_{k=1}^{M} sim(x_i, y_k)} \quad (6)$$

W_{ij} denotes the similarity weight of the labeled sample x_i and unlabeled sample y_j. M is the number of unlabeled samples which are classified into the positive class or negative class by original classifier. The denominator of formula is the normalization factor.

The most likely unlabeled sample that belongs to one class is the sample with strongest similarity weight. Suppose W_{+0} denotes the similarity weight of labeled positive sample to unlabeled sample which are classify into positive class by initial classifier. W_{-0} denotes the similarity weight of labeled negative sample to unlabeled samples which are classify into negative class by initial classifier. Then Eq. 7 is the labeling condition:

$$\begin{cases} Max(W_{+0}) \\ Max(W_{-0}) \end{cases} \quad (7)$$

Description of PSTSVM: Based on the similarity computing method, most likely unlabeled sample is labeled first. The idea is progressive labeling and dynamical adjusting to find optimal hyperplane. Following is the detail description of PSTSVM:

- Execute an initial training on labeled samples and generate an initial classifier. Classify the unlabeled samples into positive and negative class
- Compute the similarity weight of labeled sample and unlabeled sample based on Eq. 6. Label one positive unlabeled sample and one unlabeled negative sample pairwise with largest similarity weight based on Eq. 7
- Retrain the support vector machine over all labeled samples. Classify the unlabeled samples into positive and negative class
- Return to step 2. If there are not positive unlabeled or negative unlabeled samples, label all remaining unlabeled samples with current separating hyperplane and output the result

Table 1: Comparative results of PSTSVM and TSVM algorithm

	No. of Unlabeled samples		
Algorithm	Positive	Negative	Aaccuracy (%)
TSVM	10	20	83.33
PSTSVM	10	20	96.67
TSVM	10	30	67.50
PSTSVM	10	30	90.00

EXPERIMENTS

To test the effect of PSTSVM, Reuter dataset were collected. Reuter's dataset was collected from Reuters's newswire in 1987 which was one typical experimental dataset for text classification. Five positive samples and five negative samples were labeled beforehand. Accuracy of PSTSVM was compared with that of TSVM. The classification results were showed in Table 1.

From Table 1, we could see that the classification accuracy of PSTVM was higher than that of TSVM. The reason is that TSVM estimate the proportion between positive and negative labeled samples. That is, the parameter N is 0.5 in the experiment. Obviously, the parameter is not accurate. PSTSVM makes use of the information about the data distribution implicitly carried in the unlabeled samples more reasonably. In Table 1, TSVM's accuracy was 83.33% and PSTVM's accuracy was 96.67% on condition that 10 unlabeled positive samples and 20 negative samples. However, TSVM's accuracy was 67.5% and PSTVM's accuracy was 90% on condition that 10 unlabeled positive samples and 30 negative samples. Therefore, the proportion between positive and negative unlabeled samples is more unbalance; accuracy of PSTSVM is much higher than TSVM. The accuracy of PSTSVM exceeded 90%. The experimental results showed that PSTSVM algorithm was effective on a mixed training set of unlabeled samples and labeled samples.

CONCLUSION

To solve small sample text classification problem and estimation of parameter N, one semi-supervised algorithm named Progressive Similarity Transductive Support Vector Machine (PSTSVM) was introduced in this study. The algorithm labels most likely unlabeled samples pairwise by similarity computing and retrains to adjust the

hyperplane. Experiments on reuters dataset showed that classification accuracy of PSTSVM was higher than TSVM. Further, the proportion between positive and negative unlabeled samples is more unbalance; accuracy of PSTSVM is much higher than TSVM. Experimental results showed that PSTSVM algorithm was effective on a mixed training set of unlabeled samples and labeled samples.

ACKNOWLEDGMENT

This study was supported by grants from Department of Education of Hebei Province (No. QN20131150), Agricultural University of Hebei (No. LG20110502). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers which have improved the presentation.

REFERENCES

Chen, Y.S., G.P. Wang and S. Dong, 2003. Learning with progressive transductive support vector machine. Pattern Recogn. Lett., 24: 1845-1855.

Chen, Y.S., G.P. Wang and S.H. Dong, 2003. A progressive transductive inference algorithm based on support vector machine. J. Software, 14: 451-460.

Drucker, H., D. Wu and V.N. Vapnik, 1999. Support vector machines for spam categorization. IEEE Trans. Neural Network, 10: 1048-1054.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Joachims, T., 1999. Transductive inference for text classification using support vector machines. Proceedings of the 16th International Conference on Machine Learning, June 27-30, 1999, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA., pp: 200-209.

Joachims, T., 2001. A statistical learning model of text classification for support vector machines. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, USA., pp: 128-136.

Ma, J.B., G.F. Teng, Y.X. Zhang, Y.L. Li and Y. Li, 2009. A cybercrime forensic method for chinese web information authorship analysis. Proceedings of 2009 Pacific Asia Workshop on Intelligence and Security Informatics, April 27, 2009, Bangkok, Thailand, pp: 14-24.

Ma, J.B., Y. Li, G.F. Teng and Y.X. Zhang, 2013. An authorship attribution forensic method for web information. ICIC Exp. Lett., 7: 2609-2613.

Ren, G.B., J. Zhang, Y. Ma and P.J. Song, 2010. An unlabeled samples labeling method of TSVM for remote sensing image. Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology, July 9-11, 2010, Chengdu, China, pp: 286-290.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York, USA.

Wang, Y. and Z. Gong, 2008. Hierarchical classification of web pages using support vector machine. Proceedings of the 11th International Conference on Asian Digital Libraries, December 2-5, 2008, Bali, Indonesia, pp: 12-21.