International Journal of Computational Intelligence Systems and Applications. Volume 4, Issue 1, (2016) 8-16
http://www.cayley-nielson.com/ijcisa/

Research on Semi-Definite Programming Support Vector Machines

Hou Jingzhong[1,2], Xia Kewen[1,2],Yang Fan[1,2],Cui Dongyan[1,2]

[1] School of Electronics & Information Engineering, Hebei University of Technology, Tianjin, China

[2] Tianjin Key Laboratory of Electronic Materials and Devices, Tianjin, China

[2]Corresponding author: Xia Kewen

E-mail:[1]*houjingzhong2002@163.com,*[2]*kwxia@hebut.edu.cn,*[3]*yangfan@hebut.edu.cn,* [4]*cdy_xzx@163.com*

Abstract

The superior performance in pattern recognition [1] and classification have been shown in the Support Vector Machine algorithm. The Support Vector Machine have been wildly used in many different fields. However, the Support Vector Machine has it’s own defect, such as the computing efficiency decreases when Support Vector Machine in the face of higher dimensional data, in other hands, an unique Kernel function and it’s parameter must been selected at first to make sure that the Support Vector Machine running properly. In fact, the kernel function and it’s parameter selection is the key of the Support Vector Machine algorithm.Based on study on working principle and algorithm steps of the traditional support vector machine, aim to solve the parameter selection of the SVM. Research the SDP-SVM algorithm, design the SDP-SVM model, elaborate the steps of the algorithm. Then optimize the kernel function using semidefinite programming method. This method can be used to find the best parameter of the SVM. Finally simulate the heart_scale data in the UCI (University of California Irvine) database with the SDP-SVM model. After compare the experiment results of the SDP-SVM and the traditional SVM, it shows that the generalization capability and classification accuracy of the SDP-SVM algorithm do have been improved greatly.

Keywords: Support Vector Machines，Semi-Definite Programming，Kernel Function，Strip-Steel Defect Image.

1. Introduction

Support vector machines (SVM)[2-4] is a kind of algorithms which can overcome the inherent defects of traditional machine learning algorithms based on empirical risk minimization. SVM showing with better generalization ability than traditional machine learning method. SVM vector can be mapped to high-dimensional space, then establish a maximum interval hyperplane. The total classification error is minimum when the distance from points to plane is maximized. So as to deal with data and recognize with the combination of geometry theory. When nonlinear problems is encountered, in order to solve complex calculation problems of high-dimensional vector, high-dimensional feature space can mapped into linearly separable space using SVM algorithm, that is kernel function[5] mapping method.

However, there also exist inherent flaws in the support vector machine (SVM): 1) the operation cost and operation time of SVM are greatly increased when faced with high order characteristic matrix, it is difficult to use the traditional SVM on large-scale training sample; 2) the traditional SVM algorithm using binary classification algorithms, but in the practical application multi classification problems are often

ISSN 2331-1924 / Copyright © 2016 CAYLEY NIELSON PRESS, INC.
April, 2016

encountered, generally multiple two class SVM is combined to solve the problem[6];3) when using the SVM algorithm a specific kernel function and it’s parameter must be selected, but choosing the optimal parameter has been an important problem of SVM.

It is obvious from the present research that the key to solve the problem of SVM classification is how to select the parameter, and the key in the parameter selection is how to select the model. So it is urgent need to find a model making the SVM parameter selection more effective and more intelligent. Due to the SVM is widely used, continue to study and optimize the model of SVM algorithm is of profound significance. Considering the inherent flaws in the SVM model and the advantages of semi-definite programming (SDP) method, it can improve the efficiency of searching for the optimal combination if the semi-definite programming method is introduced into the SVM algorithm model. Continuously distinguish parameter effectiveness ,update the kernel matrix and optimize the parameter of kernel function through the semi-definite programming combination coefficient, so as to improve the classification accuracy and generalization ability of SVM model. This method effectively solve the parameter optimal selection problem of SVM, and reduces the size of training samples. The potential significance of researching the algorithm is profound.

To solve this problem, using semi-definite programming method to find the optimal combination coefficient of SVM, distinguish the validity of coefficients, improve the performance of SVM. Semidefinite programming support vector machines (SVM) is applied to heartscale database classification experiments, then detect and classify the heartscale data.

2. Traditional Support Vector Machine

2.1 SVM principle

SVM theory is based on the structural risk minimization principle[7], appropriately select the subset function and it’s discriminant function, to ensure that still maintain the classifier error is smaller under the condition of limited training samples.

SVM is a kind of effective tools used to solve classification problems in actual application, while in practice the classification problem most non-linear,

Therefore, it is necessary to mapping the nonlinear classification problem into a linear problem in high dimensional space by means of the transformation, and then find the optimal separating hyperplane through the linear classification method[8]. Therefore, when solving the problem of nonlinear classification, first design a nonlinear to linear mapping: $x \to \varphi(x)$; then defined for the Lagrangian dual as follows:

$$L(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}a_i\left\{y_i\left[\left(w^T x\right)+b\right]-1\right\} \qquad (1)$$

Where $a_i \ge 0$ is the Lagrange coefficient for each sample. Seeking the minimum value of formula (1) under the constraints conditions:

$$\sum_{i=1}^{n} y_i a_i = 0 \qquad (2)$$

$$a_i \geq 0, (i = 1,2,...,n) \tag{3}$$

Seeking a_i when the value is maximum:

$$Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j \varphi(x_i)\cdot\varphi(y_i) \tag{4}$$

If there is a function K can make $K(x_i, y_j) = \varphi(x_i)\cdot\varphi(y_j)$, then equation (4) becomes:

$$Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j K(x_i, y_j) \tag{5}$$

A support vector machine is obtained by using the optimal classification plane $f(x) = \operatorname{sgn}\left\{\sum_{x_i \in SV} a_i^* y_i K(x_i, x) + b^*\right\}$.

2.2 SVM Algorithm Steps

Before the introduction of the SVM algorithm steps, we first discuss some parameters in SVM.

(1) The penalty factor C

In the case of linear inseparable, seek the optimal classification surface is equivalent to solve the problem:

$$\text{Minimize } \phi(w,\xi) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right) \tag{6}$$

$$\text{The constraint is: } y_i\left[\left(w^T x_i\right) + b\right] - 1 + \xi_i \geq 0, (\xi_i \geq 0, i = 1,2,...,n) \tag{7}$$

Converted by Lagrangian method:

$$\text{Maximize } Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j \varphi(x_i)\cdot\varphi(y_i) \tag{8}$$

$$\text{Constraint is: } \sum_{i=1}^{n} y_i a_i = 0, \quad a_i \geq 0, (i = 1,2,...,n) \tag{9}$$

Since C is inversely proportional to a_i, so C is also inversely proportional to $\|w\|$. The smaller the C is, the greater the classification interval is, the stronger the generalization ability of SVM is.

As a result, SVM algorithm steps[9-11] in practical application are:（I）obtain learning samples $(x_i, y_i), i = 1,2,...,n$;（II）select the appropriate penalty factor C, and further transform the problem into a quadratic programming problem;（III）use the algorithm to optimize the problem;（IV）obtain SVM related parameters a , a^* , b ;（V）classify and predict the test set data.

2.3 Simulate the traditional SMV

The binary data set heart_scale which integrated in the libsvm toolbox is selected in this experiment[12,13], this data set is actually from the UCI (University of California Irvine) database. The

UCI database has been widely used in machine learning benchmark test, which is recognized one of the standard library in the field of machine learning. The heart_scale standard data set contains a total of 270 data samples, each data sample contains 13 latitude. The front 150 samples as a training set and the other 120 samples as a test set to classify. Simulate the data set with the traditional SVM algorithm, classification result shown in Fig.1, the ROC curve shown in Fig.2.

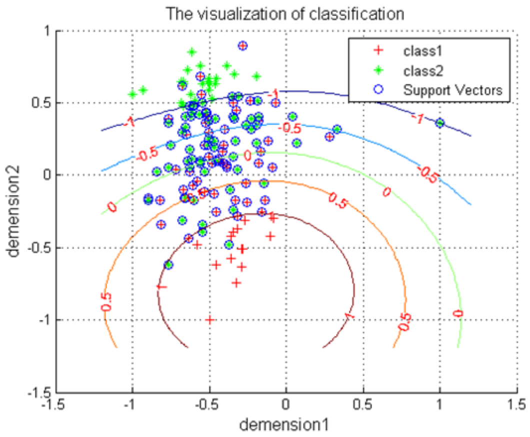

Fig1 Classification result of traditional SVM

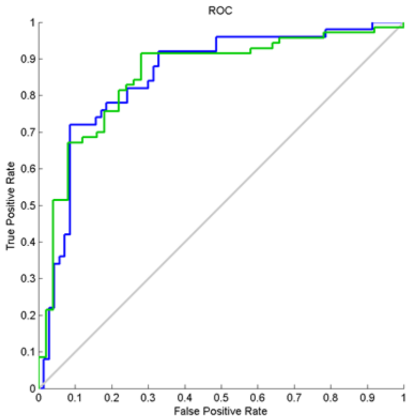

Fig2 Classification ROC of traditional SVM

In the classification result figure of traditional SVM, the plane can only show two attributes of a sample, so when it comes to multidimensional attribute classification only two of these attributes of the samples will be chosen to drawing. The horizontal axis and vertical axis in Fig.1 are the fifth and the eighth attributes inside 13 attributes of the heart_scale data set.The red cross and green asterisk in Fig.1 represent two categories, the blue circle represent support vector.

Zero calibration in the figure is the optimal separating hyperplane,and the samples within a distance of 1 line before and after the zero calibration are support vectors. It is these discrete samples support the middle optimal zero calibration. The green lines and blue lines in Fig.2 respectively represent category 1 and category 2 of the data set. It can be seen from the figure that the ROC curve of the traditional support vector machine algorithm is close to the upper left corner, its classification performance is relatively high.

3 Semi-definite Programming Support Vector Machine(SDP-SVM)

3.1 SDP method principle

The Semi-definite Programming（SDP）[14-16] is a kind of convex optimization problem which makes the linear function maximal or minimal, the constraint is that affine combination of symmetric matrix is positive semi-definite.

The basic model of SDP:

$$\begin{aligned} &\min C \bullet X \\ &s.t.\ A_i \bullet X = b_i \quad i = 1,\ldots,m \\ &X \geq 0 \end{aligned} \qquad (10)$$

Here b_i is a real number, C and A_i are n order real matrix.

Make $AX = \begin{pmatrix} A_1 \bullet X \\ A_2 \bullet X \\ \vdots \\ A_n \bullet X \end{pmatrix}$， $b = (b_1, b_2, ..., b_m)^T$, then the SDP model is converted to:

$$\begin{aligned} &\min C \bullet X \\ &s.t.\ AX = b \\ &X \geq 0 \end{aligned} \qquad (11)$$

Considering the optimization problem, the SDP model is converted to the following form:

$$\begin{aligned} &\min \quad C^T \cdot X \\ &s.t. \quad F(x) \geq 0 \end{aligned} \qquad (12)$$

Here, $F(x) \geq 0$ represent the semi-definite programming of $F(x)$, that is $F(x) \geq 0$ is a linear matrix inequality(LMI), the condition is $z \in R^n, z^T F(x) z \geq 0$. The commonly used model of $F(x) \geq 0$ is as follows:

$$\begin{aligned} &\min m \\ &s.t. \begin{bmatrix} mI & A(x) \\ A(x)^T & mI \end{bmatrix} \geq 0 \\ &X \in R^k \ and \ m \in R \end{aligned} \qquad (13)$$

In the use of Lagrange the formula (13) generate semidefinite programming duality model:

$$\begin{aligned} &\max \quad b^T y \\ &s.t. \quad A^T y + Z = C \\ &\qquad\quad Z \geq 0, y \in R^m \end{aligned} \qquad (14)$$

The optimal conditions for the semi-definite programming are (11) and (14) have strictly feasible solutions, X is the optimal solution of formula (11), if and only if $(X, Z) \in S^n \times S^n$ make the following formula holds:

$$\begin{cases} AX = b, \qquad X > 0 \\ A^t y + Z = C, \ Z > 0 \\ XZ = 0 \end{cases} \qquad (15)$$

3.2 SDP-SVM Model Design

On the classifier parameters setting, better parameter selection can not only maximize the classification accuracy of SVM, but also can maintain the good generalization ability. Not only that, in order to simplify SVM operation, you can use the SDP method for selecting parameters to optimize SVM. In some references discussion, using SDP to optimize SVM kernel function is proposed. On this basis, we put forward SDP-SVM algorithm, an improved SVM.

Taking into account a binary classification problem, $\{x_i, y_i\}_{i=1}^n$ represent data set, $y_i \in \{-1,+1\}$ represent classification set. For the SDP-SVM method, it is converted into the following model:

$$\begin{aligned} \min_{\mathrm{w},b,\xi} \quad & \frac{\|w\|_2^2}{2} + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y_i\left(\mathrm{w}^T\phi(x_i)+b\right) \geq 1-\xi_i \\ & \xi_i \geq 0 \end{aligned} \qquad (16)$$

The training set x_i is mapped to a high dimensional space by the function ϕ. C means the penalty factor of the number of error classification. translate it into the dual problem:

$$\begin{aligned} \max_{a} \quad & -\frac{1}{2}\sum_{i,j=1}^{N} y_i y_j K(\mathrm{x}_i, \mathrm{x}_j) a_i a_j + \sum_{j=1}^{N} a_j \\ s.t. \quad & \sum_{i=1}^{N} a_i y_i = 0 \\ & 0 \leq a_i \leq C, \forall i \end{aligned} \qquad (17)$$

The purpose of $K(x_i, y_i) = \phi(x_i)^T \phi(x_j)$ is to get the maximum value of a from the formula (17).

$$y(x) = f(x,a) = sign\left(\sum_i a_i y_i K(x_i, x) + b\right) \qquad (18)$$

In order to introduce the SDP model to the kernel function optimization problems, as one of the most commonly used means of optimization, kernel function calibration play an important role between the two kernel or between kernel function and the objective function. An unclassified test data $S = \{x_i\}_{i=1}^n,\ x_i \in R^m$ is given, the inner product of two nuclear matrix is defined:

$$\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^{n} k_1(x_i, x_j) k_2(x_i, x_j) \qquad (19)$$

Here, K_i is the kernel function of sample S. The kernel calibration of the predicted sample S defined as follows:

$$\hat{A}(S, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \qquad (20)$$

When the classified sample label $y_i = \pm 1,\ i = 1,\ldots,n$ is unknown, $K_2 = yy^T$ in the condition of $y_i = y_j,\ k_2(x_i, x_j) = -1,\ y_i \neq y_j$, the target kernel function $k(x_i, x_j) = 1$. And the nuclear calibration can be expressed by closeness degree formula of sample S between the ideal kernel function and the selected kernel function K :

$$\hat{A}\left(S, K_1, yy^T\right) = \frac{\left\langle K_1, yy^T \right\rangle}{\sqrt{\left\langle K_1, K_1 \right\rangle \left\langle yy^T, yy^T \right\rangle}} = \frac{\left\langle K_1, yy^T \right\rangle}{m\sqrt{\left\langle K_1, K_1 \right\rangle}} \qquad (21)$$

The greater the value $\hat{A}\left(S, K_1, yy^T\right)$ is, the more closer to the optimal kernel function, based on the theory and the definition of the kernel calibration above, use SDP to optimize the kernel function and classification set, the kernel function can be modified using SDP method.

Assuming a sample set S, specify multiple kernels function $k_1, ..., k_2$, respectively calculate the kernel matrix of k_i on the sample set S:

$$K_i = \begin{bmatrix} k_i(x_1, x_1) & k_i(x_1, x_2) & \cdots & k_i(x_1, x_n) \\ \vdots & \vdots & & \vdots \\ k_i(x_n, x_1) & k_i(x_n, x_2) & \cdots & k_i(x_n, x_n) \end{bmatrix} \qquad (22)$$

Linear combine K_i, the combination coefficient μ_i is obtained, which is between 0 and 1, and the bigger the combination coefficient is the better the kernel function is.

The combined kernel matrix K_i is introduced into the SDP model:

$$\begin{aligned} &\omega(K_i) = \max \quad e^T a - \frac{1}{2} a^T G(K_i) a \\ &s.t. \qquad y^T a = 0 \\ &\qquad\qquad C - a \geq 0 \\ &\qquad\qquad a \geq 0 \\ &\qquad\qquad 0 < \mu_i < 1 \\ &\qquad\qquad \textstyle\sum_{i=1}^{l} \mu_i = 1 \end{aligned} \qquad (23)$$

Here, $G(K_i) = diag(y) \cdot K_i \cdot diag(y)$, minimized $\omega(K_i)$ and introduce variable t, to make $t \geq \omega(K_i)$, so at this time the model, when meeting the constraints, is converted to calculate the minimum value of t.

3.3 Simulation Experiment

The experimental simulation is still using the heart_scale data set, and all the experimental parameters are consistent with the previous chapter. The model of support vector machine after semi-definite programming is called SDP-SVM, classification result shown in Fig.3, the ROC curve shown in Fig.4.

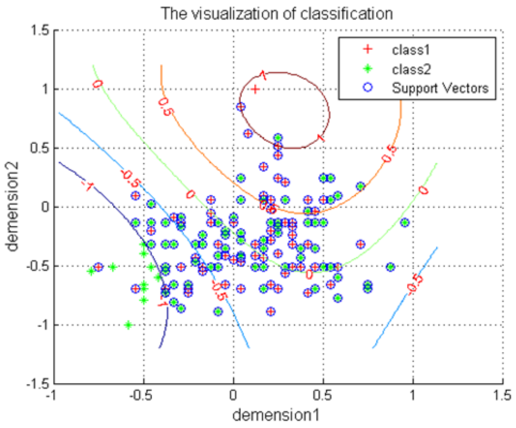

Fig3 Classification result of SDP- SVM

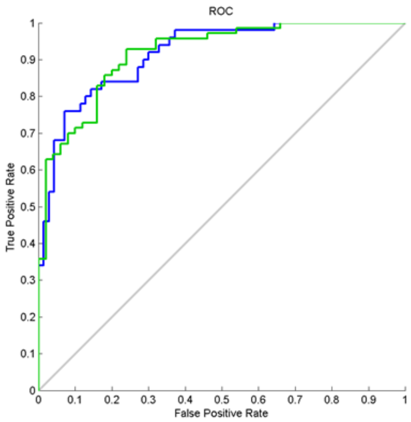

Fig4 Classification ROC of SDP-SVM

The description of the classification result is similar to previous chapter, so not details here. The zero calibration in the figure is the optimal classification hyperplane, two selected properties is 1 and 4.

3.4 Comparative Analysis

It can be seen from table 1 that the SVM kernel optimization combination by SDP method helps to to improve the SVM classification accuracy and obtain better training model, and can achieve higher classification prediction effect. SDP-SVM method can get better classification effect compared with the traditional SVM method.

Table 1 Classification effect comparison between traditional SVM and SDP-SVM

model	accuracy	AUC	time	weight
traditional SVM	88.3333%	0.8534	0.361200	1：2：7
SDP-SVM	95.1667%	0.9094	0.280314	5：4：1

4 Conclusion

SVM itself has great advantages in classification and identification, but the disadvantage is also clear, when make full use of the known information the amount of calculation will be increased, and there has not been an effective unified approach on the kernel function selection. Operating parameter selection of kernel function is the key factor to influence the generalization ability of SVM classifier.

Using semi-definite programming to determine the effectiveness of operating parameters of the kernel function, which has a good mathematical theory foundation, the essence of the SVM model is an optimal combination nuclear model. The SVM with SDP optimize the kernel function make it's kernel optimization better, improved it's classification accuracy and generalization ability.

The simulation experiment result shows the advantages of the SDP-SVM compared to the traditional SVM. It is obvious that the proposed method in the paper is effective and feasible, and the classification accuracy of the obtained SDP-SVM model is higher than that of the traditional SVM.

Acknowledgement：

This work was supported by the National Natural Science Foundation of China (No.51208168), Hebei Province Natural Science Foundation (No.E2016202341) and Tianjin Natural Science Foundation (No.13JCYBJC37700).

References

[1] R.P.W. Duin. Four scientific approaches to pattern recognition. Fourth Quinquennial Review 1996-2001. Dutch Society for Pattern Recognition and Image Processing. NVPHBV, Delft, 2001:331 – 337.

[2] Pal Sankar K.A Probabilistic Active Support Vector Learning Algorithm [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26(3): 413-418.

[3] Bennett K, Blue J. A support vector machine approach to decision trees[R]. Rensselaer Polytechnic Institute, Troy, NY: R. P. I Math Report , 1997 :97-100.

[4] D.M.J. Tax and R.P.W. Duin. Support vector data description. Machine Learning, 2004, 54(1):45 – 56

[5] Haasdonk B.Feature space interpretation of SVMs with indefinite kernels [J]. Pattern Analysis and Machine Intelligence,2005,27(4):482-492.

[6] I.M. de Diego, J.M. Moguerza, and A. Munoz. Combining kernel information for support vector classification. In Multiple Classifier Systems, pages 102 – 111. Springer-Verlag, 2004.

[7] O.Chapelle, V Vapnik, O Bousquet, et al. Choosing multiple Parameters for support vector machines[J]. Machine Learning, 2002, 46(l):131-159

[8] Krebel Ulrich H G. Pairwise classification and support vector machines [A] School kopf Bernhard(edi.). Advances in Kernel Methods: Sup2, port Vector Learning[C]. Massachusetts, The MIT Press , 1999 :255-268.

[9] John Shawe-Taylor, Shiliang Sun. A review of optimization methodologies in support vector machines[J]. Neurocomputing, 2011, 74(10):3609-3618.

[10] S. Lessmann, R. Stahlbock. Genetic Algorithm for Support Vector Machine Model Selection[C]. International Joint Conference on Neural Networks, 2006

[11] R Stoean, M Preuss, C Stoean, et al. Concerning the Potential of evolutionary support vector machines[C]. Evolutionary Computation CEC2007 IEEE Congress on, 22-28 Sept 2007: 1436-1443.

[12] V.S. Cherkassky and F. Mulier. Learning from data: Concepts, Theory and Methods. John Wiley & Sons, Inc., New York, NY, USA, 1998.

[13] Helena G. Ramos, Tiago Rocha, Jakub Král, Dário Pasadas, et al. An SVM approach with electromagnetic methods to assess metal plate thickness[J]. Measurement, 2014,54(8):201-206.

[14] Wei Wu, Jianyun Nie, Guanglai Gao. An Improved SVM-based Multiple Features Fusion Method for Image Annotation[J]. Journal of Information and Computational Science,2014,11(14):4987-4997.

[15] Ling Jian, Zhonghang Xia, Xijun Liang, Chuanhou Gao. Design of a multiple kernel learning algorithm for LS-SVM by convex programming[J]. Neural Networks, 2011,24(7):476-483.

[16] Domenico Conforti, Rosita Guido. Kernel based support vector machine via semidefinite programming: Application to medical diagnosis[J]. Computers & Operations Research, 2010,37(8):1389-1394.