International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 11 (2014), pp. 181-188
http://dx.doi.org/10.14257/ijmue.2014.9.11.18

Application of Improved Grid Search Algorithm on SVM for Classification of Tumor Gene

Li Wenwen[1,2], Xing Xiaoxue[1,3], Liu Fu[1*] and Zhang Yu[1]

[1]College of Communications Engineering, Jilin University, Changchun 130022, China
[2]College of Mechanical Engineering, Baicheng Normal University, Baicheng 137000, China
[3]Collegeof Information Engineering, Changchun University, Changchun 130022, China
liwenwen1017@126.com(LI Wenwen)
** liufu@jlu.edu.cn(LIU Fu)*

Abstract

According to the merits and shortcomings of the traditional gridsearch algorithm in parameters optimization of support vector machine (SVM), an improved grid search algorithm is proposed. Dichotomous search algorithm is used to reduce target searching range. First, searching range is determined roughly, and a set of parameters are obtained. Then fine search is applied in reduction the range for searching, and searching the optimum parameters.Three kinds of famous tumor gene data set are used in the comparison experiments to validate the classification accuracy of principal component analysis (PCA)-SVM and kernel principal component analysis (KPCA)-SVM. Experiment results and data analysis shows that, comparing with traditional gridsearch algorithm, the proposed method has higher classification accuracy and less search time.

Keywords: *Artificial intelligence; Gridsearch; Support vector machine; Tumor gene data set; Principal component analysis*

1. Introduction

Support vector machine (SVM) method have great advantages in solving high-dimension and nonlinear pattern recognition problem of small sample, such as strong adaptability, short training time, global optimum and strong generalization ability. And SVM has become the hotspot of study, especially the wide use in identification of tumor genetic [1]. Extended performance of the SVM depends on the parameter of kernel function in some extent, and when the kernel function is determined, we just need to determine the penalty factor parameter C. Therefore, selection of kernel parameters and penalty factor will not only affect the classification accuracy and learning ability, but also affect the performance of SVM. Grid-search algorithm is one of the most common methods for parameter optimization. For the classification of tumor gene, the PCA is used to decrease the dimensions of data and reduce the amount of sample information firstly, and then the sample is trained by SVM. Finally, the grid-search algorithm is used to determine the optimizing parameters of SVM.

The advantage of grid-search algorithm is that it can search the independent parameters in parallel, and consume less time when limited parameters are available. On the other hand, the

[*] Corresponding Author

ISSN: 1975-0080 IJMUE

disadvantages of grid-search are large calculation, especially when more parameter needed and difficulty to estimate the range of parameters and penalty factor [2]. This paper presents an improved grid search method of SVM parameters optimization, and its application in the classification of tumor gene data set. The method improves the accuracy of classification, and can also reduce the time of the parameter optimization.

2. SVM

Support vector machine (SVM) is a kind of novel machine learning methods based on statistical learning theory. And it based on VC dimension theory [3] and the structural risk minimization principle [4] to find the optimal separation hyperplane. SVM can directly weighs the balance among learning accuracy and learning ability of the training sample and has good generalization ability.

Support vector machine (SVM) was originally developed from the linearly separable problem, and many jobs are related to two types of problems. The set of training sample are $\{\mathbf{x_i}, i = 1, 2, \ldots, n\}$, and every sample belongs to a class in two class of ω_1 and ω_2, so the linear discriminant function can be expressed as

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 \quad (1)$$

And the decision rules is shown as the following formula:

$$\mathbf{w}^T \mathbf{x} + w_0 \begin{cases} > 0 \\ < 0 \end{cases} \Rightarrow \mathbf{x} \in \begin{cases} \omega_1, & y_i = +1 \\ \omega_2, & y_i = -1 \end{cases} (2)$$

Solving the most optimal hyperplane problem is converted into the optimization problem with inequality constraints [5]:

$$\begin{aligned} &\min_{\mathbf{w}, w_0} \frac{1}{2} \|\mathbf{w}\|^2 \\ &s.t. \quad y_i \left(\mathbf{w}^T \mathbf{x}_i + w_0 \right) \geq 1, i = 1, 2, \ldots. n \end{aligned} \quad (3)$$

Because the training samples of classification error wouldadversely affect classifier, It is necessary to introduce penalty factor C in order to reduce the $\sum_{i=1}^{n} \xi_i$ as much as possible.So the optimization problem is transformed to following formula:

$$\begin{aligned} &\min_{\mathbf{w}, w_0, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \\ &s.t. \quad y_i \left(\mathbf{w}^T \mathbf{x}_i + w_0 \right) \geq 1 - \xi_i \qquad i = 1, 2 \ldots, n \; (4) \\ &\qquad \xi_i \geq 0, i = 1, 2, \ldots, n \end{aligned}$$

Where C is adjustable parameter, the value of C will effect thepenalty factors of the error classification samples directly, and value of C is proportional tothepenalty factors.

In practical problem, the key point of non-linear SVM is selection of kernel function and searching of the optimal parameters, which is the most widely used method. Original feature space of training samples are mapped into a high dimensional feature space by using a nonlinear algebraic mapping, and the transformed high dimensional sample set can be classified by linear classification. Therefore characteristics of the input sample can be transformed via nonlinear transformation into the space which can be solved by linear method.

ϕ is a non-linear transformation functions, and we use a kernel function K to replace the dotproduct of feature vector after the non-linear transformation,

So samples X_i is related to X_j as follows:

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right)=\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) \quad (5)$$

The resulting discrimination function is shown as follows:

$$g\left(\mathbf{x}\right)=\sum_{i\in sv}\alpha_i y_i K\left(\mathbf{x}_i,\mathbf{x}\right)+w_0 \quad (6)$$

In nonlinear classify algorithms, different SVM model can be derived from different kernel functions.

The radialbasisfunction (RBF) has the advantages of little error and extensive applicable range. In this paper, RBF function is used as kernel function in experiments[6]. The classifiers are nonlinear due to the nonlinear kernel function, and this method is also termed as C-SVM, i.e. standard Support vector machine[7].

3. The Improved Grid Search Algorithm

The basic principle of grid search is as follws: select the search area of the punishment factor C and kernel parameter σ firstly [8]. Then set C and σ step length, and search the parameter according to the step length, thus the 2-d plane grid is formed, and each intersection node on the grid corresponds to a set of (C,σ) values. Usually, in order to try to get the global optimal solution of parameters, search range of the parameters is very large, so it takes a lot of time. Figure 1 shows the flow diagram of the model of the method based on grid search. The improved idea in this paper is to confirm a rough parameter search area firstly, and select a larger step length, so as to get a set of (C,σ), on this basis, select a precise search area to get the (C,σ) in the end [9]. The steps of optimizing the parameter based on improved method of grid search are described as follows:

1) Rough searching: set the searching range of penalty factor C and kernel parameter σ as $[a_0,b_0]$ and $[c_0,d_0]$ respectively, for example, $[2^{-10},2^{10}]$ or $[2^{-15},2^{15}]$ and the initial step-size is set to 8. Then the most excellent parameters $(BestC_0, Best\sigma_0)$ can be found using traditional methods of grid search, with which cross validation can get the highest accuracy. In order to overcome the problems of extra-learning and weak normalization capability, we choose the (C,σ) corresponding to the minimum value of penalty factor C as the best values, if the accuracy of cross validation from severalgroupsofparameters are equal. And if the minimum value of C corresponds to several σ, we need to choose the closest (C,σ) in order to obtain the preliminary values of $BestC_0$ and $Best\sigma_0$.

2) Fine searching: the searching range of C is set as $[a_0,(a_0+b_0)/2]$,if $BestC_0 \le (a_0+b_0)/2$, otherwise $[(a_0+b_0)/2,b_0]$. And the searching range of σ is set as $[c_0,(c_0+d_0)/2]$, if $Best\sigma_0 \le (c_0+d_0)/2$, otherwise $[(c_0+d_0)/2,d_0]$. After the comparisonofparameters, assume the searching range of C and σ are $[a_1,b_1]$ and $[c_1,d_1]$ respectively, then the searching step size is set as 2 and the traditional grid search method is used to find the the most excellent parameters $(BestC_1, Best\sigma_1)$ by following the method

describedin step1. So the finalpunishment factor C and kernel parameter σ are $BestC = BestC_1$ and $Best\sigma = Best\sigma_1$ respectively.

For most dataset, the optimalparameters of SVM can be obtained by convergent iterative procedure.Inrarecases, convergencyvalue is the approximateoptimalsolution, but it is a solution much closer toglobaloptimal one [10].

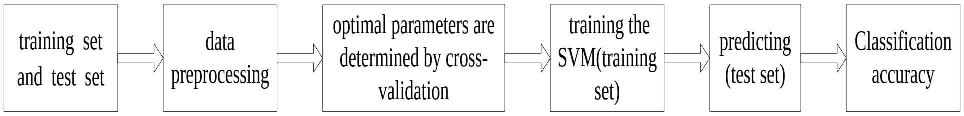

training set and test set
data preprocessing
optimal parameters are determined by cross-validation
training the SVM(training set)
predicting (test set)
Classification accuracy

Figure 1. Flow Chart of C-SVM Model based on the Grid Search

4. Experimental Results

This paper studies the binary-class of SVM, and it has good scalability in high dimension and small sample. So it is of great significance to use the method in classification of tumor gene. This paper selected three sets of data, which are famous public resources and open to society. A contrast experiment research was made between the SVM based on PCA and the SVM based on KPCA [11]. And we analyzed the classification accuracy of the two methods based on the improved grid search for selecting the optimal parameters. The datasets we selected are as follows.

Singh-2002

Tissue samples of Singh-2002 have102 tissue samples consisting of 12600 genes, and the data sets make up of 52 cancerous prostate(PR)tissues and 50 normal prostate(N)tissues. We choose 26 trained samples and 26 test samples in the cancerous prostate tissues, and 25 trained samples and 25 test samples in the normal prostate tissues. There are 339 genes remaining for experiment after processing under the condition of missing partial data.

Armstrong-2002-v1

The data sets have72 tissue samples consisting of12582 genes, and Armstrong-2002-v1 make up of 24 acute lymphoblastic leukemia(ALL) and 48 acutemyeloidleukemia(AML). We choose 12 trained sample sand 12 test samples in the acute lymphoblastic leukemia, and 24 trained samples and 24 tests amples in the acutemyeloidleukemia. There are 1081 genes remaining for experiment after processing under the condition of missing partial data.

Lahio-2007

The data sets have37 colon tissue samples consisting of 22883 genes, and Lahio-2007 make up of 8 serrated colorectal cancer(Serrated CRC) and 29 conventional colorectal cancer(conventional CRC) We choose 4 trainedsamplesand 4 testsamples in the serrated colorectal cancer, and 15 trainedsamplesand 14 testsamples in the conventional colorectal cancer. There are 2202 genes remaining for experiment after processingunder the conditionofmissing partialdata and noisepoints.

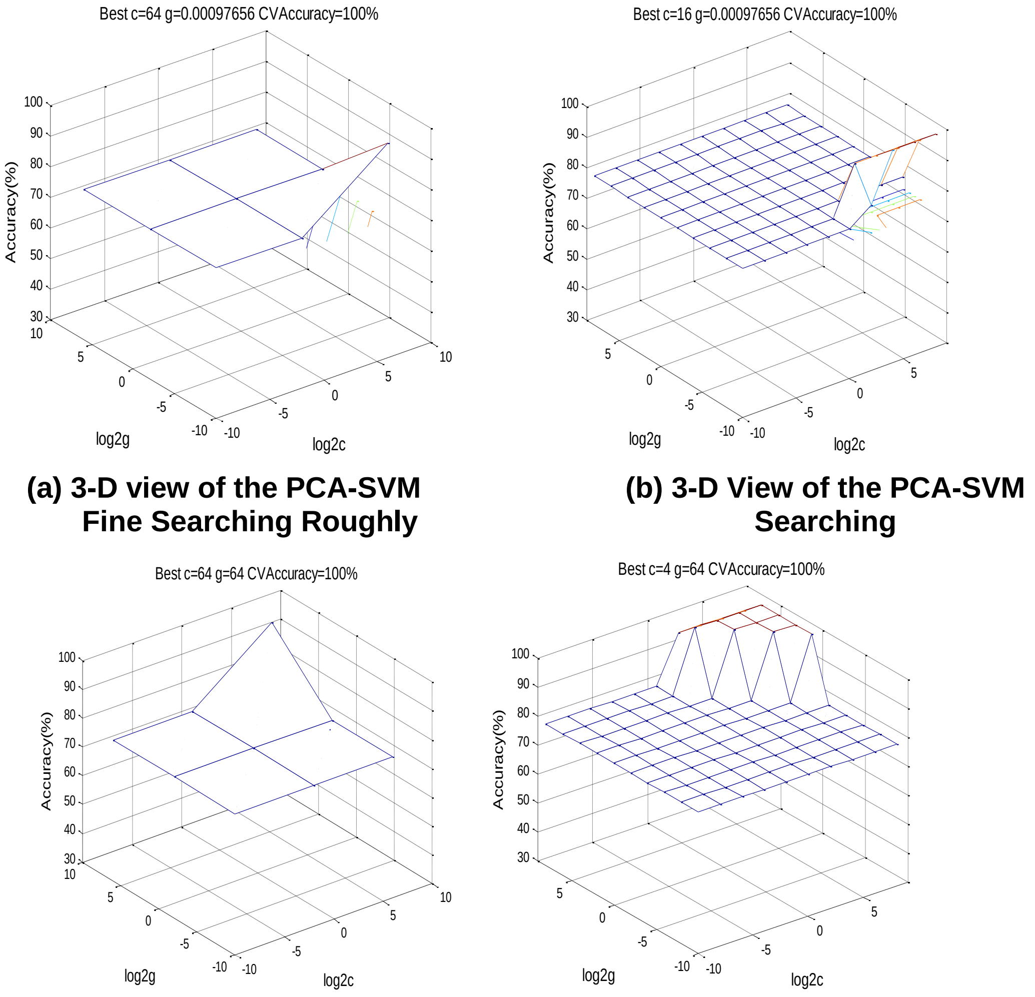
Cumulativepercentage in the PCA-SVM based on the improved grid search algorithm is 90%, and the featuredimension of the data sets after data dimension reduction are 20,51,29 respectively. Cumulativepercentage in the KPCA-SVM based on the improved grid search algorithm is 100%, and the featuredimension of the data sets after data dimension reduction are 50,35,18 respectively. In experiment1, both the searchingrange of C and σ are $\left[2^{-10},2^{10}\right]$, and in experiment2, both the searchingrange of C and σ are $\left[2^{-15},2^{15}\right]$. The result of experiment is shown inTable1 and Table2. Figure2 shows three-dimensionalview of PCA-SVM and KPCA-SVM searchingprocess of parameters based on the improved grid search algorithm. Table 3 shows the results of PCA-CSVM optimizing the parameter in different ranges based on different grid search. Table 4 shows the results of KPCA-CSVM optimizing the parameter in different ranges based on different grid search.

Table 1. Comparison of PCA-CSV Moptimizing Parameters in Different Ranges based on the Improved Grid Search

Performance index		Singh-2002	Armstrong-2002-v1	Lahio-2007
Experiment1	Searchingrange of C	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$
	Searchingrange of σ	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$
	Best C	16	16	16
	Best σ	0.0039	0.0039	0.0009765
	Accuracy(%)	95.0908	100	94.59459
	Time (s)	2.811333	1.613956	0.596087
Experiment 2	Searchingrange of C	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$
	Searchingrange of σ	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$
	Best C	8	8	8
	Best σ	0.0078	0.0078	0.0020
	Accuracy(%)	95.0908	100	94.59454
	Time (s)	6.156422	3.247822	0.644178

Table 2. Comparison of KPCA-CSVM Optimizing Parameters in Different Ranges based on the Improved Grid Search

Performance index		Singh-2002	Armstrong-2002-v1	Lahio-2007
Experiment1	Searchingrange of C	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$
	Searchingrange of σ	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$	$\left[2^{-10},2^{10}\right]$
	Best C	256	16	4
	Best σ	256	256	46
	Accuracy(%)	94.11764	98.6111	91.89189
	Time (s)	3.324356	1.120378	0.243011
Experiment 2	Searchingrange of C	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$
	Searchingrange of σ	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$	$\left[2^{-15},2^{15}\right]$
	Best C	512	8	8
	Best σ	128	8	32
	Accuracy(%)	94.11764	98.6111	91.89189
	Time (s)	7.568079	2.092555	0.716548



(a) 3-D view of the PCA-SVM Fine Searching Roughly

(b) 3-D View of the PCA-SVM Searching

(c) 3-D view of the KPCA-SVM Searching Roughly

(d) 3-D View of the KPCA-SVM Fine Searching

Figure 2. 3-D View of the Parameter Searching based on the Improved Grid Search

Table 3. Comparison of PCA-CSVM Optimizing Parameters in Different Ranges based on Different Grid Search

	Data sets	Grid search		Improved grid search	
		Accuracy （%）	Time （s）	Accuracy （%）	Time （s）
Experiment1	Singh-2002	93.1372	11.249312	94.59459	2.811333
	Armstrong-2002-v1	100	21.682194	100	1.613956
	Lahio-2007	91.89189	3.731399	94.59454	0.596087
Experiment 2	Singh-2002	95.0908	22.868459	95.0908	6.156422
	Armstrong-2002-v1	100	42.299089	100	3.247822
	Lahio-2007	91.89189	6.922925	94.59454	0.644178

Table 4. Comparison of KPCA-CSVM Optimizing Parameters in Different Ranges based on Different Grid Search

	Data sets	Grid search		Improved grid search	
		Accuracy （%）	Time （s）	Accuracy （%）	Time （s）
Experiment1	Singh-2002	93.1372	11.884880	94.11764	3.324356
	Armstrong-2002-v1	98.6111	11.06798	98.6111	1.120378
	Lahio-2007	94.59459	2.148789	91.89189	0.243011
Experiment 2	Singh-2002	95.0908	26.385434	94.11764	7.568079
	Armstrong-2002-v1	98.6111	30.23595	98.6111	2.092555
	Lahio-2007	94.59459	4.550801	91.89189	0.716548

It can be seen from the comparison of table 3 and table4 that, compared with the traditional grid searching, improved grid search algorithm has obvious advantages in searching time and accuracy, except that for Lahio-2007, the accuracy is reduced by 2.7% using the KPCA-CSVM based on improved grid searching. The results of experiments show that, the proposed improved grid search algorithm greatly reduces the searching time, and improves the accuracy of classification.

5. Conclusion

In this paper, we classify the datasets using PCA-SVM and KPCA-SVM, and improve the grid search algorithm based on the traditional SVM parameter optimization. Dichotomous search algorithm is used to improve parameter optimization efficiency greatly. The comparative experiment results indicate that the proposed method improve the classification accuracy of SVM and reduce the time of parameter optimization and classification of datasets. The new method can classify the genetic data sets accurately and quickly and have promising applications in many fields.

Acknowledgement

The authors would like to thank for the developers of the datasets:
http://algorithmics.molgen.mpg.de/Supplements/CompCancer/.

References

[1] C. D. Qin and S. Y. Liu, “Tumor geneselection based on double regularized support vector machine”, Journal of Jilin University(Engineering and Technology Edition),vol. 43, no.1, **(2013)**, pp.192-197.

[2] S. Chakrabortya and R. Guob, “A Bayesian hybrid Huberized support vector machine and its applications in high-dimensional medical data”, Computational Statistics and Data Analysis, vol. 55, no. 3, **(2011)**, pp. 1342-1356.

[3] V. Vapnik, E. Levin and Y. Le Cun, “Measuring the VC-dimension of learning Machines”, Neural Computation, vol. 6, no.5, **(1994)**, pp. 851-876.

[4] Z. Y. Chen, J. P. Li and L. W. Wei,“A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue”, Artificial intelligence in medicine, vol. 41, no. 2, **(2007)**, pp. 161-175.

[5] C. F. Lin and S. D.Wang, “Fuzzy Support Vector Machine”, IEEE Transactions on Netural Networks, vol.13, no. 2, **(2002)**, pp.464-471.

[6] Z. S. Yao, C. F. Shao, Z. H. Xiong and H. Yue, “Short-term traffic volumes forecasting of road network based on principal component analysis and support vector machine”, Journal of Jilin University (Engineering and Technology Edition),vol. 38, no.1, **(2008)**, pp. 48-52.

[7] G. Wang, Y. F. Qiu and H. X. Li, “CSVM and its application in the Chinese theme classification”,2010 International Conference on Optics, Beijing, China, **(2010)**.

[8] J. W. Han, T. P. Breckon, D. A. Randell and G.Landini, “The application of support vector machine classification to detect cell nuclei for automated microscopy”, Machine Vision and Applications, vol. 23, no. 1, **(2012)**, pp. 15-24.

[9] J. F. Wang, L. Zhang, G. X. Chen and X. W. He, “A parameter optimization method for an SVM based on improved grid search algorithm”, Applied Science and Technology, vol. 3, **(2012)**, pp. 28-31.

[10] T. Nakagawa, Y. lwahori and M. K. Bhuyan, “Defect Classification of Electronic Board Using Multiple Classifiers and Grid Search of SVM Parameters”, Computer and Information Science, vol. 493, **(2013)**, pp. 115-127.

[11] Z. Liu and Z. Wei, “Image Classification Optimization Algorithm based on SVM”, Journal of Multimedia,vol.8, no. 5,**(2013)**, pp. 496-502.