Available online at www.sciencedirect.com

SciVerse ScienceDirect

AASRI Procedia 1 (2012) 384 – 388

AASRI Procedia

www.elsevier.com/locate/procedia

2012 AASRI Conference on Computational Intelligence and Bioinformatics

A Transductive Support Vector Machine Algorithm Based on Spectral Clustering

Xu Yu*, Jing Yang, Jian-pei Zhang

School of Computer Science, Harbin Engineering University,145, Nantong street,Nangang district, Harbin, 150001, China

Abstract

As transductive inference makes full use of the distribution information of unlabeled samples, so compared with traditional inductive inference, it is always more precise. For the question that transductive support vector machine algorithm (TSVM) is with a low classification precision and an unstable learning performance, in this paper, a transductive support vector machine algorithm based on spectral clustering (TSVMSC) is proposed. Firstly, the algorithm utilizes spectral clustering algorithm to divide the unlabeled samples into several clusters, then labels them with the same class, finally makes transductive inference on the mixed data set composed by both labeled and unlabeled samples. It is not necessary to estimate the ration of the positive samples to the negative samples, so the stability can be improved largely. Meanwhile, the prior distribution information of the unlabeled samples can further be extracted, so the classification precision can be improved effectively. The experiments show that TSVMSC exhibits a superiority over TSVM both at the stable performance and classification precision.

Keywords: Transductive inference; Spectral clustering; Support vector machine; Unlabeled samples

1. Introduction

Statistic learning theory is proposed by Vapnik (Vapnik, 1995), which is a theory on small sample problem. SVM is a new generation of leaning algorithms based on Statistic learning theory (Cortes and Vapnik, 1995),

* Corresponding author. Tel.: +86-13624508906; fax: +86-0451-82519602.
E-mail address: yuxub09@hrbeu.edu.cn.

doi:10.1016/j.aasri.2012.06.059

and it is the most practical and youngest part. Currently SVM develops constantly. Although traditional SVM can deal with high-dimensional, non-linear and small sample data set, it solves a more complex problem in the classification process of finite samples. Thus it obeyed the following basic principle that when we solved a given problem, we should avoid to solve a more general problem in the middle process. Based on this idea, Vapnic proposed a transductive thought (Gammerman et al., 1998) and present a transductive support vector machine algorithm which can improve the performance of classifier effectively.

This article is a further study on transductive learning, trying to find a more common transductive learning algorithm than the existing methods. According to the inherent characteristics of support vector machine classification, this article design a transductive support vector machine algorithm based on spectral clustering (Shi and Malik, 2000). The present classifier first uses spectral clustering to cluster the similar non-label samples based on the good results of spectral clustering. Then it marks all the samples in the same cluster with the same label. Finally the classifier uses transductive learning method to do classification. As this method uses spectral clustering to mine the effective priori distribution information of the unlabeled samples, so it can effectively improving the performance of the transductive support vector machine algorithm.

Section 2 introduces the idea of transductive learning and described the transductive support vector machine classification algorithm presented by Joachims (Joachims, 1999). Based on the characteristics of traditional support vector machine classifiers and the good result of spectral clustering, we present a transductive support vector machine learning algorithm based on spectral clustering and give the detailed processes of the algorithm. In Section 4, we give the experimental results and make a detailed analysis. In Section 5, we conclude the whole paper and point out that further research directions and ideas.

2. Overview of transductive inference

Transductive inference was first proposed by Vapnik. The difference between transductive learning and traditional learning is that the traditional classifier first obtained the functions according to the training sample set, then judge the class label of test samples according to the classification functions, while transductive classifiers directly use the training sample set to judge the class labels of the test samples.

Let training samples be $(\mathbf{x}_1, y_1),\cdots,(\mathbf{x}_l, y_l)$, $y \in \{-1,1\}$, which come from the same distribution independently. Let $\mathbf{x}_1^*,\cdots,\mathbf{x}_k^*$ be non-labeled samples following the same distribution. Under the non-linear separable condition, the transductive learning based on support vector machines can be described as the following optimization problems.

$$\begin{aligned} \min_{y_1^*,\cdots,y_k^*,\mathbf{w},b,\xi_1,\cdots,\xi_n,\xi_1^*,\cdots,\xi_k^*} \quad & \frac{1}{2}\|\mathbf{w}\| + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{k}\xi_j^* \\ y_i((\mathbf{w}\cdot\mathbf{x}_i)+b) \geq 1-\xi_i \quad & i=1,...,n \\ y_j((\mathbf{w}\cdot\mathbf{x}_j^*)+b) \geq 1-\xi_j^* \quad & j=1,...,k \\ \xi_i \geq 0, \quad & i=1,...,n \\ \xi_j^* \geq 0, \quad & j=1,...,k \end{aligned} \tag{1}$$

In general, the exact solution of this problem requires a search of 2^k cases in sample set, which is an NP problem. For a small number of samples, this process can be completed. For a large number of unlabeled samples, however, it is necessary to use a variety of approximate optimization algorithm to obtain the solution.

Currently, the most important outcome in this respect is Transductive Support Vector Machine (TSVM) proposed by Joachims. Below, we will briefly introduce the principle and realization of TSVM algorithm.

The training process of TSVM is the process of solving the above optimization problem. The training algorithm can be divided into the following steps:

Step 1: Set parameter *C* and *C**, and use inductive SVM on the training data to obtain an initial classifier. Set the number of positive label samples, *N*, according to a rule.

Step 2: Compute the decision function values of all the unlabeled examples with the initial classifier. Label N examples with the largest decision function values as positive, and the others as negative. Set a temporary effect factor C_{tmp}^{*}.

Step 3: Retrain the support vector machine over all the examples. For the newly yielded classifier, switch labels of one pair of different-labeled unlabeled examples using a certain rule to make the value of the objective function in (1) decrease as much as possible. This step is repeated until no pair of examples satisfying the switching condition is found.

Step 4: Increase the value of C_{tmp}^{*} slightly and return to Process 3. When $C_{tmp}^{*} > C^{*}$, the algorithm is finished and the result is output.

3. A transductive support vector machine algorithm based on spectral clustering

As TSVM algorithms use the idea of transductive learning effectively, so it can combine the implied distribution information of unlabeled samples and training samples well. Therefore, compared with the traditional support vector machine algorithm, TSVM algorithm is with a higher classification accuracy.

However, TSVM algorithm still exists a number of shortcomings, such as the number N of positive label samples in the unlabeled samples for TSVM algorithm must be artificially specified, but N value is usually very difficult to make a reasonable estimate.

TSVM algorithm uses a simple method to estimate the value of N, which is to estimate the proportion of samples with positive labels to all the unlabeled samples according to the proportion of samples with positive labels to all the labeled samples. However, when the number of samples with labels is small, the method is difficult to estimate a more accurate value of N. Once the pre-set value of N is quite different from the actual number of samples with positive labels, the performance of the TSVM algorithm will become very unstable, and moreover the classification accuracy of the algorithm can not be assured effectively.

As the performance of the TSVM algorithm is very unstable, so in this paper, we propose a Transductive Support Vector Machine based on Spectral Clustering, referred to as TSVMSC.

The basic though of TSVMSC is as follows. First of all, cluster the similar unlabeled sample effectively according to the good clustering performance of spectral clustering. Then mark all the samples in the same cluster with the same label. Finally learn with transductive support vector machine on all the labeled and unlabeled samples, which is to find a solution of optimization problem (1). Below we present the detailed steps of TSVMSC.

Step 1: Specify clustering number k (usually set $3<k<7$), and specify the parameters C and C^{*}.

Step 2: Use spectral clustering algorithm to cluster the sample set, and divide it into k clusters.

Step 3: Mark all the samples in the same cluster with the same label, and search 2^{k} cases in sample set to solve optimization problem (1).

Step 4: Output the class labels of unlabeled samples.

4. Experimental results and analysis

In order to test the performance of TSVMSC proposed in this paper, we make an experiment on the Reuters newswire data sets, the Reuters-21578. The dataset comes from Reuters newswire in 1987, which is a typical text classification dataset (Joachims, 2001). Our experimental data is obtained by adapting Joachims's pre-treated data set.

Each word is represented by a feature vector in high dimensional space, and each feature corresponds to a word root. The objective of this experiment is to train a classifier to find the texts related with corporate acquisition. This experiment selects Radial Basis Function (RBF) as kernel function (Burges, 1998):

$$K(\mathbf{x},\mathbf{y}) = \exp(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}) \quad (2)$$

where σ is a width parameter, "**x**" and "**y**" are n-dimensional vectors in the original feature space.

In all experiments, the training set is composed by 5 positive labeled samples and 5 negative labeled samples. For TSVM algorithm, select different number of unlabeled samples to test the performance of TSVM algorithm in different data distributions of unlabeled samples. For TSVMSC algorithm, set k=3,5,7 in the spectral clustering algorithm. In all the experiments, we use the same test set that contains 300 positive labeled samples and 300 negative labeled samples. Table 1 shows the learning results of TSVM algorithm and TSVMSC algorithm when k=5, and Table 2 shows the learning results of TSVMSC algorithm when k=3,5,7, respectively.

Table 1. Learning results comparision on Reuters data set between TSVM and TSVMSC

Number of unlabeled samples	Algorithm	Training time (s)	Test precision (%)
Pos=500	TSVM	133.66	89.16
Neg=500	TSVMSC	105.32	90.21
Pos=500	TSVM	156.03	95.26
Neg=1000	TSVMSC	115.35	92.19
Pos=1000	TSVM	160.96	72.69
Neg=500	TSVMSC	115.35	92.19
Pos=1000	TSVM	300.56	91.10
Neg=1000	TSVMSC	236.11	93.56

Table 1 shows that the TSVM algorithm is not very stable, because for TSVM algorithm the number of positive labeled samples, N, in the unlabeled samples set must be artificially specified. However, the value N is often difficult to make a reasonable estimate. If N is estimated accurately, the accuracy of the TSVM algorithm is relatively high, but if N is not estimated very accurately, the performance of the algorithm can be greatly affected. However TSVMSC algorithm does not need to estimate the values of N, so the stability of TSVMSC algorithm is greatly improved compared to TSVM algorithm. As TSVMSC algorithm uses spectral clustering algorithm to cluster similar samples, and marks them with the same label, so actually this algorithm use the priori distribution information of unlabeled samples effectively, so the algorithm can reach a higher level accuracy.

Table 2. Learning results comparision of TSVMSC among different cluster number

k	Number of unlabeled samples	Training time (s)	Test precision (%)
3		75.60	86.37
5	1000	105.32	90.21
7		165.23	95.91
3		108.16	88.18
5	1500	115.35	92.19
7		181.22	96.77
3		198.77	89.67
5	2000	236.11	93.56
7		310.16	95.52

It can be shown from Table 2 that with the increase of k, the running time of TSVMSC algorithm becomes longer, which is very easy to understand as the transductive support vector machine learning algorithm needs to solve the optimization problem (1). With the increase of k, the accuracy of TSVMSC algorithm on the test set can be increased. This is mainly because the optimization problem can be solved more accurately with the increase of k, so the generalization ability of the algorithm is increased, and correspondingly the accuracy on the test set can be improved.

5. Conclusion

After a brief introduction of the thought of transductive learning and transductive support vector machine learning algorithm, TSVM, we proposed a Transductive Support Vector Machine based on Spectral Clustering, referred to as TSVMSC.

The algorithm avoids initially estimating the ratio of positive labeled samples and negative labeled samples in the unlabeled sample set. Moreover, the spectral clustering algorithm can mine the priori distribution information of the unlabeled samples effectively. The experimental results show that TSVMSC algorithm can achieve a good effect on stability and accuracy.

As transductive learning is a relatively new area of research, so many issues are worthy of studying furtherly. The next research direction is how to determine the number of clusters k more rationally, and how to improve TSVMSC algorithm furtherly according to the data distribution characteristics and the thoughts of different algorithms.

Acknowledgements

This project is supported by the National Natural Science Foundation of China (Nos. 60873037,61073041,61073043), the Natural Science Foundation of Heilongjiang Province of China (No. F200901), Harbin Outstanding Academic Leader Foundation of Heilongjiang Province of China (No. 2011RFXXG015), and Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20112304110011).

References

[1] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995.
[2] Cortes C, Vapnik VN. Support vector networks. Machine Learning. 1995;20:273-297.
[3] Gammerman A, Vapnik VN, Vowk V. Learning by transduction. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Wisconsin; 1998;148-156.
[4] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;22:888-905.
[5] Joachims T. Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning (ICML). San Francisco: Morgan Kaufmann; 1999;200-209.
[6] Joachims T. SVMlight. http://www-ai.cs.uni-ortmund.de/SOFTWARE/SVM_LIGHT/svm_light.eng.html, 2001.
[7] Burges CJC. A tutorial on support vector machine for pattern recognition. Data Mining and Knowledge Discovery. 1998;2:121-167.