ICIC Express Letters
Part B: Applications
Volume **8**, Number **5**, May **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 889–893

THE PREDICTION MODEL FOR COLLEGE ADMISSION SCORE BASED ON SUPPORT VECTOR MACHINE

Cheng Li[1,2], Zhisheng Ma[3], Honglie Zhang[1] and Yanju Liu[3]

[1]College of Computer and Control Engineering
[3]Modern Education Technology Center
Qiqihar University
No. 42, Wenhua Street, Qiqihar 161006, P. R. China
{ lcrb406; zhang_honglie; liuyanjuzhb }@163.com; liuyanjv@tom.com

[2]College of Computer Science and Technology
Harbin Engineering University
No. 145, Nantong Street, Nangang District, Harbin 151000, P. R. China

Received October 2016; accepted January 2017

Abstract. *The college admission score prediction is with the high difficulty, the low accuracy and few related researches, because of its high randomicity and its multiple influencing factors. Aiming at the situation, based on the support vector machine (SVM) theory, the support vector regression (SVR) is constructed to carry out the preliminary prediction for the college admission score, which is combined with the admission mode of Chinese college entrance examination in the paper. The simulation experiments are carried on by Libsvm-3.21 toolkit of Matlab R2010b software. By the quantitative analysis of the prediction results, the ideal prediction effect is achieved, and the mean absolute prediction error is 5.6. The change tendency of the prediction score is correct, and the feasibility of the support vector machine prediction model is verified for the college admission score prediction.*

Keywords: Support vector machine, Support vector regression, Quantitative analysis, Prediction model

1. **Introduction.** The college entrance examination is the main way for most senior middle school students to enter their ideal universities and colleges in China, and it has also been one of social focuses. In general, for the examinee, the experience from the teacher and the parents and their own comprehension is mostly relied on when the volunteer intention for the college is filled in. However, actually, it is fairly inaccurate to solve the issue only on the basis of human experience. On the one hand, the resources from the teacher and the parents are very limited, and it is impossible to carry out the extremely comprehensive analysis; on the other hand, there exist a list of the subjective factors when the teacher and the parents are faced with the historical situation of the college admission. Therefore, how to select the suitable college for their examination scores has been a hot and difficult problem.

The lowest control score line for all the provinces is determined by the number of examinees, according to the enrollment plan and the examinees' score. Generally, the enrollment number is determined slightly more than the plan number, whose times is 1.1-1.2. The score of the examinees is arranged according to the score ranking in the province. The province's lowest control line is determined when the plan number is reached, and only the examinee who has reached the score is eligible to participate in the college admission. The standard of the admission score line for all the colleges is designated basically the same way, which is determined according to the examinee score and the enrollment number of that year for that college admission. Therefore, the enrollment is nearly ended up when the lowest control score line for all the provinces is released.

The college admission score has been impacted on by many factors, including the enrollment policy, the enrollment scale, the number of applicants, the examinee quality, the score difference over the years, the school popularity, and so on, in which there exists the obvious randomness to some extent. At present, the related researches are divided into three categories: the examinee college entrance examination score prediction [1], the college enrollment situation prediction [2,3], the decision support system of the volunteer intention for the college [4,5]. The related technologies are adopted, including the neural network [6], the gray model [7,8], the decision tree [9,10], and support vector machine [11]. Comparatively, the advantage of support vector machine is to be adopted under the condition of the small sample. In the paper, based on the SVM theory, the prediction model for the college admission score is constructed to carry out the preliminary prediction, which is combined with the admission mode of Chinese college entrance examination.

The remainder of the paper is organized as follows. The SVM prediction model is constructed in Section 2. Then, the simulation experiments are carried out, and the experimental results are demonstrated and analyzed in Section 3. Finally, conclusions are given with the importance and the practical value of the prediction algorithm.

2. **SVM Prediction Model Construction.** Support vector machine is the machine learning algorithm, which is proposed by Vapnik et al. Its basic idea is that the kernel function is used to input the sample space and mapped to the high-dimensional feature space by a new type of machine, which is based on the statistical learning theory [12]. In the high-dimension space, an optimal classification plane is got, and the nonlinear relationship between input and output variables is obtained then. The support vector machine algorithm is a convex quadratic optimization problem, and the solution is the global optimal solution. It is suitable for solving the small sample and the nonlinear and high-dimensional pattern recognition problem. And it can be applied to other machine learning problems, such as the function fitting.

The college admission score prediction can be regarded as the calculation of result y by the given dependent variable x, and thus desirable ε-SVR is selected to be the prediction model of the paper. The value of loss function ε is set as 0.01. The model is constructed in detail as follows.

The given training sample is set, which is shown in Formula (1).

$$D = \{(x_1, y_1), \cdots, (x_l, y_l),\ x_i \in R^n,\ y \in R,\ i = 1, \cdots, l\} \tag{1}$$

The regression function $f(x) = w \cdot \phi(x_i) + b$ is searched to fit these training points for obtaining the minimum of $R[f]$, where w is the coefficient vector, and b is the constant. The formula of $R[f]$ is expressed as follows.

$$R[f] = \int c(x, y, f) dP(x, y) \tag{2}$$

where $P(x, y)$ is some kind of probability distribution of the generated sample. $P(x, y)$ is unknown and the minimum of Formula (2) cannot be solved directly. Therefore, the optimal regression function is inverted into solving the minimum of Formula (3).

$$\Phi(w, \xi, \xi^*) = \frac{1}{2}(w, w) + C\frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{3}$$

where C is the penalty factor, and ξ, ξ_i are the maximum and minimum limit of slack variable. The constraint conditions are increased, in which the error between the observed value y of point x and the predicted value $f(x)$ is not over the smaller given positive ε.

And Formula (3) is inverted into the optimal problem, which is shown as follows.

$$\begin{cases} \min\limits_{w,\xi_i,\xi_i^*,b} & \frac{1}{2}(w\cdot w)+C\cdot\frac{1}{l}\sum\limits_{i=1}^{l}(\xi_i+\xi_i^*) \\ \text{s.t.} & (w\cdot\phi(x_i)+b)-y_i\le\varepsilon+\xi_i \\ & y_i-(w\cdot\phi(x_i)+b)\le\varepsilon+\xi_i^* \\ & \xi_i,\xi_i^*\ge 0 \end{cases} \tag{4}$$

If the kernel function is defined as $K(x_i,x_j)=\phi(x_i)\phi(x_j)$, the Lagrange function and KKT theorem are adopted to transform Formula (4) into the antithetical parallelism problem, which is shown as follows.

$$\begin{cases} \max\limits_{a,a^*} & \sum\limits_{i=i}^{l}\left[a_i^*(y_i-\varepsilon)-a_i(y_i+\varepsilon)\right]-\frac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}(a_i-a_i^*)(a_j-a_i^*)K(x_i,x_j) \\ \text{s.t.} & 0\le a_i,\ a_i^*\le C/l,\ \ \sum\limits_{i=1}^{l}(a_i-a_i^*)=0,\ \ i=1,2,\cdots,l. \end{cases} \tag{5}$$

According to Formula (5), the solution is $(\bar{a},\bar{a}^*)$, and the regression function is expressed as follows.

$$f(x)=w\cdot\phi(x)+b=\sum_{SV}(\bar{a}-\bar{a}^*)K(x_i\cdot x)+\bar{b} \tag{6}$$

At last, the prediction is carried out by Formula (6).

The Mercer condition must be satisfied by the kernel function in the model. Therefore, the Gaussian kernel is chosen in this research, which is shown as follows.

$$K(x,y)=\exp\left(-\gamma\left|x-y\right|^2\right) \tag{7}$$

where the value of nuclear parameter γ and the value of penalty parameter C in Formula (4) are both the empirical value and the experimental value. That is to say, in the simulation experiments, the two parameter values are revised time and again, and the values with the favorable effects are selected. $C=15$ and $\gamma=2^{-1}$ are selected after the experiments for many times.

3. **Simulation Experiments and Result Analysis.**

3.1. **Data sample.** In 2013, the parallel volunteer intention for the college is implemented in Heilongjiang province. In the simulation experiments, the lowest scores are selected as sample data from 2013 to 2016 and randomly from 10 universities which have enrolled the key undergraduate science students in Heilongjiang province. The data from 2013 to 2015 is the training sample, and that of 2016 is the test sample.

The data samples from 2013 to 2016 are shown in Table 1.

Table 1. Data sample

College code	A	B	C	D	E	F	G	H	I	J
2013	683	632	533	527	590	651	594	646	623	592
2014	690	641	532	529	606	655	620	653	635	605
2015	681	625	499	483	588	651	607	648	620	585
2016	683	630	503	486	597	658	618	656	626	592

3.2. **Simulation experiment.** The data from 2013 to 2015 is used to train the established model, and to predict the lowest admission scores of the 10 colleges and universities in 2016. Then, the prediction scores are compared and tested with the true data of 2016 to calculate the relative score difference (True score subtracts the prediction score). Finally, the mean absolute error (MAE) is adopted to evaluate the effect of the prediction model of the paper.

Libsvm-3.21 toolkit of Matlab R2010b software is installed to carry out the prediction simulation experiments. The main code is as follows.

Main Code

```
tic;
x = [x1;x2]; y = [x2;x3];
model = svmtrain (y, x, ′ − s 3 −t 2 −c 15 −h 0 −g 0.5 −p 0.01′);
[py,mse,∼] = svmpredict (y,x,model);
testx = [x3]
display(‘true data’);
Testy = [x4]
[ptesty,tmse,∼] = svmpredict (testy,testx,model);
display(‘predicted data’);
ptesty
toc
```

The prediction results are shown as follows.

Table 2. Prediction result

College code	The lowest score in 2015	The lowest score in 2016	The predicted score in 2016	The relative score difference σ
A	681	683	686.1077	−3.1077
B	625	630	633.0000	−3.0000
C	499	503	515.5000	−12.5000
D	483	486	506.0000	−20.0000
E	588	597	598.2167	−1.2167
F	651	658	654.9900	3.0100
G	607	618	613.5000	4.5000
H	648	656	651.8370	4.1630
I	620	626	627.5832	−1.5832
J	585	592	595.0000	−3.0000

3.3. **Result analysis.** MAE is used to evaluate the prediction result, which is expressed as follows.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\sigma| = 5.6 \tag{8}$$

In Table 2, it is shown that the relative prediction error is bigger only in a few prediction points, the relative prediction error is comparatively smaller in most prediction points, and MAE is just 5.6 points. At the same time, the change tendency of the prediction score is very correct. The lowest admission scores of 10 colleges and universities of 2016 are higher than those of 2015, to which the prediction results are the same completely. It is proved that the prediction result is ideal and that it is feasible and practical to adopt the support vector machine theory to predict the college admission score.

4. **Conclusions.** The greatest advantage of support vector machine is to be used in the machine learning under the condition of the small sample. In the paper, based on the support vector machine theory, the lowest admission score data of each college is selected over the years to construct the college admission score prediction model for predicting the lowest line of the college admission scores for next year. Through the quantitative analysis of the prediction result, it is proved that the prediction result is ideal, and it is verified that the support vector machine prediction model is feasible to predict the college admission score. And the prediction model improvement, the qualitative analysis of the prediction result, and the prediction accuracy improvement are the further research direction, which have the great practical value to some extent.

Acknowledgment. This work was jointly supported by National Natural Science Foundation of China (Grant No. 61403222), Natural Science Foundation of Heilongjiang Province (Grant No. F201439, F201440), and Technology Program of Qiqihar (Grant No. GYGG-201408).

REFERENCES

[1] J. P. Wu, Students' score prediction of college entrance examination based on BP neural networks, *Public Communication of Science & Technology*, no.20, pp.164-165, 2015.

[2] R. H. Guo, X. Wan and P. F. Wu, Prediction of the number of college admission based on gray theory, *Journal of Hubei Engineering University*, vol.33, no.6, pp.48-51, 2013.

[3] J. J. Yu, An improved discrete grey model and its colleges and universities enrollment forecast modeling, *Journal of Science of Teachers' College and University*, vol.30, no.2, pp.46-49, 2010.

[4] G. Q. Xu and Y. Lin, Applying to college aided decision support system based on data mining, *Computing Technology and Automation*, vol.33, no.4, pp.106-109, 2014.

[5] B. Joo, S. Shim and H. Bae, Utilization of sequential data for machine learning in process control, *ICIC Express Letters*, vol.10, no.3, pp.535-540, 2016.

[6] Q. P. Huang, X. Zhou and Y. J. Gan, Application of SVM and neural network model in the stock prediction research, *Microcomputer & Its Applications*, vol.5134, no.5, pp.88-90, 2015.

[7] A. J. Zhou, Prediction research on the college admission score line based on GM(1,1) model gray system, *Fujian Computer*, vol.30, no.3, pp.25-26, 2014.

[8] Y. Huang, G. Huang and J. Ren, An improved grey model for urban air quality, *ICIC Express Letters, Part B: Applications*, vol.7, no.1, pp.97-103, 2016.

[9] Y. F. Miao and X. H. Zhang, Improvement and application of C4.5 decision tree algorithm, *Computer Engineering and Applications*, vol.51, no.13, pp.255-258, 2015.

[10] H. Yan, J. P. Lu and Q. Y. Qian, A nonlinear combined model for wind power forecasting based on multiple attribute decision making and support vector machine, *Automation of Electric Power Systems*, vol.37, no.10, pp.29-34, 2013.

[11] X. J. Chen, Z. A. Yao and W. Huang, Support vector machine in forecasting the supply of college graduates, *Acta Scientiarum Naturalium Universitatis Sunyatseni*, vol.52, no.1, pp.68-73, 2013.

[12] M. Han and D. F. Guo, Application research on college career prediction based on SVM selection, *Journal of Hangzhou Normal University (Natural Science Edition)*, vol.8, no.5, pp.358-362, 2009.