ISSN : 0974 - 7435

Volume 10 Issue 11

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(11), 2014 [5408-5414]

City innovation capability evaluation method based on support vector machine

Yong-li Zhang[1], Yan-wei Zhu[2]

[1]Qinggong College, Hebei United University, Tangshan 06300, (CHINA)

[2]Department of Mathematics and Information science, Tangshan Normal University, Tangshan, (CHINA)

ABSTRACT

China will develop into an innovative country in 2020. It has become an important topic that study on evaluation method of innovation ability. But the science and technology innovation capacity determination is complex, there are many factors affecting the innovation ability, there are a non-linear relationship, uncertainty and ambiguity. Support vector machine is a statistical learning method based on small samples, using structural risk minimization principle, and it is good generalization ability. This paper uses support vector regression algorithm to evaluate the ability of innovation of science and technology, get the support vector machine regression model, Through the 2013 yearbook data analysis of the experimental results, this method is achieved very good results in evaluation of regional innovation capacity.

KEYWORDS

Innovation ability; Support vector machine; Two regression.

INTRODUCTION

China has plans to be an innovative country in 2020. As the key point of building an innovation oriented country and the construction of regional innovation system, enhance the construction of innovative city on the overall innovation capability in China plays an important role, Tangshan City is one of the national innovation pilot city, Tangshan has made some achievements in the exploration of innovation of science and to promote the development of economy and realize the transformation and development of traditional resources type city and industrial town. In order to scientific and technological innovation capability development process inspection, monitoring and evaluation of the city, it is an important research direction of evaluation index system and evaluation method of the effective. It is complex to determine the technology innovation ability. There are many factors to influence and reaction innovation ability. There is a non-linear relation ship between each factor and uncertainty and ambiguity. These are barrier to evaluation on the innovation capability.

About evaluation of the innovation ability of science and technology, To determine the weight of indexes are evaluated by using AHP and entropy method in reference[1], Factor analysis and fuzzy C means clustering method (FCM) evaluate the ability of innovation of science and technology in reference[2]. But these traditional methods can not solve the complex nonlinear relationship between the evaluation factors and innovation ability level, the utility function, the weights of evaluation in the process of manually, limits the universality evaluation mode, also affect the reliability of the results. Evaluation method of literature by[1-4] neural network, can overcome the deficiencies of these methods, but there are three very difficult to overcomes[5]: slow training speed, difficult to meet the many requirements prior learning situations; the learning process is very easy to converge to local minima, the learning accuracy is difficult to guarantee; network generalization ability is poor, can not guarantee the sample after training on the training set with network the good application effect. Support vector machine method can overcome the above all kinds of lack, which is developed in recent years based on the statistical learning theory. It will provide a new way of innovation ability evaluation. Support vector machine was first proposed by Professor Vapnik and his collaborators, due to its excellent properties, this method has aroused the concern of many researchers. Support vector machine is mainly used in pattern recognition and function fitting. Support vector machine is a statistical learning method based on small samples, using structural risk minimization principle, and has good generalization ability[6]. This paper uses support vector regression algorithm to evaluate the ability of innovation of science and technology, get the support vector machine regression model, the results show that the method is feasible and effective.

THE BASIC THEORY OF SUPPORT VECTOR REGRESSION MACHINE

Support vector machine is proposed by Boser, Guyon and Vapnik, and for the first time in computational learning theory (COLT) has been proposed in 1992 Conference Papers. The samples from the low dimensional feature space was mapped to high dimension space, the nonlinear problem can be transformed into a linear separable problem solving. It can solve the structural problems of finite sample of high dimensional model. it become an important tool for solving nonlinear problems. Support vector regression is developed on the basis of support vector machine. The following first introduce support vector machine two class classification machine[7].

Support vector machine classification machine

Suppose $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ is a sample set, where $x \in R,\ y \in \{1,-1\}$ let :

$$\begin{array}{ll} w\cdot x_i+b\ge 1 & y_i=1 \\ w\cdot x_i+b\le -1 & y_i=-1 \end{array} \qquad (1)$$

w is a classification hyper plane vector method, $\frac{|b|}{\|w\|}$ is the distance form origin to the hyper plane[8].

(1) can be expressed as the merger : $y_i(w\cdot x_i+b)\ge 1$. the hyper plane $H: w\cdot x_i+b=0$. origin to the hyper plane $H_1: w\cdot x_i+b=1$ distance is $\frac{|b-1|}{\|w\|}$, origin to the hyper plane $H_2: w\cdot x_i+b=-1$ distance is $\frac{|b+1|}{\|w\|}$. In the hyper plane H_1 or H_2 points are called support vector. Class interval is $\frac{2}{\|w\|}$,it is distance between two hyper plane H_1 and H_2. The between class distance is greater, the better the classification capabilities of machine. The whole learning process can be considered as the maximum $\frac{2}{\|w\|}$, which will minimize $\frac{\|w\|^2}{2}$.

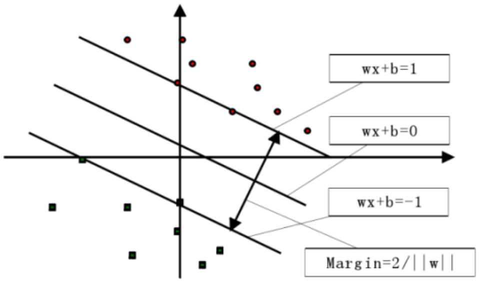

Figure 1 : The optimal separating hyper plane is the one with maximal margin from the data in two dimensional space

SVM expression will be :

$$\begin{array}{l} \min \Phi(w,b)=\frac{1}{2}\|w\|^2=\frac{1}{2}\left(w\cdot w\right) \\ s.t. \quad y_i\left[\left(w\cdot x_i\right)-b\right]\ge 1, i=1,2,\cdots,n \end{array} \qquad (2)$$

Where n is the number of training samples.

The problem described above is called the original problem. The problem can be solved used two times planning. Usually can use the Lagrange method, the original problem is converted into a dual problem [9-11] :

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i x_j)-\sum_{i=1}^{l}\alpha_i \qquad (3)$$

$$s.t. \quad \sum_{i=1}^{l}\alpha_i y_i=0 \quad \alpha_i\ge 0 \quad i=1,2,\cdots,l$$

To solve the above expression, get $\alpha^* = \left(\alpha_1^*, \cdots, \alpha_l^*\right)^T$, then $w^* = \sum_{i=1}^{l} y_i \alpha * x_i$, let α^*>0 component. For example α_j^* corresponding sample $\left(x_j, y_j\right)$(This is a sample of support vector)[12],Then the expression $b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* \left(x_i \cdot x_j\right)$, count $w$and b, Ultimately determine the optimal hyper plane $\left(w^* \cdot x\right) + b^* = 0$,

Get the decision function $f(x) = \operatorname{sgn}(w^* \cdot x + b^*)$ or $f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i^* y_i (x \cdot x_i) + b^*\right)$.

Support vector regression hyper plane

Support vector regression is proposed based on two classification problems. A linearly separable training sets, there is a plane:

$$y_i = (w \cdot x_i) + b \tag{4}$$

The training samples are divided into two types by fitting. One is the training samples: $y_i - (w \cdot x_i) + b > 0$,The other is to satisfy the $y_i - (w \cdot x_i) + b < 0$ training samples. Assuming that there are no training samples: $y_i - (w \cdot x_i) + b = 0$.

If (4) is the two class of training samples to establish maximum interval hyper plane, the hyper plane can be used to solve the problem of support vector regression. It is to look for the support vector regression machine. It is reduced to a convex optimization problem of minimizing a linear inequality constraint two functions[5,6,13].

That is: given a linearly separable training samples $S = (y_1, x_1), \cdots, (y_l, x_l)$, to solve the problem of optimization

$$\begin{cases} \text{Minimise} & (w \cdot w) \\ \text{Subject to} & ((w \cdot x_i) + b) - y_i < 0 \\ & y_i - ((w \cdot x_i) + b) < 0 \\ & i = 1, \cdots, l \end{cases}$$

draw into Lagrange Multiplier $a_i, a_i^* > 0$, Lagrange function optimization problem is

$$L(w, b, a) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{l} (a_i - a_i^*)[y_i - (w \cdot x_i) + b]$$

Solving the optimization problem, $a_i, a_i^* \neq 0$ These vectors are support vectors. Support vector regression linear separable problem is:

$$f(x) = \sum_{i=1}^{l} (a_i^* - a_i)(x_i \cdot x) + b$$

When the training sample is not linearly separable, from the low dimensional space was mapped to high dimension space, transformed into the linear separable problem solving [14].

According to the Mercer theorem, the definition of kernel function:

$K(x_i \cdot x) = (\varphi(x_i) \cdot \varphi(x))$, By introducing the kernel function, support vector machine :

$$f(x) = \sum_{i=1}^{l} (a_i^* - a_i) K(x_i \cdot x) + b$$

SUPPORT VECTOR MACHINE MODEL OF INNOVATIVE CITY EVALUATION AND ITS APPLICATION

The accuracy of the evaluation results of regional innovation ability is affected by Scientific, objectivity and rationality of the design on the evaluation index system of innovative city.

The evaluation index system of innovative city

Regional innovation ability refers to the ability of an area. Knowledge is transformed into new products, new processes, new services, which can be used to evaluate the regional innovation system is appropriate and effective tool. The evaluation index system of Chinese regional innovation ability was studied. The present researches about regional innovation ability of discussion and research were summarized. The monitoring index system is designed. It includes the innovation environment, innovation resources, innovation, innovation output and innovation effect of 5 first level indexes and 53 two level indexes.

Selection of experimental data

This paper comes from the rankings in 2013 Chinese city innovation ability of randomly selected 30 as the experimental data. In order to facilitate the experimental and data collection, 32 indexes second level were selected from 53 second level indicators.

These data come from the People's Republic of China National Bureau of Statistics web site Chinese statistical yearbook 2013[15], Chinese statistical yearbook on science and technology of 2013[16] and the national database. There are 30 cities, each city has 32 second level index. The original values of 32 second level indexes were normalized.

Following formula:

$$\boldsymbol{y_{ij}} = \frac{\boldsymbol{x_{ij}} - \min \boldsymbol{x_{ij}}}{\max \boldsymbol{x_{ij}} - \min \boldsymbol{x_{ij}}} \qquad \textbf{(5)}$$

Where $\boldsymbol{i} = 1, \cdots, 30;\ \ \boldsymbol{j} = 1, \cdots, 32$.

20 cities were randomly selected as training data, the remaining 10 cities as the testing sample. The data is trained by SVM. Matlab2012b software and Lssvmlab1.80 tool are used.

Lssvmlab1.80 software is used to predict the test sample. In order to prediction method using support vector machine to compare to other methods, the training sample is regressed by MATLAB (consider cross impact index system)It can also predict the test sample values. Prediction of the two methods and the error values obtained were compared, as shown in TABLE 1.

TABLE 1 : Comparison of SVM regression method and two regression forecast method

true value		SVM		two regression	
		Predicted values	Error	Predicted values	Error
Dongguan	97.4891	97.4879	0.0012	97.5239	-0.0348
Wuxi	92.8944	92.8936	0.0008	92.8066	0.0878
Jiaxing	83.1787	83.1798	-0.0011	83.3528	-0.1741
Huangshan	76.8226	76.8217	0.0009	76.7972	0.0254

Huzhou	74.2984	74.2979	0.0005	74.298	0.0004
Yantai	73.5283	73.5275	0.0008	73.4969	0.0314
Baoding	66.5678	66.5704	-0.0026	66.5834	-0.0156
Panjin	65.4133	65.4127	0.0006	65.4863	-0.073
Tangshan	62.461	62.4605	0.0005	62.4521	0.0089
Fushun	62.2677	62.2668	0.0009	62.2599	0.0078

The residual sum of squares is 0.04657438 based on support vector regression. The residual sum of squares is 0.00001317 based on two regressions. SVM prediction error is smaller than two regression prediction. The above analysis shows that, the overall effect of prediction model SVM regression is better than multiple regression models.

CONCLUSIONS

Support vector regression model was established using the principle of support vector machine. The 10 test samples were predicted and there is very good results. Compared with the two regression forecast results, prediction method of support vector regression accuracy is very high. In the evaluation of innovative city this method of application is successful. In addition, the method is extended; it can be applied in many similar problems.

There are the following characteristics of Support vector machine: It has the advantages of simple structure, training and learning the unsolved classification, function fitting, easy application and good flexibility. Through the choice of different kernel function and parameters SVM can obtain different characteristics and properties of machine learning. Especially from the theory there is the stronger generalization ability. Model complexity is independent of the dimensionality of the input data. To avoid the curse of dimensionality, it is suitable for high dimensional problems. The problem can be transformed into an optimization problem. In theory uniqueness of solution is existed. It avoid falling into local minima. Looking for support vector finite model is selected in the input data of neural network computation, than the whole sample iteration speed.

Research on support vector machine has attracted more and more attention. In data mining, pattern classification and regression are the basic research contents, regression occupy a very important position. In the computer age regression has been from the traditional statistical penetration into the control theory, prediction, mathematical modeling, pattern recognition, physics, cosmology, life science and economics and social science and other fields. Because the support vector regression machine has good nonlinear characteristics and generalization, we can solve these problems in the field; it will play a more important role in these areas.

ACKNOWLEDGMENT

This work is supported by Tangshan City science and technology research and development program (No.2013CX-1), and Research on science and technology to guide the project of higher education of Hebei Province (No.Z2011328).

REFERENCE

[1] Yan Wang; To establish the evaluation system of regional science and technology innovation ability [J]. Statistics and Management, **08**, **(2013)**.

[2] Xiaoyun Du; The research on innovation system of science and technology of the 17 cities Shandong Province [D]Shandong University of Finance and Economics, **(2013)**.

[3] Xiangwei Kong; Based On Support Vector Machine The Study on Innovative Capacity Evaluation Index System Of Household Appliances Enterprises In China [D]. Beijing Jiaotong University, **(2009)**.

[4] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, Sayan Mukherjee; Choosing Multiple Parameters for

Support Vector Machines[J]. Machine Learning, 1-3 **(2002)**.

[5] V.N.Vapnik; Statistical learning theory. New York : Springer-Verlag, **(1995)**.

[6] C.Cortes, V.Vapnik; Support vector networks. Machine Learning, **20(3)**, 273-297 **(1995)**.

[7] Deng Naiyang; Support vector Machine Theory, algorithms and Development. Beijing: Science Press, (in Chinese), 176 **(2009)**.

[8] Yongli Zhang, Yanwei Zhu, Shufei Lin, Xiaohong Liu; Application of Least Squares Support Vector Machine in Fault Diagnosis. International Conference on Information Computing and Applications (ICICA2011).

[9] Zhang Zhancheng, Wang Shitong, Deng Zhaohong, Chung Fuli; A fast decision algorithm of support vector machine. Control and Decision, (in Chinese), **3**, **(2012)**.

[10] Tzong-Huei Lin; A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models[J]. Neurocomputing, **16**, **(2009)**.

[11] Z.C.Yang; User-Online Load Movement Forecasting for Social Network Site Based on BP Artificial Neural Network, Journal of Computers, **8**, 3176-3183 **(2013)**.

[12] Yuan Ping; Research on Clustering and Text Categorization Based on Support Vector Machine [D]. Beijing University of Posts and Telecommunication, **(2012)**.

[13] Julia Neumann, Christoph Schnörr, Gabriele Steidl; Combined SVM-Based Feature Selection and Classification[J]. Machine Learning, 1-3 **(2005)**.

[14] V.Vapnik write, Zhangxuegong translation; The nature of statistical learning theory[M]. Beijing: Tsinghua University Press, **(2000)**.

[15] National Bureau of Statistics of the People's Republic of China. Chinese statistical yearbook2013 [M]. China Statistics Press, **11**, **(2013)**.

[16] National Bureau of Statistics of the People's Republic of China. Chinese statistical yearbook on science and technology 2013[M]. China Statistics Press, **11**, **(2013)**.