THE IMPLEMENTATION OF ONLINE TRANSDUCTIVE SUPPORT VECTOR MACHINE

ZHANG Xihuang,XU Wenbo
School of Information Engineering, Southern Yangtze University
214122 Wuxi, Jiangsu, P.R.China

Abstract: For many learning tasks, samples are collected over a long period of time. The distribution of the data set is likely to change. And only a little labeled training data is available at beginning of SVM training. So the collected data could not represent the whole data set. SVM should be able to adapt to such changes and situation. In order to achieve an acceptable performance with fewer labeled training samples at beginning of SVM studying, after studying transductive inference method, an Online tranductive SVM based on feedback is well organized based on TSVM. To deal with the expending size of support vector window, the $\xi\alpha$ Estimators is introduced, which will maintain the window on the training data by automatically adjustment, The estimated generalization error is minimized. The approach is both theoretically well founded as well as effective and efficient in practice. Two experiments show OTSVM is effective

Key words: support vector machine; transductive inference; $\xi\alpha$ estimator; online tranductive

1. INTRODUCTION

A major concern with machine learning and supervised learning techniques such as support vector machine (SVM) is based on very complete, absolute, and strict theories [1,2,3,4]. But it is not suitable for the situation where data is collected over an extended period of time, or the data is likely to change over time. In many on-line applications this introduces the problem that the distribution of the data is likely to change over time. To meet this problem changed over time, a traditional machine leaning system should be able to adapt to such changes. A second problem in many real world applications is that only little labeled training data is available. Since only few labeled

training samples are provided, users often take partial feedback, which could be able to achieve a good performance.

After introduce the basic support vector machine (SVM) principle, This paper improve the transductive inference method which could detect and handle concept changes with support vector machines extending the approach by using unlabeled data to reduce the need for labeled data [5,6,7]. The approach has a clear theoretical motivation and does not require complicated parameter tuning. At last, by the adjustment approach of sample window, an on-line SVM is suggested. The experiments on Hoovers Data set show that the on-line SVM approach effectively selects an appropriate window size and results in a low predictive error rate. In case of few labeled samples, the use of unlabeled data can also be expected to improve the performance of the on-line SVM. When the amount of unlabeled data is used, the performance of the system could be in an ideal level.

2. SUPPORT VECTOR MACHINES

Support vector machines are based on the structural risk minimization principle from statistical learning theory. In basic form, SVM learn linear decision rules. SVM learning can be well analyzed by using the feature vector space.

Let z be a set of training samples

$z = (x1,y1),(x2,y2),\cdots(xn,yn), xi \in Rm, yi \in \{-1,+1\}$

where $xi \in Rm$ is an m-dimensional input vector and $yi \in \{-1,+1\}$ is the label of xi.

According to Vapnik's theory about support vector machine. The solution of SVM could be described as the following optimized problem

$$\begin{aligned} &MinW(\alpha) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t\ \ y_i(w\cdot x_i + b) \geq 1-\xi_i,\ \xi_i \geq \quad 0 \leq \alpha_i \leq C,\quad \sum_{i=1}^{l}\alpha_i y_i = 0 \end{aligned} \tag{1}$$

Here, ξi is a relax parameter and C is an impact factor. $y = (w \cdot x + b)$ describes the plane to classify the samples[1,4].The solution of above problem is also given by Vapnik[1,4].

3. TRANSDUCTIVE INFERENCE LEARNING

The algorithm of transductive inference learning based on support vector machine is that both the labeled and unlabelled samples are used in the process of learning. If the distributed information of unlabelled samples is

carried into the new classifier, the new classifier will be trained better because the data characteristic of the whole samples space is depicted well. At present the transductive inference leaning based on support vector machine is at the preliminary stage. The representative one is TSVM[5,6,7,8].

The following labeled training samples are

(x1,y1),(x2,y2),⋯(xn,yn),x ∈Rm, yi ∈{-1,+1}

and another group of unlabelled samples with the same distribution are

x1,x2*, ⋯xk**. The training process of TSVM is the generalization of the training process of support vector machine. It can be described as the following optimized problem:

$$\begin{aligned} &Mini\ (y_1^*,...,y_k^*,w,b,\xi_1,...\xi_n,\xi_1^*,...\xi_k^*) \\ &\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{k}\xi_i^* \\ &s.t: \forall_{i=1}^{n}: y_i[w\cdot x_i + b] \geq 1-\xi_i \quad \forall_{j=1}^{k}: y_j[w\cdot x_j^* + b] \geq 1-\xi_j^* \\ &\quad \forall_{i=1}^{n}: \xi_i \geq 0 \quad \forall_{j=1}^{k}: \xi_j^* \geq 0 \end{aligned} \tag{2}$$

where *C* and *C**are the parameters that can be adjusted by users. The *C***ξi**is called as an impact factor in the object function. The training process of TVSM is the same as the process resolving the above-mentioned optimization question.

The main idea of the training algorithm consists of several steps. Firstly the N labeled samples makes the initialized classifier. And secondly, the k unlabeled samples is labeled by the initial classifier. At last, in succession, the final classifier will be obtained by training both the N original labeled samples and k labeled samples.

When *k=1*, the training algorithm of TVSM needs to be adjusted apparently.

Theorem: In the algorithm of TSVM, when *k=1*, if the labeled samples are classified by using the initial classifier, the classified result makes sure that the object function is minimum in the formula (2).

Prove: Let *x** is classified by the initial classifier with label +1,so *x** is in the +1 region of the division plane. The correspond loose variable $\xi+^*>0$. If *x** were labeled as the value –1, the corresponding loose variable should be $\xi+^*>0$.

$$\begin{cases} [w\cdot x^* + b] = 1-\xi_+^* \\ -[w\cdot x^* + b] = 1-\xi_-^* \end{cases}$$

$$\xi_+^* + \xi_-^* = 2$$

Since *x** is in the region of the current division plane +1, and $\xi+^*<1$,

$\xi_-^* = 2-\xi_+^* > 1$

i.e. $\xi_+^* < \xi_-^*$ or $C^*\xi_+^* < C^*\xi_-^*$

That is to say, when x* is in the region of the current division plane +1, the value of formula (2) using the plus label will be less than the one using the

-1 labels. Therefore, the less value of the objective function will be obtained when using the classified result of the initial classifier.

The theorem indicates that the training process of Joachims's TSVM in case of $k=1$ is valid. Apply the formula of $k=1$ to Joachims's TSVM continuously, you can get the online tranductive vector machine. However, the number of n supporting the vector will be increased constantly too. It will do harm to both calculating and the dynamic performance of the support vector machine.

In order to reduce the degree of calculation complexity of online tranductive vector machine, and to improve the dynamic performance, the redresses to the number of support vector must be considered.

4. ESTIMATING AND SUPPORT VECTOR WINDOW ADJUSTING

The concept of support vector window is introduced to select proper support vector quantity in OTSVM. The $\xi\alpha$ estimator could resolve the problem of selecting the proper support vector window. The $\xi\alpha$ estimator method [6,7,8] is based on LOO and has considered the selection of parameters during the estimating process. This method largely reduces the workload of computation.

We can know from last section that two $\xi\alpha$ vectors are gained after SVM training. ξ is the vector of training losses at the solution of the SVM training problem. α is the solution of the SVM training problem. Let $d=\|\{i: (\alpha_i R\Delta 2 + \xi i) \geq 1\}\|$ record the number of vector which $(\alpha_i R\Delta 2 + \xi i) \geq 1$, where $R\Delta 2$ coud be 1. Then $\xi\alpha$ estimator will be

$$Err_{\xi\alpha}^{n}(h_L) = \frac{\|\{i:(\alpha_i R_\Delta^2 + \xi_i) \geq 1\}\|}{n} = \frac{d}{n} \tag{3}$$

$\xi\alpha$ estimate can be gained from every training. In terms of $\xi\alpha$ estimate you can judge the quality of SVM. In the same way you can compute the $\xi\alpha$ estimate value of SVM in different sample sets. The algorithm to choosing the size of the window in the sample set could be written briefly as following:

Let there are two vector sets $\{x0,x1,x2, \cdots xn\}$ and $\{x0,x1,x2, \cdots xn \cdots xm\}$, here m>n, Steps of the algorithm are:

step1: Input t two vector sets

step2: Calculate

$$Err_{\xi\alpha}^{n}(h_L) = \frac{\|\{i:(\alpha_i R_\Delta^2 + \xi_i) \geq 1\}\|}{n} = \frac{d}{n}$$

and

$$Err_{\xi\alpha}^{m}(h_L) = \frac{\left\|\{i:(\alpha_i R_\Delta^2 + \xi_i) \geq 1\}\right\|}{m} = \frac{d}{m}$$

step3: If *Errnξα (hL)*< *Errmξα (hL)* then *n* is the corresponding size of vector window.

else *m* is the corresponding size of vector window.

The above algorithm could be applied to the design of online tranductive SVM based on feedback.

5. THE DESIGN AND IMPLEMENTATION

Samples could only represent a little part and the past. So it is necessary to modify the SVM classifier while using the initial classifier. This is the idea of Online tranductive SVM(OTSVM) design based on feedback. Considering labeled samples are not usually acquired at random, and samples have been handled with some representative character. So there is no reason to take the distribution of its label as a basis to estimate the whole samples. And that if there are samples which are acquired at random, it is inaccurate to use that samples for computing because the small number of samples. In this section we will apply $\xi\alpha$ estimator to accomplish the design of dymanic SVM classifier based on feedback and the transductive inference learning [9,10,11]. In this OTSVM, the possibel labels for unlabeled samples are given by the initial SVM classifier. Then the the newer labelled samples are take part in the new training with initial labelled smples. In this process, OTSVM adjusts the division plane dynamically. The theorem certified in this article ensures the validity and rationality of this dynamic classifier, and the $\xi\alpha$ estimator shows the selection of window in samples rationalitily.

In the design of OTSVM we introduced *Wmin* and *Wmax* to ensure the range of support vector which is selected rationality.

After synthesizing the transductive inference learning and the $\xi\alpha$ estimator, the algorithm of OTSVM classifier would be :

Step1: select the labeled samples set *V0*, set the value of *Wmin* and *Wmax*

Step2: initial learning is done to the labeled samples set *V0*,. An original classifier *F0* and support vector *SV0*, will be gotten.

Step3: *W=number_of[SV0]*

Step4: for (i=1;i< numbers of samples tested; i++)

{ Step5: input the unlabelled tested sample *xi**

Step6: adopt classifiers *Fi-1* to classify unlabelled and tested sample *xi**, we will get the label value *yi**;

Step7: *Vi=SVi-1 U (xi*, yi*)*

Step8: calculate $\xi\alpha$ estimate and the value of support vector window

Step9: If $W \leq Wmin$ then $W = Wmin$;if $W \geq Wmax$ then $W = Wmax$

Step 10: adjust *SVi-1*, then we get *SVi.*

Step 11: reorganize the classifier *Fi-1*, and then we get a new classifier *Fi*}
Step 12: end

6. THE EXPERIMENTS AND CONCLUSION

We have adopted the data set of experiment from Hoovers sets [12]. The data gather over 10800 Web pages from the Internet.

6.1 Experiment 1

The purpose of the experiment is to test the effect of different numbers of initial samples to OTSVM and simultaneously to draw the effect of classification from comparing it with the traditional SVM.
In Hoover's data set, four groups of label samples are used as training sets to train initial OTSVM. The number of +1 labeled samples and -1 labeled samples are 30, 50, 80 and 100 in each group respectively. The samples were classified by hand in advance. And 160 random samples used to test set are classified by hand too. The classification result of 160 random samples will be a test criterion and attend to train and study in OTSVM.

Table1: the performance of OTSVM with different initial samples

	OTSVM		SVM	
group	accurate rate	recall rate	accurate rate	recall rate
1 (30)	68.3	88.6	65.7	84.7
2 (50)	82.8	89.4	71.8	87.8
3 (80)	86.2	95.7	78.4	92.8
4 (100)	87.4	94.7	79.9	91.1
average	81.2	92.1	73.9	89.1

Table 1 shows the result of experiment. The first group of training set contains 30 label samples, the second contains 50 labeled samples, the third contains 80 labeled samples and the fourth contains 100 labeled samples. The result of classification and training from the 160 test samples indicated that: more initial training sample sets may improve classification performance; the average accurate rate of OTSVM is higher than that of SVM; the function of improving classification effect of OTSVM is not obvious when the number of initial samples is large enough (the 3rd&4th group).

6.2 Experiment 2

The purpose of this experiment is to test the effect of different initial samples to OTSVM and simultaneously to observe the study progress process of OTSVM. There are four groups; each group has 80 labeled samples in training set (labeled by hand in advance). Selected 200 random samples are classified by hand too. The result of classification by hand will be a test criterion and will attend to train and study.

Table 2 shows the results of studying method of OTSVM in the condition of the same number of initial training sample set(the result of SVM is not listed due to the limited space). The result of classification from 200 testing samples indicates that the study capability of OTSVM will be increased in the process of classification and the accurate rate will be improved too.

Table2: the result on different sample set with same number of samples and the change of performance in OTSVM training *(AR= accurate rate; RR= recall rate)*

	OTSVM							
group	AR* 1-50	RR** 51-100	AR 51-100	RR 51-100	AR 101-150	RR 101-150	AR 151-200	RR 151-200
1	83.4	92.1	86.7	93.4	89.4	93.6	91.1	94.4
2	76.5	89.4	81.7	95.2	84.5	92.4	87.3	96.5
3	81.9	94.8	83.9	91.4	88.7	97.3	92.7	95.7
4	82.7	94.2	84.1	92.7	87.6	94.7	90.8	97.2
average	81.1	92.6	84.1	93.2	87.6	94.5	90.5	96.0

We should pay more attention to the second group of the experiments. The data shows that its accurate rate is obviously lower than common SVM in the testing of former 50 data(The common SVM accurate rate is 79.1% and 92.1%). By the deep analysis of the data, there are some data distribution departures, which will cause low accurate rate. Fortunately, the accurate rate will continuously increase along with the continuously studying and adjustment of OTSVM. The fact shows that OTSVM may reduce the departure of label sample distribution departures and offset of support vector set.

We can draw the conclusion from the above two experiments that OTSVM is based on training samples and gradually changed in the process of classification and studying. Every testing sample will affect the training of OTSVM, which comes from the corporate function of the classification process and training. Training will not be finished if the classification does not end. Obviously, this kind of method can adapt to the more general data distribution regulation and then it has the attribute of extending. We also can deduce it from the experiment that the excursion trend of classification plane

direct to the right orientation that improves the capability of OTSVM in the process of classification and training.

Although we use the ξα estimator, the amount of calculation of arithmetic is still very large. It may have some relatively convenient methods. By sorting αi, the method we chose supports vector in the experiment also lead to a good result. But it should be proved strictly in theory.

REFERENCES

[1] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

[2] Stitson MO, Weston JAE, Gammerman A, Vovk V, Vapnik V. Theory of support vector machines. Technical Report, CSD-TR-96-17, Computational Intelligence Group, Royal Holloway: University of London, 1996

[3] Cortes C, Vapnik V. Support vector networks. Machine Learning, 1995,20:273~297

[4] Vapnik V. Statistical Learning Theory. John Wiley and Sons, 1998

[5] Branson K. A naive Bayes classifier using transductive inference for text classification. 2001. http://www-cse.ucsd.edu/

[6] 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 Publishers, 1999. 200~209

[7] Chen YS, Wang GP, Dong SH. A progressive transductive inference algorithm based on support vector machine. Journal of Software, 2003,14(3): 451~460

[8] Klinkenberg. Ralf Using Labeled and Unlabeled Data to Learn Drifting Concepts. http://www-ai.cs.uni-dortmund.de/ (2001)

[9] Ralf Klinkenberg ,Thorsten Joachims. Detecting concept drift with support vector machines. Proceedings of the Seventeenth International Conference on Machine Learning (ICML),San Francisco, 2000. Morgan Kaufmann

[10] James Allan. Incremental relevance feedback for information filtering. In H. P. Frei, editor, Proceedings of the Nineteenth ACM Conference on Research and Development in Information Retrieval, pages 270–278, New

[11] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell,T. M. (2000). Text classi_cation from labeled and unlabeled documents using EM. Machine Learning,39, 103-134

[12] www.hoovers.com