International Journal of Database Theory and Application
Vol.8, No.2 (2015), pp.71-80
http://dx.doi.org/10.14257/ijdta.2015.8.2.08

Large-Scale Data Classification Method Based on Machine Learning Model

Hao Jia

Department of Electrical Engineering, Dalian Institute of Science and Technology, Dalian 116052 China
jiangdao1979@yeah.net

Abstract

Classification is to map the data item in the database into a given class. It is an important research direction in data mining. In allusion to the shortcomings of traditional classification methods, such as the decision tree, K nearest neighbor, Bayes , fuzzy logic, genetic algorithms and neural networks and so on, the support vector machine with perfect theory, strong adaptability, global optimization, short training time, good generalization performance is introduced into the classification, a machine learning model based on the SMO algorithm and RBF kernel function of the SVM is proposed to realize a classification method in this paper. This method transforms the nonlinear classification problem into linear classification problem by improving the data dimension. It can better solve the problems of the minimum error in the training set and the larger error in the test set in the traditional algorithm. Application of UCI classification experiment shows that the proposed method takes on the better convergence, faster training speed and higher classification accuracy.

Keywords: *classification, support vector machine, sequential minimal optimization algorithm, RBF kernel function, machine learning model*

1. Introduction

Data mining is regarded as multidisciplinary science of statistics, artificial intelligence and machine learning [1]. It originated in the mid-1990s of twentieth Century and is young, active and hot research field in recent years [2]. Classification technology is one of the most practical technologies in data mining. It is a learning process for mapping the data samples into a defined class. That's to say, it will classify a given set of input vectors and their attributes by using the inductive learning method based on the classification. With producing of a large amount of information, the data are increasing by millions. It is becoming a new challenge for us how to mine useful information from these data.

Data classification is an important concept in the data mining [3]. Its process generally divided into two steps: the first step is to establish the classification model, which describes the predetermined data set or concept set. The mode is constructed by analyzing the database of attribute describing. The second step is to use the classification to classify new data set. It mainly involves the accuracy of classification rules. A good classification rule set should be higher accurate, lesser contradictions division and smaller rule sets for the new data set. Support vector machine (SVM) is based on statistical learning theory. It is a new machine learning method based on the finite sample and overcome the shortcomings of the neural network and traditional classification methods, such as over learning, local extreme and the curse of dimensionality and so on. The SVM can effectively solve the small sample, nonlinear problem and is an effective tool for solving classification problems and regression estimation [4-5]. So it widely concerned by more and more researchers.

ISSN: 2005-4270 IJDTA

However, because the scale of the data are more and more, the support vector machine need use a lot of memory in the learning process, the searching speed is very slow. So the support vector machine shows the bottleneck problem of the slow training speed for large scale data sets. Many researchers have tried to solve the bottleneck of the slow training speed for large scale data sets in these years. Their proposed methods can be divided into two categories: (1) improved SVM algorithms. For example, Huang *et. al.*[6] proposed a new, simple, and efficient network architecture which consists of several SVM each trained on a small subregion of the whole data sampling space and the same number of simple neural quantizer modules which inhibit the outputs of all the remote SVM and only allow a single local SVM to fire (produce actual output) at any time. The experiments on a few real large complex benchmark problems demonstrate that our method can be significantly faster than single SVM without losing much generalization performance. Dong *et. al.*.[7] proposed an fast SVM training algorithm to solve this problem. This method introduces a parallel optimization step to quickly remove most of the nonsupport vectors. Some effective strategies such as kernel caching and efficient computation of kernel matrix are integrated to speed up the training process... Chen *et. al.*. [8] studied sequential minimal optimization type decomposition methods under a general and flexible way of choosing the two-element working set in order to improve the training speed of the classification. These methods can improve the speed of training samples on a certain extent. But they are not very ideal for large data sets. (2) reduce the data size and training samples by using some other algorithms. For example, Cervantes *et. al.*. [9] proposed a new method, SVM classification based on fuzzy clustering. The proposed approach is scalable to large data sets with high classification accuracy and fast convergence speed. Empirical studies show that the proposed approach achieves good performance for large data sets. Cervantes *et. al.*.[10] introduced a novel two-stage SVM classification approach for large data sets: minimum enclosing ball (MEB) clustering is introduced to select the training data from the original data set for the first stage SVM, and a de-clustering technique is then proposed to recover the training data for the second stage SVM. The proposed method was applied in several benchmark problems, experimental results demonstrate that our approach have good classification accuracy while the training is significantly faster than other SVM classifiers. Li *et. al.*.[11] proposed a clustering algorithm for efficient learning. The method mainly categorizes data into clusters, and finds critical data in clusters as a substitute for the original data to reduce the computational complexity. The computational experiments presented in this paper show that the clustering algorithm significantly advances SVM learning efficiency.

In allusion to the existing shortcomings of the SVM and improved SVM in classification, a machine learning model based on the SMO algorithm and RBF kernel function of the SVM is proposed by in-depth studying and researching the related theory of machine learning training algorithm and kernel function of the SVM in this paper. The method can correctly and effectively realize the classification tool of the SVM for the data set. This method transforms the nonlinear classification problem into linear classification problem by improving the data dimension. It can better solve the problems of the minimum error in the training set and the larger error in the test set in the traditional algorithm. Application of UCI classification experiment shows that the proposed method takes on the better convergence, faster training speed. The classification accuracy is improved.

2. Support Vector Machine

Support vector machine (SVM), introduced by Vapnik [12], is one of the most popular tools in bioinformatics for a supervised machine learning methods based on structural risk minimization. The basic characteristic of SVM is to map the original nonlinear data into a higher-dimensional feature space where a hyperplane is constructed to bisect two classes

of data and maximize the margin of separation between itself and those points lying nearest to it (the support vectors). The hyperplane should be used as the basis for classifying unknown data. So SVM was widely applied in pattern recognition, nonlinear system identification, modeling, predication and control and so on. The SVM is mainly used to solve the binary classification problem. The theory was originally derived from data classification. The SVM is to find one division plane with meeting the given requirement in order to keep the point of the training set far away the plane. In other words, it is to find one split plane to keep the largest classification interval (margin). The SVM originated from the optimal classification surface from the linearly separable circumstance. It is used to solve the linear constrained quadratic programming problem by mapping the input space into the high dimensional inner product space in order to obtain the global optimal solution to guarantee convergence speed and avoid the local minimum problem.

The Given the training sample is $\{x_i, y_i \mid i = 1,2,3,\cdots,m\}$, m is the number of samples, the set $\{x_i\} \in R_n$ represents the input vector, $y \in \{+1,-1\}$ indicates the corresponding desired output vector, the input data is mapped into the high dimensional feature space by using nonlinear mapping function $\phi(\bullet)$. In the high-dimensional feature space, the constructed optimal classification hyperplane may be separated by one hyperplane $w^* x + b = 0$. Each sample point is satisfied by the followed expression:

$$y_i[w^* \phi(x_i) + b] - 1 \geq 0, i = 1,2,3,\cdots,m \qquad (1)$$

where w represents the weight vector, b is the threshold value. At this time, the classification interval (Δ) is $2/w$. So the maximum interval is equivalent to the minimum of the $\| w \|^2$. It meets the equation (8), and the optimal classification plane is the classification plane of the smallest $\| w \|^2$. The slack variables of the ξ_i and ξ_i are used to measure the distance between the actual value y_i and the support vector machine. The optimization problem of data separation plane is transformed into the following optimization problem:

$$\begin{cases} \min \dfrac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i \qquad i = 1,2,3,\cdots,m \\ s.t. \begin{cases} y_i (w^* x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \end{cases} \qquad (2)$$

where C is penalty parameter, which is used to control the punish degree.

The multiplier α_i and kernel function $k(x_i, y_i) = \phi(x_i)\phi(x_j)$ of the Lagrangian is introduced to transform the above optimization problem into the quadratic programming optimization problem.

$$\begin{cases} \max L(\alpha) = \sum_{i=1}^{m} \alpha_i - \dfrac{1}{2} \sum_{ij}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1,2,3,\cdots,m \end{cases} \qquad (3)$$

The corresponding point $\alpha_i > 0$ is called support vector, the number of support vectors is less than the number of training samples in general. The classification decision function is obtained by the following expression.

$$f(x) = sign[\sum_{i,j=1}^{m} \alpha_i y_i k(x_i, x_j) + b] \quad (4)$$

In the SVM, the used kernel functions have radial basis function (RBF) $k(x_i, x_j) = \exp(-\| x_i - x_j \| / 2\sigma^2)$ (σ is the parameter of RBF), the polynomial function $k(x_i, x_j) = (x_i x_j + b)^d$, S function $k(x_i, x_j) = \tanh[k(x_i, x_j) + v]$ $(k > 0, v < 0)$ .In this paper, the radial basis function(RBF) is selected as the inner product kernel function.

3. Machine Learning Model Based on Support Vector Machine

A machine learning algorithm model based on support vector machine is proposed in this paper. A training algorithm of sequential minimal optimization (SMO) algorithm and radial basis function kernel function are researched and used to implement the machine learning algorithm model based on support vector machine. The sequential minimal optimization (SMO) algorithm and radial basis function kernel function are selected to implement the classification function of data set. The machine learning algorithm model is required to deal with linear and nonlinear data in order to obtain the correct classification result and analyze the classification effect.

The main goal of the support vector machine is to find a hyperplane, which is used to rightly divide data set into two classes of data. At the same time, the separate data points are most far to the classification plane. The linear and nonlinear training sample sets are respectively discussed in here. And the nonlinear classification problem is transformed into the linear classification problem for solving. The machine learning algorithm model based on support vector machine is shown in Figure1.

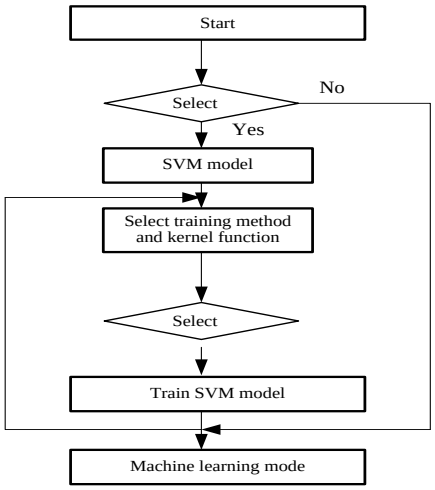


Figure 1. The Machine Learning Algorithm Model Based on support Vector Machine

4. The Key Technology Analysis of Machine Learning Mode

4.1. The Selection of Kernel Function

An attention characteristic of support vector machine is to use the kernel function of the satisfied Mercer conditions to replace the inner product operation among vectors in order to realize the nonlinear transform. There does not need the specific

form of the nonlinear. According to the above idea, some researchers improved the classical linear algorithm and proposed the corresponding nonlinear form based on the kernel function. The most important parameter of support vector machine model is the kernel function. What kernel function is selected, it will mean that the training samples are mapped into what kind of space in order to realize the linear division. In high dimensional feature space, the dot product operation is required by using the function of the original space. Transformation form is being not needed know. According the relevant theories of the functional, a kernel function can satisfy Mercer condition, it will correspond to the dot product in the transformation space. Therefore, he appropriate dot product function is used to achieve the linear classification in the optimal classification. And the computational complexity is not increased.

There are some common kernel functions, such as linear kernel, polynomial kernel function, the radial basis kernel function, sigmoid kernel function, Fourier kernel function and so on. The different system process data are executed regression estimation; there have the corresponding optimal kernel function. Because the radial basis function takes on the simple form, radial symmetry, smoothness and good analysis, it is widely applied. So the radial basis function is selected as the kernel in the regression model. The specific form is described as follows:

$$K(x, x_i) = \exp(-\|x - x_i\|^2 / \sigma^2) \qquad (5)$$

where, x is the dimension of input vector m, x_i is the i^{th} center of radial basis function, it has the same dimension with x, σ the standard parameter, which is used to determine the center point width of the function, $\|x - x_i\|$ is the norm of vector $x - x_i$, which is used to described the distance between x and x_i.

The coefficient of kernel width (σ) reflects the correlation degree among support vectors. It is related to the input space of learning samples. If the input space of learning samples is larger, the value (σ) is larger. If the value (σ) is smaller, the connection among support vectors is more relaxed, and the learning machine is relatively complex and the generalization ability is not guaranteed. If the value (σ) is larger and the influence among support vectors is stronger, the regression model is difficult to achieve the sufficient accuracy.

4.2. The SVM Based on SMO Algorithm

Sequential minimal optimization (SMO) algorithm is a simple algorithm; it can quickly solve the quadratic programming problem of the SVM. According to Osuna theory, under guaranteeing the convergence, the quadratic programming problem of the SVM is decomposed into a series of sub problem to solve. Compared with other algorithms, the SMO method selects a minimal optimization problem to solve in each step. For the optimization problem of the standard SVM, the minimum optimization problem is the optimization problem of two Lagrange multipliers. In each step, the SMO method selects two Lagrange multipliers to optimize. Then the SVM is updated in order to reflect the new optimal value. The SMO method consists of two steps: one is to use analysis method for solving a simple optimization problem; the other is to select the Lagrange multipliers strategy of the optimizing.

4.2.1. The Optimization Solution of Two Large Multipliers

In order to understand the optimization problem of two multipliers, the SMO method is firstly used to calculate its constraint, and then solve the constrained

minimization problem. For the quadratic programming problem of the SVM, two multipliers are considered, that's (i, j). Definite auxiliary variables $s = y_i y_j$, the pattern recognition problem is $y_i \in \{1, -1\}$. For function approximation problems, there must be four different situations: $(\alpha_i, \alpha_j), (\alpha_i, \alpha_j^*), (\alpha_i^*, \alpha_j), (\alpha_i^*, \alpha_j^*)$. For (α_i, α_j) and (α_i^*, α_j^*), suppose s = 1. For the other conditions, suppose s = − 1. So, for the pattern recognition problems, the following constraints are obtained:

$$s\alpha_i + \alpha_j = s\alpha_i^{old} + \alpha_j^{old} = \gamma \quad (6)$$

For function approximation problems, the constraints are:

$$(\alpha_i - \alpha_i^*) + (\alpha_j - \alpha_j^*) = (\alpha_i^{old} - \alpha_i^{*old}) + (\alpha_j^{old} - \alpha_j^{*old}) = \gamma \quad (7)$$

$\alpha_j^{(*)} \in [0, C_j^{(*)}]$ is used to obtain $\alpha_i^{(*)} \in [L, H]$.

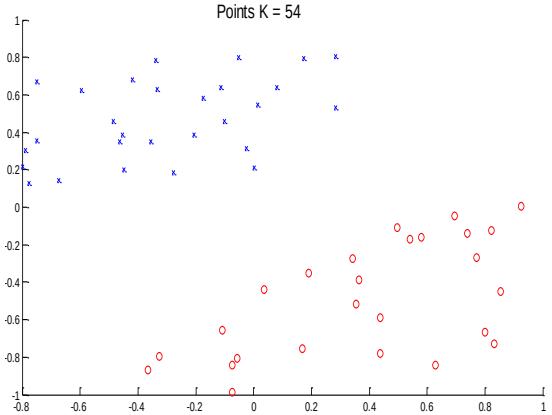
4.2.2. Select the Lagrange Multipliers Strategy

After the SMO method solves the problem of two multipliers, Lagrange multiplier is updated in each step. In order to speed up the convergence, the SMO method uses one outer loop to realize he first multiplier selection. The outer loop decides whether each sample meets KKT condition in the whole training set. If one does not meet the KKT condition, it will be selected to optimize. After training samples in the whole training set meet the above conditions, all samples of the located boundaries are inspected to select the second multiplier. The SMO selects the minimal multiplier of the objective function, which is regarded as the second multiplier to optimize. If this method fails, the SMO will search the all non boundary samples in order to find out the minimal multiplier of the objective function.

5. Analysis of the Experimental Simulation and Application of Machine Learning Model

5.1. Classify by Using the SMO Algorithm and Kernel Function

The data set is linear separable data, the classification results is obtained by using the SMO algorithm and kernel function:



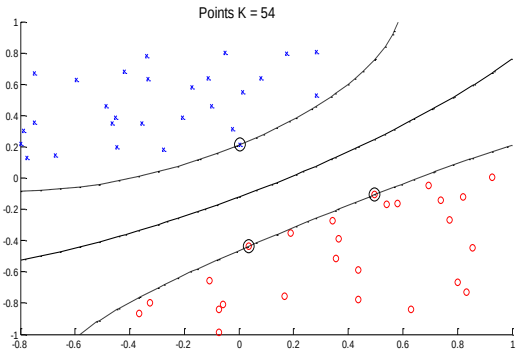


Figure 3. The Classification Results (points K=54)

5.2. Analysis of Large-Scale Data Classification Based on Machine Learning Model

In order to analyze the effectiveness of large-scale data classification method based on machine learning model, five data sets from the UCI are selected to test and verify the proposed method. The basic information of five data sets are show in *Table*.1. In this experiment, Matlab R2010b is used to realize support vector machine algorithm. When the classification method of the SVM carried out the training, the kernel function used the radial basis function (RBF). the value of σ is 0.5.The value of the penalty parameter(C) is obtained by using the 5-fold cross validation method.

Table 1. The Detail Information of Five Data Sets from the UCI

Index	Data set	Samples	Attributes	Classification
1	adult	48842	14	2
2	letter	20000	16	28
3	pima	768	8	2
4	statlog	6435	36	6
5	wine	178	13	3

The five data sets of adult, letter, Pima, statlog and wine randomly selected training samples, which are 8000, 4000, 350, 2000 and 63. These samples are classified by using the machine learning model of the large-scale data classification, then the obtained classification accuracy were 83.6%, 94.1%, 84.4%, 93.9% and 100%. The classification results are shown in Table 2.

Table 2. The Classification Results of the Training Samples from the UCI

Index	Data set	Training samples	Classification accuracy(%)
1	adult	8000	83.6
2	letter	4000	94.1
3	pima	350	84.4
4	statlog	2000	93.9

5	wine	63	100.0

In order to prove the effectiveness of large-scale data classification method based on machine learning model, the proposed method is compared with other representative SVM and WSVM method [13]. The experiment results are shown in *Table*.3.

Table 3. Comparing Classification Results of Three Methods

Index	Data set	SVM		WSVM		The proposed method	
		Training time(s)	Accuracy (%)	Training time(s)	Accuracy (%)	Training time(s)	Accuracy (%)
1	adult	4028.4	78.4	3012.3	81.5	2319.1	83.6
2	letter	3976.2	86.2	3546.2	91.3	3020.4	94.1
3	pima	50.2	78.6	38.2	80.8	30.3	84.4
4	statlog	2106.3	86.3	1767.3	89.2	1582.7	93.9
5	wine	17.2	89.3	8.2	94.3	2.7	100.0

From Table 3, we can see that the proposed large-scale data classification method based on machine learning model has better classification accuracy than the SVM and WSVM methods for five data sets from the UCI. In the training time, the proposed large-scale data classification method based on machine learning model uses less time than the SVM and WSVM methods for training samples. In general, under the same number of training samples, the proposed large-scale data classification method based on machine learning model takes on the better convergence, faster training speed and higher classification accuracy.

6. Conclusion

Data classification is an important concept in the data mining. Its process generally divided into two steps: the first step is to es*Table*lish the classification model, which describes the predetermined data set or concept set. The second step is to use the classification to classify new data set. It mainly involves the accuracy of classification rules. Support vector machine (SVM) is based on statistical learning theory. It can effectively solve the small sample, nonlinear problem and is an effective tool for solving classification problems and regression estimation. In allusion to the existing shortcomings of the SVM and improved SVM in classification, we proposed a machine learning model based on the SMO algorithm and RBF kernel function. The method can correctly and effectively realize the classification tool of the SVM for the data set. This method transforms the nonlinear classification problem into linear classification problem by improving the data dimension. It can better solve the problems of the minimum error in the training set and the larger error in the test set in the traditional algorithm.

Acknowledgements

The authors would like to thank all the reviewers for their constructive comments. This research was supported by the National Natural Science Foundation of China (U1433124), Open Project Program of State Key Laboratory of Software

Engineering(SKLSE) (SKLSE2012-09-27), Open Project Program of Provincial Key Laboratory for Computer Information Processing Technology, Soochow University (KJS1326), the Open Project Program of Artificial Intelligence Key Laboratory of Sichuan Province (Sichuan University of Science and Engineering) (2014RYJ01,2014RYJ02).

References

[1] V. Nedic, S. Cvetanovic, D. Despotovic, "Data mining with various optimization methods. Expert Systems with Applications, vol. 41, no. 8, **(2014)**, pp. 3993-3999.
[2] W. Y. Feng, Q. L. Zhang, "Mining network data for intrusion detection through combining SVMs with ant colony networks", Future Generation Computer Systems, vol. 37, **(2014)**, pp. 127-140.
[3] K. B. Duan, J. C. Rajapakse, "Multiple SVM-RFE for gene selection in cancer classification with expression data", IEEE Transactions on Nan bioscience, vol. 4, no. 3, **(2005)**, pp. 228-233.
[4] C. P. Hou, F. P. Nie, "Multiple rank multi-linear SVM for matrix data classification", Pattern Recognition, vol. 47, no. 1, **(2014)**, pp. 454-469.
[5] P. Mahesh, "Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data", Transactions of the ASME IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 5, no. 5, **(2012)**, pp. 1344-1355.
[6] G. B. Huang, K. Z. Mao, C. K. Siew, "Fast modular network implementation for support vector machines", IEEE Transactions on Neural networks, vol. 16, no. 6, **(2005)**, pp. 1651-1663.
[7] J. X. Dong, A. Krzyzak and Y. Suen, "Fast SVM training algorithm with decomposition on very large data sets", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 4, **(2005)**, pp. 603-618.
[8] P. H. Chen, R. E. Fan and C. J. Lin, "A study on SMO-type decomposition methods for support vector machines", IEEE Transactions on Neural Networks, no. 17, no. 4, **(2006)**, pp. 893-908.
[9] J. Cervabtes, X. Li and W. Yu, "Support vector machine classification based on fuzzy clustering for large data sets", MICAI'06: Proceedings of the 5th Mexican International Conference on Artificial Intelligence, LNCS 4293.Berlin: Springer: **(2006)**, pp. 572-582.
[10] J. Cervabtes, X. Li, W. Yu, "Multi-class support vector machine for large data sets via minimum enclosing ball clustering", Proceeding of the 4th International Conference on Electrical and Electronics Engineering, Piscataway: IEEE, **(2007)**, pp. 146-149.
[11] D. C. Li and Y. H. Fang, "An algorithm to cluster data for efficient classification of support vector machines", Expert Systems with Applications, vol. 34, no. 3, **(2008)**, pp. 2013-2018.
[12] V. Vapnik, "Statistical Learning Theory", New York: John Wiley & Sons, **(1998)**, pp. 253-256.
[13] C. X. Wang, T. Z. Tao and C. S. Ma, "Resolution of classification for imbalanced dataset based on cluster-weight and grading-SVM algorithm", Computer Engineering and Applications (Doi:10.3778/j.issn.1002-8331.1311-0145), **(2014)**.

Author

Hao Jia, Engineer, received the Master degree in electrical engineering from Dalian Jiao tong University in 2005, Dalian, China. The main research directions: Artificial Intelligence, Electrical Engineering.

