Research on SVM Improved Algorithm for Large Data Classification

Hong Dai
College of International Finance & Bank
Liaoning University of Science and Technology
Anshan LiaoNing
e-mail: dear_red9@163.com

***Abstract*—In view of the two problems of the SVM algorithm in processing large data, the paper proposed a weighted Euclidean distance, radial integral kernel function SVM and dimensionality reduction algorithm for large data packet classification. The SVM cannot handle multi classification and time of building model is long. The algorithm solved these problems. The improved algorithm reconstructs the data feature space, makes the boundary of different data samples clearer, shortens the modeling time, and improves the accuracy of classification. The proposed method verified the feasibility and effectiveness with experiments. The experimental results show that the improved algorithm can achieve better results when multi-duplicated samples and large data capacity are used for multi classification.**

Keywords- support vector machine (SVM); large data; multi-classification;Euclidean distance; radial integral kernel function

I. INTRODUCTION

The rapid development of network technology makes a huge amount of data every day. The rapid and accurate classification of the vast amounts of data collected is necessary to extract comprehensible knowledge. According to forecasting by market research firm IDC, global data will exceed 40ZB by 2020[1]. Many industries have provided storage systems with capacity ranging from tens of gigabytes to hundreds of terabytes, or even petabytes. But nearly 60% of the data is repeated, which not only increases data storage, processing time, but also leads to higher and higher costs of data analysis and classification. The efficient and accurate classification algorithm is one of the hot issues in current industry research. There are some common classification algorithms. For example, K-Nearest Neighbor 、 Native Bayes、Neural Net、Support Vector Machine and Linear Least Square Fit and so on[2].

Support vector machine (SVM) algorithm is a kind of machine learning method based on VC dimension theory in statistical learning theory and structural risk minimum principle. It has excellent data classification and regression processing ability[3]. The support vector method was first proposed by Vapnik to solve the problem of pattern recognition. It selects a set of characteristic subsets from the training samples, so that the classification of the characteristic subset is equivalent to the division of the whole dataset. The characteristic subset is called the support vector (SV). Due to its excellent learning ability, the application scope is very wide. For example, intrusion detection, facial expression classification, Time series prediction, speech recognition, signal processing, Gene detection, text classification, font recognition, Fault diagnosis, chemical analysis, image recognition and other fields. SVM algorithm has some obvious advantages in solving classification problems. It has a shorter forecast time. The global optimal solution can guarantee the accuracy of the target detection classifier in the classification. But there are some disadvantages, such as the detection model is established for a long time. Time complexity and space complexity increase linearly with the increase of data when processing large scale data.

The data objects are often large data sets in the emerging fields of data mining, document classification and multimedia indexing. The number of attributes and the number of records are very large resulting in poor execution of the processing algorithm[4]. The classifier is only determined based on support vector machine by support vector. The complexity of the classifier is not related to the number of training samples. It only has to do with the number of support vectors[5]. In the paper, propose a weighted Euclidean distance, the radial product kernel function and the decreasing dimension packet support vector machine method, reduces the data dimension, remove redundant feature attributes and duplicate data. A classification model with better generalization ability is obtained by using less support vectors. Reducing storage and processing of data resources, speed up the classification model established time, solve big data classification problems.

II. THE WEIGHTED EUCLIDEAN DISTANCE AND THE RADIAL PRODUCT KERNEL FUNCTION SVM

A. SVM Theory

SVM developed from the optimal classification of linearly separable problems. The support vector machine algorithm is to find the classification line to make the classification interval the largest. The two types of samples are correctly separated and extended to the spatial form to find the optimal classification surface.

Linear support vector machine classification algorithm is described as follows:

(1) Known training set T=$\{(x_1,y_1),\cdots,(x_i,y_i)\in(\boldsymbol{X}\times\boldsymbol{Y})'\}$, Among them $x_i\in\boldsymbol{X}=\boldsymbol{R^n}$, $y_i\in\boldsymbol{Y}=\{1,-1\},i=1,\cdots n;$

(2) To solve the optimization problem, the optimal solution is obtained $a^* = (a_1^*, \cdots, a_n^*)^T$.

(3) Calculate $w^* = \sum_{i=1}^{n} a_i^* y_i x_i$, choose a positive component a*j of a*, and compute $b^* = y_j - \sum_{i=1}^{n} y_i a_i^* (x_i \cdot x_j)$

(4) Put $w^* = \sum_{i=1}^{n} a_i^* y_i x_i$ into the hyperplane equation to get the optimal function. That is the decision function:

$$f(x) = \operatorname{sgn}\{(w * x) + b\} = \operatorname{sgn}\left\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i * x) + b^*\right\}$$

sgn () is the symbol function. a_i^* is Lagrange multiplier for each sample. b^* is the classification threshold. $K(X_i, X)$ transforms the inner product of space. For nonlinear problem, the problem is transformed into a linear problem of high dimensional space by nonlinear transformation. Solve the optimal classification surface in the optimal classification surface. Both the optimization and classification problems involve only the inner product operation in the dual problem. According to the relevant theory of the function, only one kernel function is required to satisfy the Mercer condition. It corresponds to the inner product of a transformation space. Use the inner product function $K(x_i, x_j)$ to realize the optimal classification surface. The optimal classification surface is realized by the inner product function $K(x_i, x_j)$. The corresponding objective function is

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j),$$

Its classification function is

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*\right\} \qquad (1)$$

This is the support vector machine.

The support vector machine is the nonlinear transformation of the inner product function. Map low dimensional input space to high dimensional space. Find the optimal classification surface in high - dimensional space and classify the sample. The output is a linear combination of intermediate nodes, and an intermediate node corresponds to a support vector. The core idea of SVM is to find the optimal classification hyperplane to meet the classification requirements. The noise points near the edge of the class are often mixed with valid samples in the SVM algorithm. Therefore, the classification surface is often not optimal. The performance of the classifier is not optimal. In the paper, propose a weighted Euclidean distance and radial product kernel function and the descending dimension packet SVM. Remove noise from the disturbance near the edge of the class. Reconstructing the eigenspace and reduce the dimension of feature space and further improve the classifier performance.

B. The Weighted Euclidean Distance and the Radial Product Kernel Function SVM

The support vector machine algorithm is to map the input vector to the higher-dimensional space through a kernel function. This space is called Reproducing Kernel Hilbert Space. The given kernel defines the corresponding Hilbert space. Definition of inner product is below in RKHS:

$$< f, g >= \sum_i \sum_j \alpha_i \beta_j K(x_i, x_j)$$

Nonlinear transformation can be obtained. $\langle \Phi(x), \Phi(y) \rangle = \langle k(\bullet, x), k(\bullet, y) \rangle = k(x, y)$.

Theorem 1（Mercer theorem）

(X, μ) is a finite measure space. Use $k \in L_\infty(X^2)$ construct integral operator of square integrable function on χ if it is a symmetric real valued function $T_k : L_2(X) \to L_2(X)$

$$(T_k f)(x) := \int_{\chi} k(x, y) f(y) d\mu(y) \qquad (2)$$

That is positive semidefinite $\forall f \in L_2(X)$ $\int_{\chi} k(x, y) f(x) f(y) d\mu(x) d\mu(y) \geq 0$.

$\varphi_j \in L_2(X)$ is standard characteristic function of T_K. The corresponding eigenvalues are $\lambda_j > 0$ arranged in descending order. There is $k(x, y) = \sum_{j=1}^{N_F} \lambda_j \varphi_j(x) \varphi_j(y)$ on X^2. $N_F \in N$ or $N_F = \infty$. $k(x, y)$ corresponding vectors of $l_2^{N_F}$, Warp transformation

$$\begin{aligned} \Phi : X &\to l_2^{N_F} \\ x &\to \left(\sqrt{\lambda_j} \varphi_j(x)\right)_j \end{aligned} \qquad (3)$$

The inner product can be represented as $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Thus, linear learning is performed in the feature space F. Replace the inner product of the linear algorithm with the kernel function that satisfies the Mercer theorem. Don't have to know the exact form of the nonlinear transformation.

Theorem 2 The training and training results of support vector machines have nothing to do with non-support vectors. The classification function (1) of support vector machine can be seen. The determinant of a classification hyperplane is a sample of the corresponding Largrange multiplier greater than 0 in the classification method of SVM. The calculation of the classification hyperplane is only related to these samples, which is not related to other non-support vectors in theorem 2.

The problem of vector machine is to use decomposition method to decompose large scale problems into small scale problems. Solve the sub-problem iteratively and iteratively. Due to the sparse feature of solution (1), the non-support vector consumes a lot of time of SVM modeling. If these non-support vectors can be artificially removed before the model is established, the time for the test model built will be greatly improved. The influence of different attributes results is different on the classification. Give a certain weight to each attribute of the data set according to the criteria. That is the property weighting. Modify the standard Euclidean distance by using the attribute weight w.

$$d^w(x_i, x_j) = \sqrt{\sum_{k=1}^{n} w_k \left| x_{ik} - x_{jk} \right|^2} \qquad (4)$$

$d^w(x_i, x_j)$ is the weighted Euclidean distance between the two sample points x_i and x_j. x_{ik} is the first k attribute value of the sample. $w = (w_1, \cdots, w_n)$ is weight vector. That

is, the sample attribute value is stretched or contracted according to the weight vector in Euclidean space. Weight vector is the importance measure of each attribute. For the weight vector w use the information gain to describe. Information gain can be used to measure the amount of information contained in an attribute. The greater the attribute information gain, the greater the effect of this attribute on classification. Each sample (tag number removal) is described by n attributes $(f_1, \cdots, f_n)$ in the dataset D. Then the weight vector can be expressed as $w = \sqrt{G} = (\sqrt{Gain(f_1)}, \cdots, \sqrt{Gain(f_n)})$. Among them, Gain (fi) is the sample attribute information gain. K is the kernel function in the sample space $X \times X$ in improving algorithm of support vector machine based on attribute weighting. $X \subseteq R^n$. P is the n order linear transformation matrix of given input space. N is the dimension of input space. Attribute weighted kernel function is defined as $K_p(x_i, x_j) = k(x_i^T P, x_j^T P)$. Attribute weighted radial product kernel function is

$$\begin{aligned} k_p(x_i, x_j) &= \exp\left(-\frac{\left|x_i^T P - x_j^T P\right|^2}{\sigma^2}\right) \\ &= \exp\left(-\frac{(x_i - x_j)^T PP^T (x_i - x_j)}{\sigma^2}\right) \end{aligned} \quad (5)$$

The attribute weighting matrix used in this paper is n order diagonal matrix, and the meridional transformation matrix P is shown in formula (6).

$$P = \begin{bmatrix} W_1 & & & \\ & W_2 & & \\ & & \ddots & \\ & & & W_n \end{bmatrix} \quad (6)$$

The attributes are linear transformed in order to make the sample SVM classification achieve a better effect. The shape of the feature space changes as well. Better linear classification hyperplane can be found in the new feature space. Kernel function calculation can also effectively avoid the influence of some weak correlation or irrelevant features in order to improve the performance of SVM classification.

The support vector machine algorithm cannot handle multi-classification problem. If the noise in the raw data is not properly handled, the characteristics will be extracted into the feature space. Similarly, the large data generated are also typical high-dimensional data such as image, sound and video by multimedia applications[6]. In the paper, use the method of reducing the feature disturbance to deal with the large-scale data problem. Under the condition of not reducing the performance of classifier, the training samples are divided into block learning by point - point voting classification. The modeling time is accelerated by reducing the number of training samples.

The data characteristic space of the sample is classified in the process of constructing classifier. For C classification, construct $C(C-1)/2$ second classifier. $D_{ij}(x)$ is the optimal classification hyperplane of class i and class j. The direction of the hyperplane is $D_{ij}(x) = -D_{ji}(x)$, and for the discriminant category vector x,

$$D_i(x) = \sum_{i \neq j, j=1}^{c} \operatorname{sgn}(D_{ij}(x)) \quad (7)$$

Among them, $\operatorname{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$

It belongs to category: $\arg \max_{i=1,\cdots,C}(D_i(x))$ (8)

If the maximum is the same, define its category as the first largest category in the order. The training data is collected together with the data of the same category, as train1、train2、train3、train i、……、train j、……、train n. Set up second classification model svm12 for training data set of train1 and train2. Set up second classification model svm13 for training data set of train1 and train3. Thus, establish the second classification model SVMij. Use model svm12, svm13,... , svmij predict a class for the test set. Get the result of the vote and finally decide the data category.

Adopt the normalization process in order to prevent large numeric attributes from flooding small numeric attributes in the paper. Each attribute is mapped to [-1, 1] with the maximum and minimum value of the characteristic property according to x*=2*(x-x mean)/(xmax-xmin). xmean is the mean of the property x. xmax is the maximum value of attribute x, and xmin is the minimum value of attribute x. The data of the normalized data are simplified. The data reduction number is determined based on the classification test accuracy of the training data set.The radial product kernel function and the descending dimension packet SVM experiment based on the weighted Euclidean distance.

The experiment is divided into two phases, the training phase and the testing phase. Construct the weighted Euclidean distance and the radial integral-kernel function support vector machine algorithm in the training phase. And the training data set is processed in a reduced dimension. Create $5 \times (5-1)/2$ a classification model. The test phase applies the classification model to predict the test data set, and then votes on the predicted results.

The experimental steps:

(1) The character attributes of the training data set are numerically processed、normalization and removal of repeated data preprocessing.

(2) Categorize the voting and calculate the information gain of each attribute in the 10 groups of training.

Construct 10 sets of attribute weight vector w and linear transform diagonal matrix P, each group form is $w = \sqrt{G} = (\sqrt{Gain(f_1)}, \cdots, \sqrt{Gain(f_n)})$, $P = diag(w)$.

(3) The eigenweighted radial integrator replace the radial integrand kernel function in the standard support vector machine.

(4) Train the training data set and establish the test model.

(5) Test the test data set and finish a point-to-point vote.

The experimental environment is Windows XP operating system. Intel(R) Core(TM) 2 Duo CPU T7250，Frequency for 2.00GHz，memory 1.00GB，MATLAB 7.8.0，libsvm-mat-2.9-1. Among them, libsvm-mat 2.9-1 is the program of libsvm in matlab under the university of Taiwan.

Make the value of the character in the property because libsvm cannot handle character attribute values. The

parameters of the radial product kernel function are set uniformly. σ is one in the radial product. The penalty coefficient is one hundred. Weight the penalty coefficients according to the proportion of samples.

The data set contains 391,458 data. It contains a lot of duplicate data. The training model can generate a large number of redundant computations. So delete duplicate data.The number of samples set up is decreased by the model after the deleted data set is 145585 data. The data is limited to a certain extent after processing. Normalize the numeric attribute. On the one hand, the speed of convergence of the program is accelerated. On the other hand, it makes it easier to process data later. Prevent small numeric attributes from being overwhelmed by large numeric attributes. Use the same normalized standard for training sets and test sets. The data of test set and training set are analyzed to obtain the maximum and minimum values of each attribute. Each property is mapped to [-1, 1] by the formula of x*=2*(x-x mean)/(xmax-xmin).

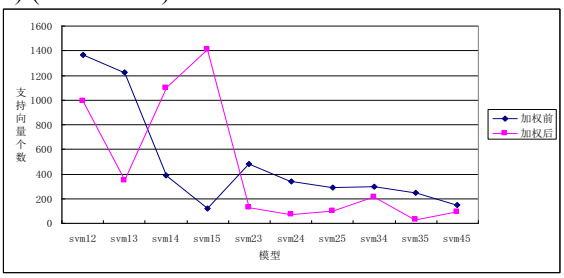

Figure 1. Support vector statistics

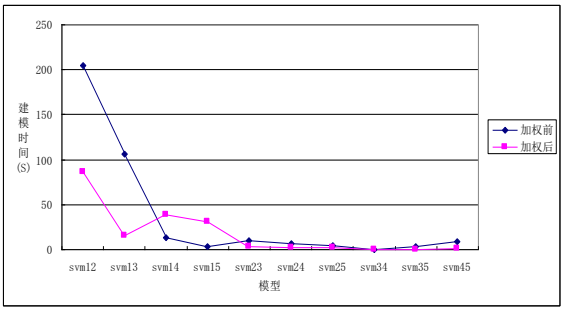

Figure 2. The test model establishes time statistics

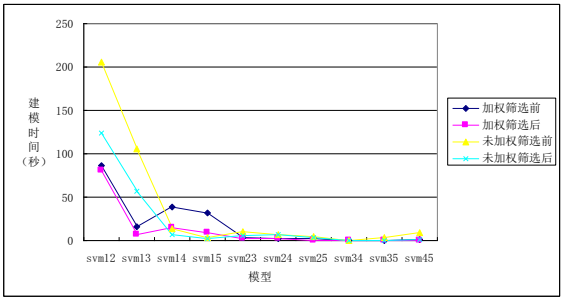

Figure 3. The test model establishes time statistics

Large-scale data screening algorithm and attribute weighted support vector selection algorithm are used to establish test model for training set. Apply the test model to test the test data set. Show the experimental comparison of modeling time in figure 3.

TABLE I. THE WEIGHTED EXPERIMENTAL RESULTS STATISTICS

	Before weighted	***After weighted***
Model time（s）	362.29	179.96
Detect time (%)	92.11	92.22

The number of support vectors used to set up the test model has decreased in different degrees before and after screening. It shows that the selection strategy is good to complete the data reduction. The properties of affecting the experimental results are highly weighted in the experiments. Emphasis on their role in classification, thus further improve classification accuracy. The calculation of kernel function avoids some weakly related or irrelevant characteristics. It can be seen from figure 1, figure 2, figure 3, and table 1 that the SVM algorithm is better than that of the non-weighted Euclidean distance and radial product kernel function in the model setting time.

The establishment time of the weighted model is significantly shortened. The weighted results show that the distribution of samples has its own characteristics. It's easier to break down from a lot of samples. The experimental results show that the accuracy of the test is improved. The overall performance of the test model is improved.

III. CONCLUSION

In the paper, a weighted Euclidean distance, the radial product kernel function and SVM algorithm are constructed to solve the problem of large data feature extraction and big data classification.

Effective reduction the data sample of establishing the test model through data filtering preprocessing. The number of support vector samples established the test model has decreased in varying degrees. The overall detection accuracy reached the highest after selecting and weighting. It is shown that the selection algorithm can effectively remove the non-significant interference samples and improve the generalization of the model. The results show that the class distribution of the weighted samples is more obvious. The training time is significantly shorter. The detection effect is further improved after weighting.

ACKNOWLEDGMENT

The This work is supported by Liaoning University of Science and Technology Reform and innovation of graduate education in 2016 -- the construction of excellent courses under contract number 2016YJSCX19.At the same time，the work is supported by Liaoning University of Science and Technology Reform of undergraduate education in 2017- under contract number XJGYB201705. The work is supported by Liaoning University of Science and Technology project of school talent in 2015- under contract number 2015RC03.

We would like to thank our colleagues and friends for their encouragement and moral support given to us. We

would also like to extend our thanks to our family members for all the help, direct or indirect.

REFERENCES

[1] Shan Wang, Liang Tan. Similar duplicated data cleaning in the Web big data environment[J]. Computer engineering and design，2017，38（3）：646~651.

[2] Li-Juan Gen, Xing-Yi Li. KNN algorithm for large data classification[J]. Computer application research，2014，31（5）：1342~1344，1373.

[3] Xue-Zu Li, Guo-Long Chen. Large data analysis and prediction based on intelligent least squares support vector machine[J]. Computer engineering,2015,41(6):38~42.

[4] Qing-He,Ning Li,Wen-Juan Luo,etc. Overview of machine learning algorithms in big data[J]. Pattern recognition and artificial intelligence，2014，27（4）：327~336.

[5] Ming-Wei Guo. An overview of target detection algorithms based on support vector machines [J]. Control and decision-making，2014，29（2）：193~200.

[6] Wei-Lin, Ting-Liu, Wei-Guo Lv. The boundary vector adjustment entropy function support vector machine research in big data [J]. Microelectronics and computers, 2016, 33(8): 149~152,157.

[7] Wei-Dong Jiao, Shu-Shen Lin. The whole improved fault diagnosis method based on support vector machine[J]. Journal of instrument and instrument, 2015，36(8):1861～1870.

[8] Guo-Jun Mao , Dian-Jun Hu, Song-Yan Xie. Large data classification model and algorithm based on distributed data flow[J]. Computer journal，2017，40（1）：161~175

[9] Guo-Jun Mao,Dian-Jun Hu,Song-Yan Xie. Large data classification model and algorithm based on distributed data flow[J]. Computer journal，2017，40（1）：161~175

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2010-11-25