ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 3 (2013) No. 2, pp. 29-34

Analysis and prediction of logistics enterprise competitiveness by using a real GA-based support vector machine

Ning Ding[1], Hanqing Li[2], Hongqi Wang[2]
1 School of human and development, China agriculcure University, Haidian District, Beijing 100083, China
[2]School of Economics and management, Beijing Jiaotong University,100044 Beijing, China
517276872@qq.com

Abstract: This research is aimed at establishing the forecast and analysis diagnosis models for competitiveness of logistics enterprise through integrating a real-valued genetic algorithm to determine the optimum parameters and SVM to perform learning and classification on data. The result of the proposed GA-SVM can satisfy a predicted accuracy of up to 95.56% for all the tested logistics enterprise competitive data. Notably, there are only twelve influential feature included in the proposed model, while the six features are ordinary and easily accessible from National Bureau of Statistics. The proposed GA-SVM is available for Objective description Forecast and evaluation of a logistics enterprise Competitiveness and stability of steady development.

Keywords: SVM, Logistics Enterprise Competitiveness, GA, Forecast

1. Introduction

Chinese logistics enterprises in recent years have just started to develop. The regulations about logistics enterprise began to implement until 2005 May 1st In China(Wang & Zeng,2008). Therefore, it is necessary to establish a feasible evaluation index system in order to evaluate the development level of Chinese logistics enterprises.

Xie(2009) pointed out that, in order to survive in an increasingly competitive marketplace, many companies are turning to data mining techniques for churn analysis. Mahesh Pal(2012) has analyzed the SVM were state of art classification algorithms and perform well in terms of classification accuracy in

comparison to multinomial logistic regression based classification algorithm as well as other classifier for land cover classifications. Liu and Li compared logistic regression (LR), probabilistic neural network (PNN) and support vector machine (SVM) classifiers for discriminating between normal and PD subjects in assessing the effects of DBS-STN on ground reaction force (GRF) with and without medication.

2. Brief Description of the Research Method

2.1. Support Vector Machine

SVM, proposed by Vapnik (1995), was mainly used to find out a separating hyperplane to separate two classes of data from the given data set(Xie & Li,2009). Let each entry of data be $(x_i, y_i), (i = 1,2,3,...,n), x \in R^d, y \in \{-1,+1\}$, and x_i is input data, y_i represents category, demotes the sample quantity, and demeans the input dimension. For any x_i on the separating hyperplane, the condition (1) Should be satisfied. As usual, (2) Denotes the decision functions, where w is the normal vector of the hyper plane and means the bias value. For any given entry of test data, if $f(x) \geq 0$, the entry of data could be classified as "+1"; if $f(x) < 0$, it could be classified as "−1". SVM could be classified as liner or non-linear based on the problem types (Burges, 1998). When data could be categorized as two types, the linear SVM could find a hyper plane with the maximum margin width $\frac{1}{2}\|w\|$ to separate the data into different types by finding the minimum of $\frac{1}{2}\|w\|^2$ subject to (3) For solving the above problem, the Lagrange optimization approach could be adopted for carrying out the resolution process easily. Constraint Eq.(3) could be replaced by Lagrange multipliers. The Lagrange function is expressed as (4) where the Lagrange multiplier $a_i \geq 0, i = 1,2,\cdots,$ n , corresponds to each inequality with the constraint Eq.(3). As such, the original problem to find the minimum $\frac{1}{2}\|w\|^2$ has been converted into finding the minimum L_p with the constraint equation $a_i \geq 0$. However, it is still difficult for the non-linear SVM to find the optimal solutions. For dealing with this situation, the Lagrange Dual Optimization Problem is used to make the solution process easier, which is formulated (5).

$$w \cdot x + b = 0 \quad (1)$$

$$f(x) = w \cdot x + b \quad (2)$$

$$y_i[(w \cdot x) + b - 1] \geq 0, i = 1,2,\cdots,n \quad (3)$$

$$L(w,b,a)=\frac{1}{2}(w\cdot w)-\sum_{\mathrm{i}=1}^{\mathrm{n}}a_i\{\mathrm{y_i}[(\mathrm{w}\cdot\mathrm{x_i})+\mathrm{b}]\}+\sum_{i=1}^{n}\alpha_i \tag{4}$$

$$MaxL_D(\alpha)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j\mathrm{y_i}\,\mathrm{y_j}(x_i\cdot\mathrm{y_j}) \tag{5}$$

$$\sum_{i=1}^{n}\alpha_i y_i=0\ \alpha_i\geq 0, i=1,2,\cdots,\mathrm{n} \tag{6}$$

$$\mathrm{y_i}[(\mathrm{w}\cdot\mathrm{x_i})+\mathrm{b}-1]+\xi_i\geq 0,\ \xi_i\geq 0,\mathrm{i}=1,2,\cdots,\mathrm{n} \tag{7}$$

$$Min\ \ \frac{1}{2}\|w\|^2+C\left|\sum_{i=1}^{n}\xi_i\right|, c>0 \tag{8}$$

$$MaxL_D(\alpha)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j\mathrm{y_i}\,\mathrm{y_j}(x_i\cdot\mathrm{y_j}) \tag{9}$$

$$\sum_{i=1}^{n}\alpha_i y_i=0 \tag{10}$$

$$\iint k(xi,xj)g(x_i)g(x_j)dx_i dx_j>0, g\in L_2 \tag{11}$$

Once α_i is found, the optimum w and b can be obtained and therefore the decision function $f(x,a,b)$ can be determined through Eq. (4). In addition, for the linear SVM to process non-separable data, Vapnik (1995) indicated that slack variables ξ_i could be added into the constraints as (4) and (7).

When errors happen to the classification of training data, ξ_i should be larger than 0. Therefore, a lower $\sum\xi$ should be preferred when determining the separating hyper plane. For this purpose, a cost parameter $c>0$ is added to control the allowable error ξ_i. The objective function should be changed from the solution for the minimum $\frac{1}{2}\|w\|^2$ into the solution for (8).For simplification, Eq.(8) could be transformed into the dual problem as (9). Based on above descriptions, it is simple to use linear SVM to separate the two different categories of data, if data could be fully separated by a linear function; otherwise, a parameter C is required to control the allowable errors. However, in the real world, not all data could be separated by linear hyper plane. Boser, Guyon, and Vapnik (1992) made the comparison between the linear and non-linear problems, and found if the original data are transferred to another feature space of high dimension (ϕ :Rd → F) through the mapping function ϕ, thereafter, the linear classification was conducted within the space and the

process could find better effect. If data (x_i, x_j) are transferred to the feature space of a high dimension, i.e., $(\phi(x_i) \cdot \phi(x_j))$, the corresponding term in the dual problem (6) should be changed. The dot product of $\phi(x_i)$ and $\phi(x_j)$ was defined by the Kernel function, $K(x_i, x_j)$, thus, the optimization by the linear or non-linear SVM finally becomes $Max\ L_D(\alpha) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \square x_j)$

Subject to (10).

2.2. Genetic Algorithm

GA coding strategies mainly include two sectors; one sector recommends the least digits for coding usage, such as binary codes; another one recommends using the real-valued coding based on calculation convenience and accuracy (Haupt & Haupt,1998). Binary codes are adopted for the decision variables in solving the discrete problems; however, it probably causes conflict between accuracy and efficiency when the problems are featured with continuity in that the calculation burden is quickly increased. The adoption of real-valued coding does not only improve the accuracy, but also significantly increases the efficiency in the larger search space, so that more practical applications adopt the real-value coding for solution.

2.3. GA-SVM Model Discriminant Analysis

As mentioned before, a kernel function is required in SVM for transforming the training data. This study adopts RBF as the kernel function to establish support vector classifiers, since the classification performance is significant when the knowledge concerning the data set is lacking. Therefore, there are two parameters, C and δ2, required within the SVM algorithm for accurate settings, since they are closely related to the learning and predicting performance. However, determining the values exactly is difficult for SVM. Tay and Cao (2001) Suggested that C should range from 10 to 1000 and δ^2 from 1 to 100, so that the established models can achieve much better results. Generally, to find the best C and δ^2 a given parameter is first fixed, and then within the value ranges another parameter is changed and cross-comparison is made using the grid-search algorithm. This method was conducted with a series of selections and comparisons, and it will face the problems of lower efficiency and inferior accuracy when conducting a wider search. However, GA for reproduction could provide the solution for this study. The scheme of an integration of real GA and SVM is shown in Fig. 1 to establish a classification model that could be used to determine whether a business crisis is approaching.

The final converged solutions should be affected by the possibility of genetic

operations or parameters For example, the possibility for mutation and population size was determined after several experimentation cycles to make the training data sets have the maximum accuracy.

3. Design of the Diagnostic Model

This article builds the appraisal model of the finance of listed companies through empirical method. The model includes two processes: the training of the model and the testing of the model. If the model which has undergone the two processes has a good recognition capability, it will be open to specific finance appraisals to further prove its validity. The structure of the model shows in fig. 1.

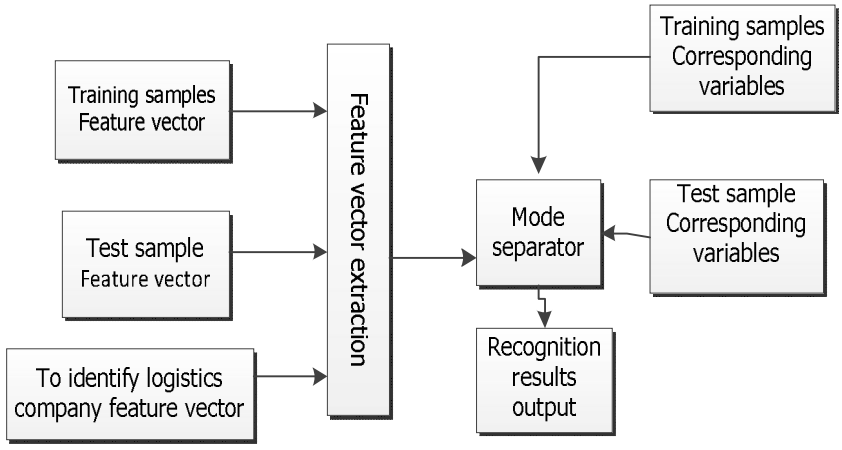

Fig.1: Competitiveness evaluation model of listed logistics companies

In the graph, black arrows represent the direction of data flow of training and testing samples; white arrows represent the direction of data flow in the actual finance appraisal process. From the graph 1, it can be concluded that the first step of building the model is to build the training sample and testing sample, and train the pattern classifier after the input of training sample. When the training is complete, the result can be confirmed by the input of training sample. Pattern classifiers which only pass the test can be used to appraise the finance condition of the listed companies. If the finance condition of the listed companies complies well with the results from the pattern classifiers, the model is proved to be successful.

4. Conclusion

This paper object of study is the finance data of logistics enterprises. Through the empirical method attracted several conclusions as follows:

1) The finance data of logistics enterprises in our country is effective. It has strong ability to predict.

2) Through an empirical test of sample and out sample, the model shows the logistics enterprises competitive power.

References

Liu, S. & Li, Y. Y.(2007).Parameter selection algorithm for support vector machines based on adaptive genetic algorithm.*Journal of Harbin Engineering University*,28(4),398-402.

Pal, M.(2012).Multinomial logistic regression-based feature selection for hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*,14(1), 214-220.

Wang, Y. & Zeng, L. B.(2008). Design on the evaluation index system of logistics enterprises' competitiveness in china. *Economy and Management,*22(11),54-57.

Xie, Y. Y. & Li, X.(2009).Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*,36(3),5445-5449.