Fast Incremental SVM Learning Algorithm based on Center Convex Vector

Dongying BAI[1]

Air and Missile Defense College,
Air Force Engineering University
Xi'an 710051, China
e-mail: choubao2012@hotmail.com

Ci ZHANG[3]

Unit 94402, Jinan 250002, China

Jun HAN[2]

Air Traffic Control and Navigation College,
Air Force Engineering University
Xi'an 710051, China
e-mail: coverkillyou@tom.com

Abstract—A fast SVM learning algorithm is proposed according to incremental learning and center convex hull operator. It is established on analyzing the relevance of support vector and convex hull from the angle of calculation geometry. The convex hull of current training samples is solved in the first place. Further, Euclidean distance elimination is applied to convex hull. Meanwhile, every time when the incremental learning is going on, the training samples should contain samples violated KKT condition in previous sample set, experiment results indicate that the algorithm effectively shortens training time while classification accuracy keep a satisfied level.

Keywords-support vector machine; incremental learning; convex hull operator;

I. INTRODUCTION

SVM(Support Vector Machine)[1] is a kind of statistical learning theory based learning algorithm, which is mainly used to solve supervised learning problems. In practice, training sample often has the characteristics of online increased, which requires classifier to retrain new samples constantly, thus, classification efficiency is greatly reduced. Incremental learning aims at obtaining the optimum solution of original training samples and the new sample set. Hence, for growing data set, Incremental Support Vector Machine (ISVM)has its unique advantage.

The study of ISVM algorithm mainly focused on how to guarantee the classification accuracy and improve learning speed. Literature[2][3][4][5]didn't analyze the impact on SV(Support Vector) set when the new samples were added, which could cause information lost of existing data set. Literature[6][7]analyzed relationship of KKT condition and sample in SVM optimization problem, considered influence on existing sample set distribution with new sample, improved learning accuracy. But samples which had little impact on subsequent training were not eliminated effectively, thus, memory is large and training time is long. Literature[8][9][10]enhanced training speed by selecting the original samples, how the original samples were reduced to obtain SV which could best represent classification information was still a thorny problem.

From above analysis, effective elimination has impact on classification accuracy and speed to some extent. On the basis of analyzing relationship of convex hull vector and support vector in computation geometry[11], a rapid SVM learning algorithm (CQSVM) is proposed using center convex hull algorithm and incremental learning. The algorithm combines solving convex hull algorithm with Euclidean center distance[12] for sample reduction, and every time the incremental training samples should contain the last training sample set against the generalized KKT condition of samples. The experiment shows the algorithm not only improves the incremental SVM learning speed, but also keeps high classification accuracy.

II. SVM CLASSIFICAITON PRINCIPLE AND GEOMETRIC RELATIONSHIP BETWEEN SV AND CONVEX HULL

A. SVM Classification Principle

SVM[1]is derived from linear optimal hyper plane, which could be described as following constrained optimization problem:

$$\min : \frac{1}{2}\|w\|^2 \qquad (1)$$

s.t. $y_i[(w \cdot x_i)+b]-1 \geq 0, (i=1,2,\ldots,n)$

Defining Lagrange function：

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w \cdot x_i + b)-1] \qquad (2)$$

Original problem is converted to following dual problem of convex quadratic programming under primary constraint conditions:

$$\max : W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \qquad (3)$$

$$s.t. \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$\alpha_i \geq 0, (i=1,2..,n)$$

Above equation is actually a quadratic function problem under constraint of inequality, it has only one optimal solution α_i^*, so

$$w^* = \sum_{i=1}^{n}\alpha_i^* y_i x_i \qquad (4)$$

When α_i^* is not zero, the acquired sample is SV. Hence, weighted coefficient vector is linear combination of SV in optimal hyper plane, b^* could be solved by constraint conditions, through that, optimal classification function can be obtained:

$$f(x) = \text{sgn}((w^*)^T x + b^*) = \text{sgn}(\sum_{i=1}^{n}\alpha_i^* y_i x_i^* x + b^*) \qquad (5)$$

B. Geometric Relationship between SV and Convex Hull

Literature[13]pointed out that, to find out the optimal hyper plane of maximal classification boundary between two classes equals seeking two nearest neighbor points in corresponding convex hulls of two classes. The fig. 1 shows geometric meaning of optimal hyper plane for two classes linear separability samples.

Under the circumstance of linearly separable, the location of SVM optimal hyper plane is decided on SV, in order to ensure that optimal hyper plane could maximize minimum interval, SV could only be generated from sample vectors which located in boundaries of two classes sample set.
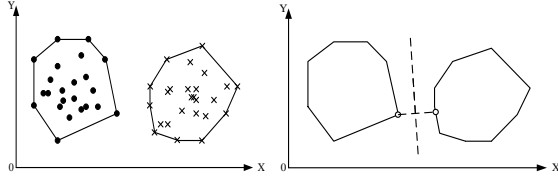

Fig.1 Geometric meaning of optimal hyper plane for two classes linearly separable samples

Convex hull refers to the edge sample of each kind of training set from the angle of geometric meaning. The SVM algorithm aims at maximizing the minimum distance between two convex sets. SV must appear on the edge of sample set. That's means SV is produced on and around the edge of the sample set. When the SV is determined, the position of optimal hyper plane has been determined.

III. KKT CONDITION AND ANALYSIS OF SV CHANGE

Karush-Kuhn-Tucher condition[1](abbreviation KKT) plays an important role in constrained optimization. According to KKT conditions, at optimal point, the product of Lagrange multiplier and constraint is zero. Linear discriminant function is then given:

$$g(x) = (\omega^* \cdot x) + b^* = \sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^* \qquad (6)$$

So each sample x meeting optimal problem KKT conditions that optimal solution $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_l]$ makes in dual problem are:

$$\alpha_i = 0 \Rightarrow y_i g(x_i) > 1 \qquad (7)$$
$$0 < \alpha_i < c \Rightarrow y_i g(x_i) = 1 \qquad (8)$$
$$\alpha_i = c \Rightarrow y_i g(x_i) < 1 \qquad (9)$$

Among them, α_i that didn't equal zero are SV. Considering function system $g(x) = h$, we can infer $g(x) = 0$ is classification plane, $g(x) = \pm 1$ is boundary of classification interval, samples on the boundary are SV.

Theorem 1 After the new sample's arrival, part of new samples which violate KKT conditions will transform to samples which satisfy KKT conditions even the SV; part of new samples which satisfy KKT conditions will transform to SV; part of non SV in original sample set which satisfy KKT condition will transform to SV; part of previous SV will transform to non SV.

The above theorem shows that if only new arriving samples and SV set in original sample set are taken into consideration, while non SV set in original sample set are ignored, important information could be lost. Therefore, every time of incremental training, training samples should contain samples which violate general KKT conditions in last training sample set.

IV. SVM ALGORITHM BASED ON CENTER CONVEX HULL AND INCREMENTAL LEARNING

A. Algorithm Principle

In order to prevent some training samples which contain classification information from eliminating quickly, meanwhile, considering that original SV set may be influenced by new arriving sample set. In the new incremental learning process, current training sample set should be included not only new samples but also samples which violate general KKT condition in last training sample set. At the same time, training samples will be furtherly reduced, the convex hull vector in two class boundary then are extracted using Euclidean center distance elimination for training sample set in ISVM training.

B. Symbol Explanation

Setting X_0 as the first training sample set; X_k is the kth incremental learning sample; I_k is new arriving sample in the kth incremental learning, $Qhull\{\}$ is convex hull operator; Ω_0 is training classifier for $Qhull\{X_0\}$; $CQhull\{\}$ is center convex hull operator, which means that, calculating distance d_i of x_k's geometric center and convex hull vector, selecting threshold r, eliminating samples $d_i > r$; Ω_0 is the final classifier.

C. Algorithm steps

- Training classifier Ω_0 with $Qhull\{X_0\}$.
- When the kth incremental learning is conducting ($k = 1, 2 \cdots$)

  a) Testing the incremental process is going on or not, if it doesn't continue, turn to step c), otherwise, executing step b)

  b) Solving center convex hull vector $CQhull\{X_{k-1}\}$ of X_{k-1}

  c) Calculating sample X_{k-1}^V which violates KKT condition in X_{k-1}

  d) Setting $X_k = I_k + CQhull\{X_{k-1}\} + X_{k-1}^V$

  e) Training X_k to get classifier Ω_k

- Termination, Ω_k outputs.

V. EXPERIMENT OUTCOME AND ANALYSIS

A. Experiment data and parameter setting

Experiment environment is Intel Core 2.5GHz, memory of PC is 2GB, Matlab7.0 platform, using libsvm-mat-2.91-1 toolbox. In order to verify validity of the algorithm, two

different classes data are selected from UCI database for test. (1) Breast-cancer-wisconsin data set, a total of 699 samples, the sample dimension is 11,the number of training samples is 300,test samples number is 399,the sample classification number is 2.(2)Artificial data set, each component in data set obeys Gauss distribution, the number of training samples is 1000,700 test samples are selected from 1000 samples, test samples number is 800,the dimension of positive samples is 5,sample classification number is 2,Euclidean distance of two classes center is 4.

B.   Relative contrasting algorithms

1) Method IncHulSVM0( ): Only with the convex hull algorithm to reduct the training sample set；every time when incremental learning is conducting, solving classifier with convex hull vector computed this time and new arriving samples.

2) Method IncHulSVM1( ): Only using convex hull algorithm to reduct training sample set; every time when incremental learning is conducting, solving classifier with convex hull vector computed this time, samples that violate general KKT condition in last training sample set and new adding samples.

3) Method IncHulCntr0( ): Using center convex hull algorithm for sample reduction; every time when incremental learning is conducting, solving classifier only with center convex hull vector computed this time and new adding samples.

4) Method Standard incremental SVM

C.   Experiment outcome and analysis

Fig.2 and fig.3 are experiment outcome in data set Breast-cancer-wisconsin. Fig.4 and fig.5 are test results based on artificial data set. RBF kernel function is used to determine training time in the experiment, while classification accuracy is the average with different kernel function（Linear、Rbf）.
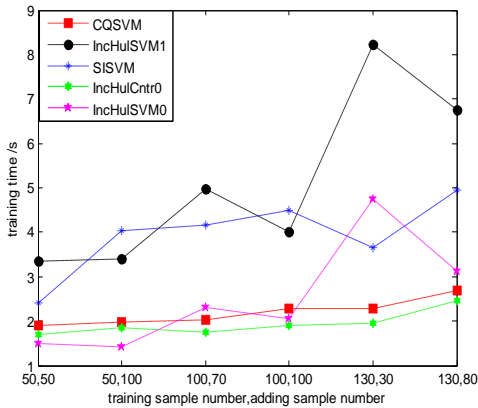

Fig.2 Training time contrast of various algorithms under Breast-cancer data set
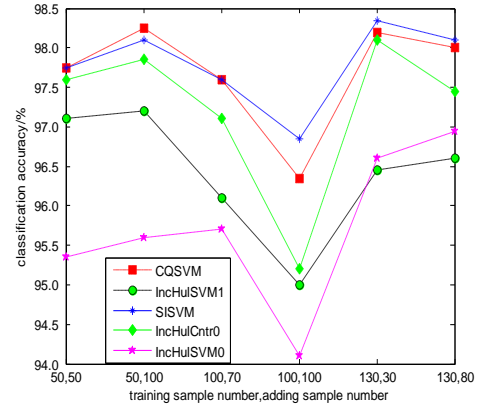

Fig. 3 Learning accuracy contrast figure of various algorithms under Breast-cancer data set
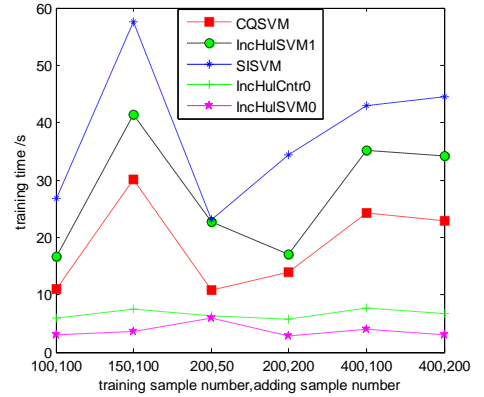

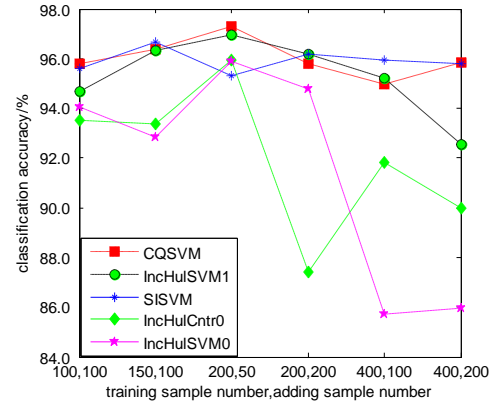Fig.4 Training time contrast of various algorithms under Artificial data set

Fig. 5 Learning accuracy contrast figure of various algorithms under Artificial data set

From training time, for the data set Breast-cancer-wisconsin, training sample number is 300.As can be seen from fig.2,training time is short for IncHulSVM0 、IncHulCntr0 and CQSVM, the cost time of standard incremental SVM and IncHulSVM1 are comparatively long, due to training sample set scale is not large, training time

overlaps to some degree for above methods, which method has the fast speed could not be determined .But for artificial data set, training sample number is 700,the data scale is larger. As is shown in fig.4,among these methods, time cost sequence from less to more is : IncHulSVM0, IncHulCntr0, CQSVM, IncHulSVM1,standard incremental SVM. From classification accuracy, as can be seen in fig.3 and fig.5, no matter we use Breast-cancer-wisconsin or artificial data set, CQSVM and standard incremental SVM has the highest classification accuracy, accuracy is similar between them. Method IncHulSVM0 costs the least time, but classification accuracy is the lowest. Methods IncHulCntr0 and IncHulSVM1 reflects approximate performance, both of them are better than IncHulSVM0，worse than CQSVM and standard incremental SVM.

The core idea of CQSVM algorithm is to boost incremental SVM learning speed, meanwhile keeping classification accuracy as far as possible. From learning speed, after reduction of training sample using convex hull algorithm, simplifying training sample with center distance ratio to enhance learning speed; from classification accuracy, for every time incremental learning, training samples should contain last training samples set which violate general KKT condition, it reflects rational use of new samples while previous classification information is not missed, so classification accuracy could be ensured not to drop as far as possible.

VI. CONCLUSION

The paper puts an emphasis on analyzing influence on SV set brought out by new adding sample under the KKT condition, proposes a new SVM algorithm based on center convex hull algorithm and incremental learning, it combines convex hull algorithm and center distance elimination, meanwhile, for every time incremental learning, training samples contain last training samples set which violate general KKT condition. The experiment shows that the algorithm can greatly improve the incremental SVM learning speed and maintain higher classification accuracy at the same time.

REFERENCES

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

[2] Syed N,Liu H,Sung K, "Incremental learning with support vector machines, " Proceedings of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence (IJCAI299), Stockholm, Sweden: Morgan Kaufmann,1999, pp. 876-892

[3] B Wen , G L Shan and X S Duan, " Research of Incremental Learning Algorithm Based on Drive Error Criterion , " Computation Technology and Automation,vol.31, 2012, pp. 100-103

[4] C M Wu, X D Wang and D Y Bai, "Fast SVM incremental learning algorithm based on between-class convex hull vector,"Computer Engineering and Application, vol.46, 2010, pp.185-187

[5] Ralaivola L, "Incremental Support Vector Machine Learning:A Local Approach,"Proc.of ICANN,Vienna,Austria:Springer, 2001, pp.322-330.

[6] Q Mou, Y K Chen and X R Bi, "Intrusion Detection Method Based on Fast Incremental SVM,"Computer Engineering, vol. 38,2012, pp. 92-94

[7] Hyunsoo Kim,HaesunPark, " Incremental and Decremental Least Squares Support Vector Machine and Its Application to Drug Design," Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference,2004

[8] Y Wang, J A Yang and H Liu, "A SVM incremental learning algorithm based on inner hull vectors," JOURNAL OF CIRCUITS AND SYSTEMS,vol.16,2011, pp.109-113

[9] GertC,Tomaso P, " Incremental and Decremental Support Vector Machine Learning," Adances in Neural Information Processing Systems(NIPS*2000),Cambridge, MA:MIT Press,2001.

[10] F S Shen,J Y Zhang and K J Wang, "SVM Incremental Learning Using Simulated Cutting Algorithm," PR& AI,vol.23,2010, pp.491-500

[11] Pei de Zhou, " Computational geometry−Algorithms Analysis and Design,"Beijing,Tsinghua University press,2000

[12] Zhang L,Zhou WD and Jiao LC, "Pre-extracting Support vectors for support vector machine ,"Proceeding of ICSP2000, pp.1432-1435

[13] Sergios Theodorids,Konstantinos Koutroumbas, " Pattern Recognition, Third Edition,"Beijing,China Machine Press,2006, pp.110-116