Prediction of Data Classification Based on Support Vector Machine

Xinghui WU[a], Yuping ZHOU[b, *]

College of Information Science, Hainan Normal University, Haikou, 571158 China

[a]email:szfwxh@qq.com, [b]email:zaifengshi@163.com, [*]corresponding author

Keywords: Data Mining, Data Classification, Support Vector Machine

Abstract. Support vector machine is an emerging technique in data mining. By analyzing principle and classification algorithm of support vector machine, a prediction model based on support vector machines was given, and the effectiveness of support vector machines in classification prediction was demonstrated with the test results.

Introduction

Support vector machine (SVM) is a new type of machine learning methods developed based on statistical learning theory in the mid 1990 of the 20th century [1-6]. Support vector machine trains learning machines with the principle of structural risk minimization (SRM), resolves the problems such as nonlinear, high dimensionality, and local minima based on rigorous theoretical basis. It has become the new research focus in the area of machine learning after neural networks [7-12]. Until now, it is only a few years from being raised to widely being payed attention on support vector machines, there are a lot of problems unresolved or not fully resolved, so it has a lot of potential in applications. Therefore, support vector machine is a very worthy to study.

The SVM's biggest advantage is subject to structural risk minimization principle, has a good generalization ability. So when it is applied to deal with nonlinear problems, by the use of a kernel instead of the high-dimensional inner product space calculation, the nonlinear function can be changed into a high-dimensional linear problem, which could resolve the problem of dimensionality and local extrema [13-18].

Principle of Support Vector Machine

Support vector machine is developed from the optimal classification in linearly separable cases [19-21]. Its essence is to find a rule to divide space point Rd into two parts. Basic ideas can be illustrated in the following diagram of simple linear separable problem (Figure 1). It is to find an optimal classification hyperplane, and maximize the distance of two types of samples from the hyperplane. In Figure 1, two samples are marked as the black point and white point separately, H for optimal classification hyperplane, H1 and H2 respectively represent the sample nearest by H and surface parallel to H, distance between them referred to as classification interval (Margin). The so-called optimal classification hyperplane or maximal margin hyperplane is required not only to separate the two types rightly (training error is 0), but also to maximize the classification interval largest.

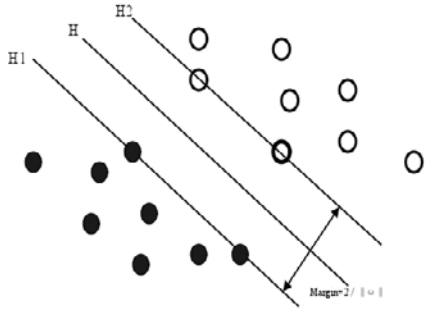

Fig. 1 Classification line under linear classification

In the case of linearly inseparable, based on Mercer nuclear expansion theorem, sample spaces can be mapped into a high-dimensional and even the infinite dimensional feature space (Hilbert space) with nonlinear mapping. Then, in the feature space, the linear learning method can be applied to deal with high nonlinear classification in the sample space.

Classification Algorithm of Support Vector Machine

Assumption a training set of size L as:

$$(x_i,y_i),\ x_i \in Rd,\ y_i \in \{+1, -1\},\ i \in \{1,, 2, ,,,l\} \tag{1}$$

It is made up of two categories, if $x_i \in Rd$ belongs to the first category, it is marked as positive (y_i=1), if not, marked as negative (y_i=−1). Support vector machine aimed at finding the classification hyperplane H:

$$\omega^T x+b=0 \tag{2}$$

To make the samples meet the need:

$$y_i\ (\omega^T x+b)-1 \geqslant 0,\ i \in \{1,,2, ,,,l\} \tag{3}$$

And then, this hyperplane can separate the points of two classes. It is to maximize the distance between the nearest points from the hyperplane in two classes.

Where, ω is the weight vector, b for bias. the vectors are called support vectors with optimal distance on both sides and the shortest plane distance.

The distance of hyperplane H1 to the origin is $|1-b|/\|\omega\|$, hyperplane $H2$ to the origin is $|-1-b|/\|\omega\|$. Therefore, the distance between the $H1$ and $H2$ is $2/\|\omega\|$, it is called classification intervals. So, to maximize classification interval is to minimize $\|\omega\|$.

The classification intervals can also discussed in the view of the VC dimension. In an N-dimensional space, to let samples distribute in the context of a hyper-sphere with radius of r, then the VC-dimension of index function set $f(x, w, b)$=sgn(<w, x>+b)(sgn is the sign function), which is formed by hyperplanes meeting the need of $\|\omega\| \leq a$ (A>0), meets the following bounds:

$$P \leq \min([R^2 A^2],\ N)+1 \tag{4}$$

To minimize the $\|\omega\|^2$ is to get the least upper bound of VC dimension, thereby to realize structural risk minimization.

Therefore, the optimal separating hyperplane can be resolved by quadratic programming as follow:

$$\text{Min } \frac{1}{2}\|\omega\|^2$$

Where, the bound is:

$$y_i\ (\omega^T x+b)-1 \geq 0,\ i \in \{1,,2, ,,,l\} \tag{5}$$

In the above algorithm, only the case of linearly separable was considered. Due to nonlinear problems can be changed into a high-dimensional linear problem by non-linear transformation, so for nonlinear classification, first using a non-linear map φ to maps the data to a high-dimensional feature space and then linear classification can be done in the high-dimensional feature space, the importing nonlinear classification is got when it is reflected back into space. In order to avoid complex calculations in a higher dimensional space, a kernel function k (x, y) is used in support vector machine instead of the inner product operation in the high dimensional space <$\Phi(x)$, $\Phi(y)$>.

In addition, considering that there may be some samples cannot be correctly classified by separating hyperplane, slack variables are used to resolve the problem, so it is optimized as:

$$\text{Min } \frac{1}{2}\|\omega\|^2+C\sum_{i=0}^{n}\xi_i \tag{6}$$

with bounds as:

$$y_i\ (\omega^T x+b) \geq 1-\xi_i,\ i \in \{1,,2, ,,,l\} \qquad \xi_i \geq 0, i \in \{1,,2, ,,,l\} \tag{7}$$

Where, C is a constant. In equation (7), the first item makes the distance between the sample and the hyper-plane as big as possible, so as to improve the generalization ability. the second item is to make classification error as small as possible.

Introducing Lagrange function:

$$L=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\alpha_i(y_i(<w, \Phi(x_i)>+b)-1+\xi_i)-\sum_{i=1}^{l}\gamma_i\xi_i \quad (8)$$

Where, $\gamma_i, \alpha_i \geq 0$, i=1,,2, ,,,*l*

Extrema of function L should meet the condition:

$$\frac{\partial}{\partial w}L=0,\ \frac{\partial}{\partial b}L=0,\ \frac{\partial}{\partial \xi}L=0 \quad (9)$$

As a result:

$$W=\sum_{i=1}^{l}\alpha_i y_i \Phi(x_i) \quad (10)$$

$$\sum_{i=1}^{l}\alpha_i y_i=0 \quad (11)$$

$$C-\alpha_i-\gamma_i=0,\ i=1,,2, ,,,l \quad (12)$$

Introducing (10)-(12) into equation (8), the dual form is optimized as:

$$\max\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \quad (13)$$

With bounds as:

$$\sum_{i=1}^{l}\alpha_i y_i=0 \quad (14)$$

$$C \geq \alpha_i \geq 0,\ i=1,,2, ,,,l \quad (15)$$

In general, most of the α_i is 0 is the characteristic of solutions, corresponding samples with other α_i are support vector.

According to the *KKT* conditions, at the saddle point:

$$\alpha_i[y_i(<w, \Phi(x_i)>+b)-1+\xi_i]=0,\ i=1,,2, ,,,l \quad (16)$$

$$(C-\alpha_i)\xi_i=0,\ i=1,,2, ,,,l \quad (17)$$

Then *b* can be calculated as follows:

$$y_i(\sum_{i=1}^{l}\alpha_j y_i K(x_j, x_i)+b)-1=0,$$

While $\alpha_i \in (0,C)$ (18)

Therefore, the value of b can be obtained through any one of the support vector. Resulting discriminant functions is as follows:

$$F(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_j y_i K(x_i, x)+b) \quad (19)$$

Prediction Model Based on Support Vector Machine

Design a support vector machine with good performance, model selection is the key. Model selection includes the kernel type selection and relative parameters selection while kernel function is determined.

A Selection of Kernel Function. The selection of kernel function is a core problem of the research on support vector machine theory, but at present there is no effective method for constructing a suitable kernel function for specific problems. In practice, by far the most common kernel functions are mainly as the following:

(1) linear function:

$$K(x,y)=x^T\cdot y \quad (20)$$

(2) polynomial function:

$$K(x,y)=[(x\cdot y)+1]^2 \quad (21)$$

(3) Gaussian radial basis function:

$$K(x, y) = e^{\frac{-\|xy\|^2}{2\sigma^2}} \quad (22)$$

(4)Sigmoid kernel function:

$$K(x,y)=\tanh\{v(x\cdot y)+c\} \quad (23)$$

In the above, Gaussian radial basis function RBF is widely used, it is a generic kernel function, through the selection of parameters, it can be applied to samples with any distribution. It has been

proved by more and more cases. RBF kernel function is the most widely used kernel function.

B Parameters Selection for Fixing Kernel Function. When the kernel function is determined, the corresponding kernel function parameters can be also determined. In this article, RBF is selected as kernel function, the corresponding parameter is (c, γ). Determination of kernel function parameters is turned into the optimization of (c, γ), it is to find the best combination of (c, γ).

Classification Prediction Examples Based on Support Vector Machine

A Characterization and Pretreatment of Data Sets. The experimental data sets of wine are from UCI database, which are physics and chemistry research data. 178 samples are involved in the data set, each sample contains 13 components (chemical composition), and is labeled separately. 50% of the 178 sample are considered as training set, and the other as a test set. Classification model can be obtained by training support vector machines with training set, and then to predict the category labels of test set using the model.

All data should be normalized before modeling prediction. It is to normalize all the values to [0,l] or [-1,l] range to reduce the complexity of numerical calculation in the process of training, as well as greater values control the training process. Sample standardized and mapping normalized are done by using the following formulas:

$$F{:}x\rightarrow y=\frac{x-x\min}{x\max-x\min} \tag{24}$$

In the formula, x for the sample data, $x_{\max}$ and $x_{\min}$ are the maximum and minimum value for sample data respectively, y for the normalized data. Where, x, $y\in R$n; $x_{\min}=\min(x)$; $x_{\max}=\max(x)$。 $y_i\in[0,1]$, i=1,2,l,…n.

B Training and Prediction. Train SVM classifiers with training set of train_wine, then to do label prediction of test set using the model, and finally the classification accuracy is 97.7528% (87/89) (classification). The main code is as follows:

```
model = svmtrain(train_wine_labels, train_wine, '-c 2 -g 1');
[predict_label]=svmpredict(test_wine_labels, test_wine, model);
```

Final classification results are shown in Figure 2.

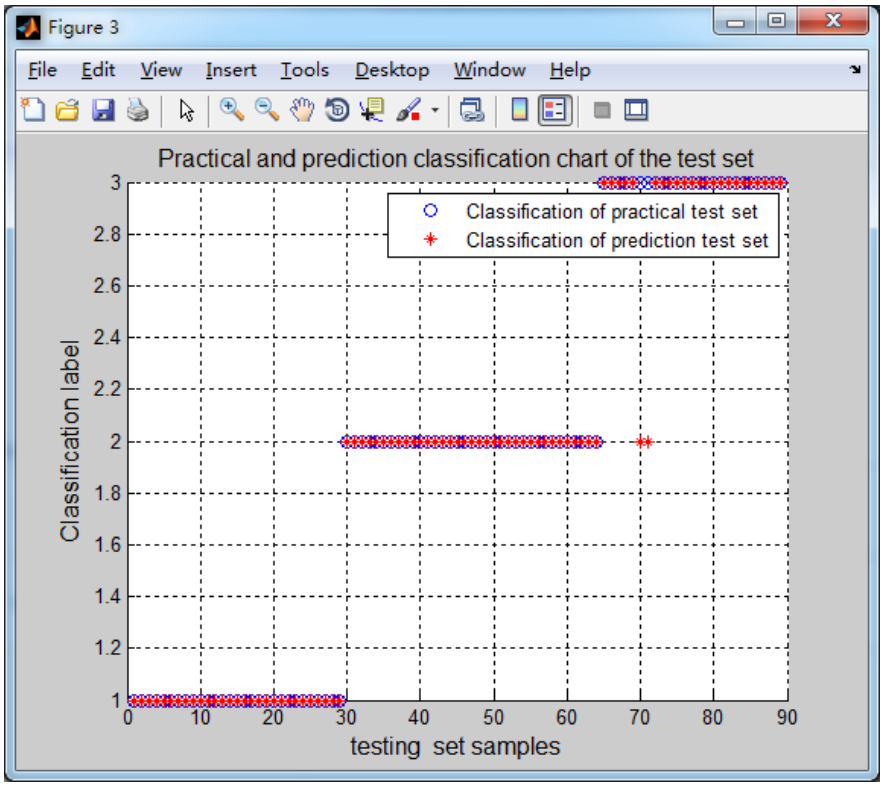

Fig. 2 Final classification results

Conclusion

In present article, classification prediction of the experimental data set of wine from UCI database were studied using support vector machines classification algorithm. Results indicated that classification algorithm of support vector machine is effective with good performance.

Acknowledgement

This work was financially supported by the Science & Technology Department of Hainan Province, P.R.China (20156242). Thanks are also given to colleagues in Hainan Normal University who have payed attentions to this article.

References

[1] H. Pei-Yi and L. Yen-Hsiu, “A New Multi-class Support Vector Machine with Multi-sphere in the Feature Space”, New Trends in Applied Artificial Intelligence, Springer-Verlag, 2007: 756-765

[2] S. Jak and J.Vandewalle, “Least Squares Support Vector Machines Classifiers”, Neural Processing Letters, 1999, 9(3): 293-300

[4] Z. Jingke and Z. Lixing, “Principal minimax support vector machine for sufficient dimension reduction with contaminated data”, Computational Statistics & Data Analysis, Vol. 94, Feb. 2016, pp. 33-48

[5] F. Kai, L. Jiangang and C. Jinshui, “Nonlinear model predictive control based on support vector machine and genetic algorithm”, Chinese Journal of Chemical Engineering, Vo. 23, Issue 12, Dec. 2015, pp. 2048-2052

[6] X. Shu-yin, XX. Zhong-yang and L. Yue-guo et al, “A method to improve support vector machine based on distance to hyperplane”, Optik - International Journal for Light and Electron Optics, Vol. 126, Issue 20, Oct. 2015, pp. 2405-2410

[7] K. Sangwook, Y. Zhibin and M. K. Rhee et al, “Deep learning of support vector machines with class probability output networks”, Neural Networks, Vol. 64, Apr. 2015, pp. 19-28

[8] F.J. Martínez López, S. Martínez Puertas and J.A. Torres Arriaza, “Training of support vector machine with the use of multivariate normalization”, Applied Soft Computing, Vo. 24, Nov. 2014, pp. 1105-1111

[9] H. Lisha, L. Shuxia and W. Xizhao, “A new and informative active learning approach for support vector machine”, Information Sciences, Vol. 244, Sep. 2013, pp. 142-160

[10] H. Chia-Hui, “A reduced support vector machine approach for interval regression analysis”, Information Sciences, Vol. 217, Dec. 2012, pp. 56-64

[11] I. Rodriguez-Lujan, C. S. Cruz and R. Huerta, “Support vector machine”, Pattern Recognition, Vol. 45, Issue 12, Dec. 2012, PP. 4414-4427

[12] E. A. Zanaty, “Support vector machine (SVMs) versus Multilayer Perception (MLP) in data classification”, Egyptian Informatics Journal, Vol. 13, Issue 3, Nov. 2012, pp. 177-183

[13] S. Ekici, “Support vector machines for classification and locating faults on transmission lines”, Applied Soft Computing, Vo. 12, Issue 6, June 2012, pp. 1650-1658

[14] L. Shih-Yen, G. Ruey-Shiang and S. Yeou-Ren, “Effective recognition of control chart patterns in autocorrelated data using a support vector machine based approach”, Computers & Industrial Engineering, Vol. 61, Issue 4, Nov. 2011, pp. 1123-1134

[15] S. T. John and S. Sun, “Support vector machines”, Neurocomputing, Vo. 74, Issue 17, Oct.

2011, pp. 3609-3618

[16] F. L. Heng and I. Dino, “Feature selection for support vector machine -based face-iris multimodal biometric system”, Expert Systems with Applications, Vol. 38, Issue 9, Sept. 2011, pp. 11105-11111

[17] K. Gayathri and N. Kumarappan, “Accurate fault location on EHV lines using both RBF based support vector machine and SCALCG based neural network”, Expert Systems with Applications, Vol. 37, Issue 12, Dec. 2010, pp. 8822-8830

[18] W. Qi and R. Law, “Fuzzy support vector regression machine with penalizing Gaussian noises on triangular fuzzy number space”, Expert Systems with Applications, Vol. 37, Issue 12, Dec. 2010, pp. 7788-7795

[19] Y. Eunseog, L. Koenig and M. K. Jeong, et al, “Support vector -based feature selection using Fisher’s linear discriminant and support vector machine”, Expert Systems with Applications, Vol. 37, Issue 9, Sept. 2010, pp. 6148-6156

[20] C.-Y. Chang, S.-J. Chen and M.-F. Tsai, “Application of support vector machine -based method for feature selection and classification of thyroid nodules in ultrasound images”, Pattern Recognition, Vol. 43, Issue 10, Oct. 2010, pp. 3494-3506

[21] A. Al-Anazi and I. D. Gates, “A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs”, Engineering Geology, Vol. 114, Issues 3–4, Aug. 2010, pp. 267-277