An Approach to Incremental SVM Learning Algorithm

Rong Xiao Jicheng Wang Fuyan Zhang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, P. R. China 210093
Cloud@graphics.nju.edu.cn

Abstract

The Classification algorithm based on Support Vector Machine (SVM) now attracts more attentions due to its perfect theoretical properties and good empirical results. In this paper, we first analyze the properties of SV set thoroughly, then introduce a new learning method, which extends the SVM Classification algorithm to incremental learning area. The theoretical bases of this algorithm are the classification equivalence of the SV set and the training set. In this algorithm, the knowledge is accumulated in the process of incremental learning. In addition, unimportant samples are discarded optimally by LRU scheme. The theoretical analysis and experimental results show that this algorithm could not only speedup the training process, but also reduce the storage cost, while the classification precision is also guaranteed.

1. Introduction

Now, the incremental data classification method has been one of the key technologies of intelligent knowledge discovering. Compared with the traditional data classification strategies, incremental data classification has two advantages: 1) It can reduce the storage cost by discarding some old samples; 2) The utilization of historical training results makes the successive learning faster.

Most incremental learning algorithms available are based on decision-tree or neural network [1, 2, 3, 4]. These algorithms are short of expected risk controlling mechanism over the whole sample space, so they always tend to over-fit in the training set. Furthermore, they couldn't discard samples optimally, so sometimes many useful samples are discarded in the successive training. SVM is a new and promising pattern recognition technique, which is developed by Dr. Vapnik and his research group [5]. It is one of the few algorithms that can solve the problem of over-fitting successfully. But traditional SVM algorithm doesn't support incremental learning. Therefore, it is important to extend SVM algorithm to incremental learning area.

The paper is organized as follows. The next section, Section 2, reviews the SVM theory and analyzes the distribution properties of the training set thoroughly. In Section 3, a new incremental learning method $\alpha-ISVM$ is presented. And the empirical result and discussion are given in Section 4 and Section 5.

2. SVM Theory

Before the discussion of the new incremental SVM classification algorithm $\alpha-ISVM$, we briefly review the SVM theory, then analyze the properties of SV. This is the theoretical basis of this paper.

2.1. Introduction of SVM

SVM theory is mainly derived from the problem of binary classification. Its main idea can be concluded as the following two points: First, it constructs a nonlinear kernel function to present an inner product of feature space, which corresponds to mapping the data from the input space into a possibly high-dimensional feature space by a nonlinear algorithm. Thus it is possible to analyze the nonlinear properties of samples in the feature space with linear algorithm. Secondly, it implements the structural risk minimization principle in statistical learning theory [6] by generalizing optimal hyper-plane with maximum margin between the two classes. Although intuitively simple, this idea actually plays the role of capacity controlling and makes the learned machine not only has small empirical risks, but also has good generalization performance. Therefore, SVM has many advantages in both theoretical base and practical prospect.

2.2. Introduction and Analysis of SV

Suppose we have the following training set:

$$(x_1, y_1),...,(x_l, y_l),\ y_i \in \{-1,1\}\ \ x_i \in R^n \qquad (1)$$

Vapnik has proved that the training result of SVM only depends on a small set of samples, which is so-called SV, and the normal of the Optimal hyper-plane w can be

1082-3409/00 $10.00 © 2000 IEEE

Authorized licensed use limited to: Pt Dwarka Prasad Mishra Indian Inst of Info Tech. Downloaded on April 16,2021 at 05:38:34 UTC from IEEE Xplore. Restrictions apply.

expressed as the linear combination of the SVs:

$$w = \sum_{SV} a_i y_i x_i, \quad 0 < a_i \le C \tag{2}$$

Therefore, the SV set can fully describe the classification characters of the whole training set. And the classification equivalence between the SV set and the training set can also be proved. Usually, the SV set is only a small portion of the training set. Thereby if we train the SVM on the SV set instead of the whole training set, the training time can be reduced greatly without much classification precision lost.

Based on the analysis of experimental results, we divide the samples in the SV set into two classes: The first class is the SVs with $a_i = C$, which is so called **BSV** (*Boundary Support Vector*), corresponding to the vectors which can not be classified correctly. Another class is the SVs with $0 < a_i < C$, which is so called **NSV** (*Normal Support Vector*).

These two kinds of SV describe different classification properties of the training set. BSV is mainly affected by noisy data when the training set is very large. And when the training set is not large enough, it describes the details about the boundary of the training set. As for NSV, it represents the classification properties of most training samples. These two kinds of SVs determine the result of the classifier together. Experiments also show that these characters do not always remain the same. When the successive training samples are introduced, BSV, NSV, and normal samples can inter-transform. For example, due to the limitation of the initial training set, the samples that represent the main classification information may be miss-classified as BSV. As the accumulation of knowledge about the classification in the successive incremental training, the classification contribution of these samples will be made clear. At the same time, these BSVs are gradually degraded to NSVs or normal samples.

3. An Incremental SVM algorithm α–*ISVM*

As we can see above, the equivalence between the SV set and training set is broken when the new samples are introduced into the training set in the procedure of the incremental learning. Thus, on one hand the original SV set can not fully describe the classification properties of the new training set. On the other hand, with the introduction of the new samples, some old samples must be discarded optimally to reduce the storage cost. Therefore, following two major problems will be met in the incremental SVM learning algorithm: 1) How to construct the new SV set from the initial training set; 2) how to discard old samples optimally.

3.1. The construction of the new classifier

The problem of constructing an incremental SVM classifier can be formally described as:

- Presupposition: Let A_0 be a history data set, $A_i (i = 1,2,...,n)$ be incremental sample sets, and $\forall i \neq j, A_i \cap A_j = \varnothing$. Let Ω^t be the SVM classifier of the data set $E_t = A_0 \cup ... \cup A_t$;
- Problem: How to find the optimal SVM classifier Ω^{t+1} on the sample set $E_{t+1} = E_t \cup A_{t+1}$ based on the classifier $\Omega^t (t = 0,1,...,n-1)$.

Traditional SVM algorithms discard all previous training results and re-train the new classifier on the whole data set in the case of incremental learning. This method is so slow that it cumbers the application of SVM learning algorithm greatly in the area of incremental learning. In this section, we describe a new scheme, which uses an iterative method to find out the optimal convergent result on the whole training set. This scheme can be described as follows: First, the old classifier is tested on the new incremental sample set. Those samples classified incorrectly is combined with the current SV set to construct a new training set, and the rest samples form a new test set. Then a new SVM classifier is trained on the new training set, and it can use the new test set to repeat previous operations. The process continues until all data points are classified correctly.

3.2. Accumulation of the sample's distribution knowledge

Using above learning algorithm, we can utilize the old training result, and accelerate the successive learning procedure. But all history data still must be kept, and this limitation not only reduces the training speed, but also increases the storage cost. The key idea to solve this problem is to discard history samples optimally. To achieve this goal, we analyze the characters of the accumulation about the distribution knowledge during the incremental training.

We start from the simple case which is shown in Figure 1, where $X^0 = \{A_t | t = 1,2,...,15\}$ denotes the initial training set, $X^1 = \{A_t | t = 16,17,...,23\}$ denotes the incremental set, and $X^0 \cup X^1$ denotes the incremental union set. First, we apply the SVM classification algorithm to the union of the initial and the incremental set. Then, we get the SV set $X^0_{sv} = \{A_t | t=8,9,10,11,12,13,14,15\}$ of the initial training set, and SV set $X^1_{sv} = X^0_{sv} \cup \{A_t | t=5,18,21,22,23\}$ of the union set. Thus, it can

Authorized licensed use limited to: Pt Dwarka Prasad Mishra Indian Inst of Info Tech. Downloaded on April 16,2021 at 05:38:34 UTC from IEEE Xplore. Restrictions apply.

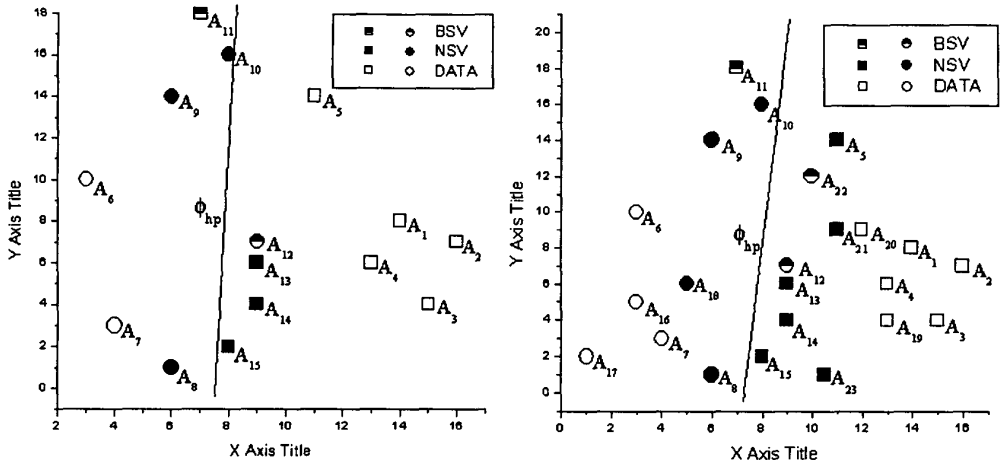

Figure 1. The classification result of the initial training set (left) and the incremental training set (right).

be intuitively concluded from figure 1 that the accumulation about the distribution knowledge has following characters:

- *The distribution properties of the samples affect the classification result directly.* For example, the set $\{A_t \mid t=1,2,3,4,6,7\}$, which is far from the separating hyper-plane, does not affect the classification result. But the point A_5, which belongs to the SV set X_{sv}^1 and is near to the separating hyper-plane, affects the classification result in a certain degree.
- As the training samples accumulate in the process of incremental learning, the description of the samples' distribution knowledge is more and more accurate. *Thus the new gained distribution knowledge is superior to the old ones.*

According to the analysis above, we define the following three concepts:

- The sample that is never selected into any SV set is called intra-sample, and the corresponding sample set is regarded as *backup set.* Generally, most samples are intra-samples.
- The sample which frequently appears in SV set is called vice-boundary sample, and the corresponding sample set is regarded as *caching set.*
- The sample in the latest SV set is called boundary sample, and the latest SV set is regarded as *working set.*

From the intuitive view of these three types of samples, the boundary sample represents the main classification knowledge of the sample space, and vice-boundary sample adds some detail information to the classification knowledge. Both of these two kinds of sample are very important to the accumulation of the samples' distribution knowledge. As to the intra-samples, they are trivial to the finial classification.

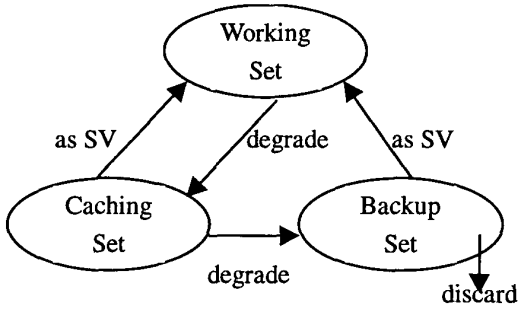

Figure 2. The relationship between the three classes

3.3. Incremental SVM Learning Algorithm $\alpha-ISVM$

Based on above analysis, our improvements are concentrated on the following points: First, the intra-samples are gradually discarded at a certain ratio, through which the training set and the storage cost of the old sample are reduced. Secondly, some vice-boundary samples are optimally introduced into the training set to accelerate the convergence of the final SV set searching procedure.

Because of the different contributions to the final classifier, the characters of these three kinds of sample sets can be used to optimize the training procedure. First, as the samples in the backup set do not affect the result of the finial classifier, they can be discarded gradually using the LRU scheme, and the remaining samples can be used as the samples of the test set to validate the new trained classifier. Secondly, as the samples in the working set are actively involved in the computation of the final classifier, it can be directly introduced into the training set of the new classifier to accelerate the training procedure. Thirdly, as the samples in the caching set affect the previous

Authorized licensed use limited to: Pt Dwarka Prasad Mishra Indian Inst of Info Tech. Downloaded on April 16,2021 at 05:38:34 UTC from IEEE Xplore. Restrictions apply.

classification result, their contribution to the final classifier is between the backup set and working set. So they can be selected into the training set of the new classifier according to a certain scheme and priority. Finally, as the incremental training proceeds, the samples' distribution knowledge accumulates gradually, and some samples' contributions to the classifier also change. Therefore, these samples can be moved from one set to another (See Figure 2).

Therefore, we present the following incremental SVM learning algorithm $\alpha-ISVM$:

The algorithm is controlled by 5 parameters: $(\alpha,\beta,\gamma,\Delta d,\delta)$, $\alpha\in[0,1]$, $\beta\in[0,1]$, $\Upsilon\in[0.5,1)$, $\Delta d>0$, $\delta\in(0,\Delta d)$. And all weights of these new samples are assumed to be zero.

1) The initializing procedure

First, a SVM classifier Ω^0 is trained on the initial sample set X^0. And its corresponding SV set X^0_{sv} and non-SV set X^0_{nsv} are preserved。Then, each sample in the SV set X^0_{sv} is assigned a value Δd. At last, the caching set C^0 is set to empty, and let X^0_{nsv} be the backup set B^0.

2) Incremental training procedure

First, the new added sample set X^{i+1} is defined to be the test set V_0^{i+1} of this training turn. Then let the working set W^i of last turn be W_0^{i+1}, let the caching set C^i of last turn be C_0^{i+1}, let the backup set B^i of last turn be B_0^{i+1}, and let the classifier Ω^i of last turn be the initial classifier Ω_0^{i+1} of this turn.

Based on above definition, an iterative searching scheme is employed to find the optimal classifier Ω^{i+1}, and the j-th searching procedure can be described as: $(j=1,2,...)$

- First, the classifier Ω_{j-1}^{i+1} is validated with the test set V_{j-1}^{i+1}. Based on the result of the validation, the set V_{j-1}^{i+1} can be divided into two classes: the set E_j^{i+1} for all of the classification errors, and the set O_j^{i+1} for all of the samples that have passed the validation. If $E_j^{i+1}\neq\varnothing$, the following steps are proceeded.
- Secondly, with the aid of the weight and β, the caching set C_{j-1}^{i+1} can be divided into two sub set $S(C_{j-1}^{i+1})$ and $NS(C_{j-1}^{i+1})$. Let $S(C_{j-1}^{i+1})\bigcup W_{j-1}^{i+1}\bigcup E_{j-1}^{i+1}$ be the new training set T_j^{i+1}, and let $NS(C_{j-1}^{i+1})\bigcup O_{j-1}^{i+1}\bigcup B_{j-1}^{i+1}$ be the new test set V_j^{i+1}. Then the traditional SVM learning algorithm is applied to the set T_j^{i+1}, the result classifier Ω_j^{i+1} and corresponding SV set $SV(T_j^{i+1})$ are preserved. Let the set $SV(T_j^{i+1})$ be the working set W_j^{i+1}, and let the set $W_{j-1}^{i+1}\bigcup C_{j-1}^{i+1}-W_j^{i+1}$ be the new caching set C_j^{i+1}. Thereafter the weight of each sample of the above-mentioned sets is adjusted using the following rules:

$$\begin{cases}\psi_s=\gamma\psi_s & \forall x_s\in C_j^{i+1}\\ \psi_s=\psi_s+\Delta d & \forall x_s\in W_j^{i+1}\end{cases}\qquad(3)$$

- Finally, a sub set $SubC$, which is defined as:

$$SubC=\{x_s\mid\forall x_s\in SubC\quad\&\psi_s\le\delta\}\qquad(4)$$

is discarded from the caching set C_j^{i+1}. It means:

$$C_j^{i+1}=C_j^{i+1}-SubC.\qquad(5)$$

Let the new backup set:

$$B_j^{i+1}=(T_j^{i+1}\bigcup V_j^{i+1})-(W_j^{i+1}\bigcup C_j^{i+1})\qquad(6)$$

By iteratively repeating the above three steps, a series of classifiers $\Omega_0^{i+1},\Omega_1^{i+1},...$ are obtained. Certain classifier Ω_t^{i+1}, whose corresponding validation precision on the test set V_t^{i+1} is 100%, will be found. It means $E_t^{i+1}=\varnothing$. At this time, the classifier Ω_t^{i+1} will be considered as the optimal classifier Ω^{i+1}. At the end of the searching procedure, some old samples are discarded with ratio α from the backup set B_j^{i+1} using the LRU scheme. Thus it is possible to reduce the storage burden by forgetting the samples that affected the classification least recently.

4. Empirical results and Discussion

Based on the research above, we have implemented the incremental algorithm $\alpha-ISVM$ on a text library, and a discussion is made at the end of this chapter.

4.1. Experiments and Results

The text library contains 1493 articles (each article is represented by a vector with 10000 features). All articles are assigned with 2 labels manually, among which 317 articles are selected as test set, the others are selected to be training set. And the text automatic classification application is realized based on the following environments: Matlab5.3 / VC6.0 / WINNT (software), PII333 / 256MB memory (hardware).

The experiments are designed as follows: First, half

Authorized licensed use limited to: Pt Dwarka Prasad Mishra Indian Inst of Info Tech. Downloaded on April 16,2021 at 05:38:34 UTC from IEEE Xplore. Restrictions apply.

Table 1. The comparison between the $\alpha - ISVM$ algorithm and the tradition algorithm $\beta = 0.5$, using 1 level polynomial SVM machine

Training Set	Samples	Experiment A		Experiment B		Experiment C		Experiment D	
		Time (s)	PREC	Time (s)	PREC	Time (s)	PREC	Time (s)	PREC
Initial	588	106.4	91.3%	106.4	91.3%	106.4	91.3%	106.4	91.3%
Inc. 1	246	77.1	92.7%	76.9	92.7%	80.4	92.7%	247	92.7%
Inc. 2	79	83.7	92.9%	87.4	92.9%	89.2	93.1%	379.5	93.1%
Inc. 3	187	75.4	93.6%	92.5	93.7%	95.1	93.9%	491.4	93.9%
Inc. 4	42	88.3	93.8%	90.5	93.8%	97.4	94.3%	582.7	94.3%
Inc. 5	34	96.5	93.5%	98.3	93.7%	109.3	94.2%	613.8	94.2%

Table 2. The effect of the discard factor α

Experiments	Discard Factor	Initial	Incremental Training Set				
			1	2	3	4	5
C	$\alpha = 0$	588	834	913	1100	1142	1176
B	$\alpha = 0.5$	588	753	532	586	591	589
A	$\alpha = 0.9$	588	689	231	283	285	279

of the 1176 articles are selected as the initial training set, another half are divided into 5 groups randomly. Then under different values of factor α, three tests are carried out (experiment A corresponds to $\alpha = 0.9$, experiment B corresponds to $\alpha = 0.5$, and experiment C corresponds to $\alpha = 0$). At last, experiment D, which using traditional algorithm, is made. The experiment results are given in the above tables.

4.2. Discussion

As we can see above, when α increases, the storage cost of old samples is reduced greatly, the precision drops slightly, and the training procedure is accelerated. Compared to experiment D, the first two experiments need much less storage space, and all of the first three experiments are trained faster than the last one while the classification precision is also guaranteed. The theoretical and experimental analysis suggests that the parameter α provide a way to control the classification precision. Theoretical analysis also shows the parameter β controls the searching procedure. A well selected β will make the searching procedure tends to converge quickly, and the training time of each searching set will be slowed down.

As to the computation complexity, we have the following result: suppose the dimension of the input space is d_l, the number of training points is L, the number of the samples in the SV set is N_{sv}. Generally, $N_{sv} / L \ll 1$, and most samples in the SV set is not BSV. Then the computation complexity of the SVM classifier is $O(N_{sv}^3 + LN_{sv}^2 + d_l LN_{sv})$ [7]. In the case of incremental learning ($\alpha = 1, \beta = 0$), suppose the incremental set has $\gamma L (\gamma > 0)$ samples, and the resulting classifier has $N_{sv}^{'}$ SVs. In most cases, we will have $N_{sv}^{'} / N_{sv} \approx 1$, and the computation complexity of this learning algorithm $O(N_{sv}^3 + d_l N_{sv}^2)$ is much less than that of the traditional algorithm $O(N_{sv}^3 + L(1+\gamma)(N_{sv}^2 + d_l N_{sv}))$.

5. Conclusion

Based on the thorough analysis on the properties of SV set, we discuss the incremental SVM learning algorithm. Further more, an incremental algorithm $\alpha - ISVM$ based on the discard factor α is presented. The algorithm fully utilizes the properties of SV set, and accumulates the distribution knowledge of the sample space through the adjustable parameters. Thus it is possible to discard samples optimally, at the same time, the training time is saved greatly. Experiments prove that this algorithm could not only improve the training speed, but also reduce the storage cost, while the classification precision is also guaranteed. The further research will mainly focus on the discarding scheme of the noisy samples.

References

[1] J. Ratsaby. "Incremental learning with sample queries", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, Aug. 1998. pp. 883 -888

[2] K. Yamauchi, N. Yamaguchi and N. Ishii. "Incremental

Authorized licensed use limited to: Pt Dwarka Prasad Mishra Indian Inst of Info Tech. Downloaded on April 16,2021 at 05:38:34 UTC from IEEE Xplore. Restrictions apply.

learning methods with retrieving of interfered patterns", IEEE Transactions on Neural Networks, Vol. 10, Nov. 1999. pp. 1351 -1365

[3] Wang, E.H.-C, and A. Kuh. "A smart algorithm for incremental learning", International Joint Conference on Neural Networks, Vol. 3, 1992. pp. 121 -126

[4] M. Veloso and D. Borrajo. "Learning strategy knowledge incrementally", Proceedings of 6th International Conference on Tools with Artificial Intelligence, 1994. pp. 484 -490

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

[6] V. Vapnik. Statistical Learning Theory. Wiley, New York, 1998.

[8] Christopher J.C.Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, Boston, 1998

Authorized licensed use limited to: Pt Dwarka Prasad Mishra Indian Inst of Info Tech. Downloaded on April 16,2021 at 05:38:34 UTC from IEEE Xplore. Restrictions apply.