Study on the medical meteorological forecast of the number of hypertension inpatient based on SVR

Guangyu Zhai [1,2], Guorong Chai[*1] and Haifeng Zhang[1]

[1]School of Management, Lanzhou University, Lanzhou 730000, China
[2] School of Economics Management, Lanzhou University of Technology, Lanzhou 730050, China

*Corresponding email: chaigr@lzu.edu.cn

Abstract. The purpose of this study is to build a hypertension prediction model by discussing the meteorological factors for hypertension incidence. The research method is selecting the standard data of relative humidity, air temperature, visibility, wind speed and air pressure of Lanzhou from 2010 to 2012(calculating the maximum, minimum and average value with 5 days as a unit) as the input variables of Support Vector Regression(SVR) and the standard data of hypertension incidence of the same period as the output dependent variables to obtain the optimal prediction parameters by cross validation algorithm, then by SVR algorithm learning and training, a SVR forecast model for hypertension incidence is built. The result shows that the hypertension prediction model is composed of 15 input independent variables, the training accuracy is 0.005, the final error is 0.0026389. The forecast accuracy based on SVR model is 97.1429%, which is higher than statistical forecast equation and neural network prediction method. It is concluded that SVR model provides a new method for hypertension prediction with its simple calculation, small error as well as higher historical sample fitting and Independent sample forecast capability.

1.Introduction

According to statistics[1], 1/3 of people in the world suffer from cardio-cerebrovascular diseases, 15 million people die of cardio- cerebrovascular diseases each year, accounting for more than 3/5 of the total deaths. In China, about 2.6 million people die of cardio- cerebrovascular diseases each year[2]. Many studies at home and abroad show that meteorological condition is one of the trigger factors for the morbidity and mortality of cardio-cerebrovascular diseases[3]. Hypertension is one of the most common cardiovascular diseases, it can result in the lesion of blood vessels, brain, heart, kidney and other important organs, meanwhile, it can arouse some serious complications like hypertensive crisis and hypertensive encephalopathy, which endanger human's life. Medical meteorological forecast, according to the relationship between weather, climate, or meteorological factors and some diseases, adopt the research method of medical term and weather report to forecast the influence of future specific meteorological conditions on the occurrence, aggravation or remission of the disease[4]. In recent years, with the improvement of people's living standard, more attention is paid to the study of medical meteorological forecast[5]. At present, the research methods on the forecast of cardio-cerebrovascular diseases adopted in China include stepwise regression method, automatic interaction detector (AID), cross-validation method, artificial neural network and so on, which set up a forecast model by establishing statistical relationship between meteorological factors and patients' number. In some cities, the forecast model of some diseases' incidence has already been set up and the medical meteorological forecast system is developed based on the weather development, which release disease

level forecast to the public.

Support Vector Regression (SVR) adopts simplified SOR algorithm[6]. Compared with other training methods, SVR has a faster convergence speed in dealing with a large amount of data, which is suitable to conduct regression analysis for large samples. This paper analyzes the effect of meteorological factors on hypertension disease, selects factors and parameters which have a significant influence on hypertension by stepwise regression method, set up forecast model by SVR and conduct forecast experiment.

2. Materials and methods

2.1. Data source

Lanzhou city (36°-36°10′04N, 103°330′-104°E), which is located the center of Gansu Province, northwest China The daily hypertension inpatient number is collected from 5 first-class hospitals from 2010 to 2012, there are totally 13,326 hospitalized patients, among which there are 3,394 in 2010, 3,887 in 2011, 6,045 in 2012. The corresponding meteorological data of the same period is from Gansu Meteorology Bureau. By calculating meteorological data of daily 8 hours from 2010 to 2012 received by MICAPS, we can get 15 meteorological factors of the previous 5 days before hospitalization such as the average temperature, maximum daily temperature, daily minimum temperature, daily average air pressure, daily highest pressure, daily lowest pressure, daily average relative humidity, highest daily humidity, daily lowest humidity, daily biggest wind speed, daily smallest wind speed, daily average visibility, daily highest visibility and daily lowest visibility.

2.2. Method

The number of daily hypertension inpatient is as response variable, the related meteorological factors of the previous 5 days before hospitalization as independent variables. SVR model is used to analyze the influence of meteorological factors' change on the number of hypertension inpatient.

The basic idea of SVR is: On the basis of SVM[7], support vector regression method is put forward[8].Suppose a set of training data is given G={（$x_{i,}$ d_i）}(x_i is the system input vector x∈ R^n , d_i is the system output vector, l is the number of samples), support vector regression approaches the regression function by the following formula:

$$y = f(x) = \omega\phi(x) + b \quad (1)$$

In the above formula, $\phi(x)$ is high dimensional feature space obtained from input space x by nonlinear transformation. The difference between SVR and common regression algorithm like linear regression and nonlinear regression is that SVR is based on structural risk minimization principle and uses insensitive loss function ε rather than square error loss function to measure the difference between return value and target value. ε as insensitive loss function is expressed as

$$\left|\xi\right|_\varepsilon = \begin{cases} 0 & if\ |\xi| < \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \quad (2)$$

Based on the above， ε as insensitive loss function can be expressed as

$$L_\varepsilon(d, y) = \left|d - y\right|_\varepsilon \quad (3)$$

Compared with square error loss function, it is insensitive to additive noise with average value being zero, symmetrical distribution and smaller overall size, which makes regression algorithm have better robustness for this kind of noise, avoiding different regression results resulted from small disturbance of the training sample set. Besides, it is impossible to make regression equation and regression objects fit completely with limited training samples. In order to get an accurate result, a large number of training samples are needed, which is hard to be meted in reality. So it is of realistic significance to adopt insensitive loss function ε .

By minimizing coefficients ω and b in expression (1),

$$R_{SRM}(C) = C\frac{1}{l}\sum_{i=1}^{l} L_{\varepsilon}(d_i, y_i) + \frac{1}{2}\|\omega\|^2 \quad (4)$$

In the above expression defining loss function, the first item $C(1/l)\sum_{i=1}^{l} L_{\varepsilon}(d_i, y_i)$ is empirical risk, which is determined by the insensitive loss function ε in expression（3）. This method of defining loss function let us use small sample points to express the decision function in expression (1). The second item $\frac{1}{2}\|\omega\|^2$ is a regularization item. C is a regularization constant, compromising between empirical risk and the regularization item. If C value is enlarged, then the influence of empirical risk on loss function will be increased. ε is pipeline value, which decides the approximation precision of training samples. C and ε value need be given in advance according to actual circumstances of questions.

Then, searching ω and b makes the question of minimizing loss function in expression (4) equal to the following: searching ω and b, minimizing them as defined by slack variable ξ_i, ξ_i^*

$$\mathrm{R}_{SRM}(\omega, \xi^{(*)}) = \frac{1}{2}\|\omega\|^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{l}(\xi_i + \xi_i^*) \quad (5)$$

The constraint condition：

$$d_i - \omega\phi(x_i) - b_i \le \varepsilon + \xi_i, \quad (6)$$

$$\omega\phi(x_i) + b_i - d_i \le \varepsilon + \xi_i, \xi^{(*)} \ge 0 \quad (7)$$

By introducing Lagrange multiplier, solving the above optimization problem with inequality type:

$$L = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l} a_i(\varepsilon + \xi_i - y_i + (\omega \bullet \phi(x_i)) + b) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{l} a_i^*(\varepsilon + \xi_i^* + y_i - (\omega \bullet \phi(x_i)) - b) \quad (8)$$

In the above expression, a_i, a_i^* are Lagrangian multipliers, meeting the condition a_i, a_i^* =0，$a_i^{(*)} \ge 0 (i = 1, \cdots, l)$. In the saddle point, seeking the differential of L on $(\omega, b.\xi_i^*)$.

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l}(a_i^* - a_i) = 0 \quad (9)$$

$$\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{l}(a_i^* - a_i)\phi(x_i) = 0 \quad (10)$$

$$\frac{\partial L}{\partial \xi_i^{(*)}} = C - a_i^* - \eta_i^* = 0 \quad (11)$$

Put expression（9）,（10）and（11）into (8), an optimized equation can be obtained

$$\max W(a_i, a_i^*) = \sum_{i=1}^{l} y_i(a_i - a_i^*) - \varepsilon\sum_{i=1}^{l}(a_i + a_i^*) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i - a_i^*)(a_j - a_j^*)K(x_i, x_j) \quad (12)$$

Meeting constraint condition：

$$\sum_{i=1}^{l}(a_i - a_i^*) = 0 \quad (13)$$

$$0 \le a_i \le C, 0 \le a_i^* \le C, i = 1, 2, \cdots, l \quad (14)$$

According to Karush-Kuhn-Tucker(KKT),only part of coefficients $(a_i - a_i^*)$ are not equal to zero in expression (12).They define the support vector in the question. Expression (10) can be written as

$$\omega = \sum_{i=1}^{l}(a_i - a_i^*)\phi(x_i) \quad (15)$$

Then the decision function defined in expression (1) can be written as the following：

$$f(x,a_i,a_i^*)=\sum_{i=1}^{l}(a_i-a_i^*)K(x,x_i)+b \qquad (16)$$

3. Results and discussions

3.1. The SVR modeling and forecast results testing

3.1.1. Modeling. The purpose of model is to use regression model based on SVM method to conduct regression fitting for the daily hypertension inpatient in Lanzhou as showed in figure. Model assumption: the meteorological factors of the previous 5 days affecting daily hypertension inpatient from Jan 1, 2010 to November, 2012 include average temperature, maximum temperature, minimum temperature, average air pressure, highest pressure, lowest pressure, average relative humidity, highest humidity, lowest humidity, biggest wind speed, smallest wind speed, average visibility, highest visibility and lowest visibility, which are regarded as independent variables. It is showed that the morbidity of hypertension patients presents no obvious relationship with the meteorological factors' change of that day, but is closely related with the meteorological factors' change of the previous five days, which demonstrates that the accumulation of the meteorological factors of the previous five days impacts significantly on the morbidity of hypertension patients[9]. Therefore the meteorological factors of the previous five days are as the input variable of the forecast model. The number of inpatient from July to November in 2012 is regarded as dependent variable to forecast. The forecast model is as follows.

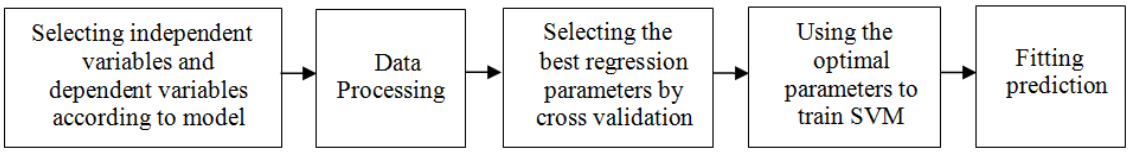

Figure1. SVR forecast model of hypertension

3.1.2. Parameter selection. There is no unified best method about the optimal selection of SVM parameters in the world. At present, the most common way is to find C (penalty parameters) and g (kernel function parameters) a value within a certain range of values by regulating parameters. For the fixed c and g, the training set is treated as the original data set and cross validation is used to get the group c and g training set to verify the classification accuracy, finally the group c and g making training set of the highest verification classification accuracy is selected as the optimal parameter. But there is one problem: how many groups of c and g are corresponding to the highest validation classification accuracy? How to deal with this kind of situation? The method is to choose the group c and g with the smallest parameter in achieving maximum validation classification accuracy as the optimal parameter. If there are several group g corresponding to c, then the first group c and g in searching are as optimal parameter.

The reason is as follows: a high c will lead to an excessive state of learning, i.e. training classification accuracy is very high but test set classification accuracy is very low (The classifier's generalization ability is reduced). So for all c and g as a group which can achieve the highest validation classification accuracy, a relative small penalty parameter c is regarded as a better choice. In the study, the number of hypertension inpatient from 2010 to 2012 is matched with the meteorological factor of the previous five days before hospitalization, parameters are obtained by cross validation c=11.3173，g=0.125，CVmse=0.0026389. Figure2 is the contour map of CV parameter selection results.

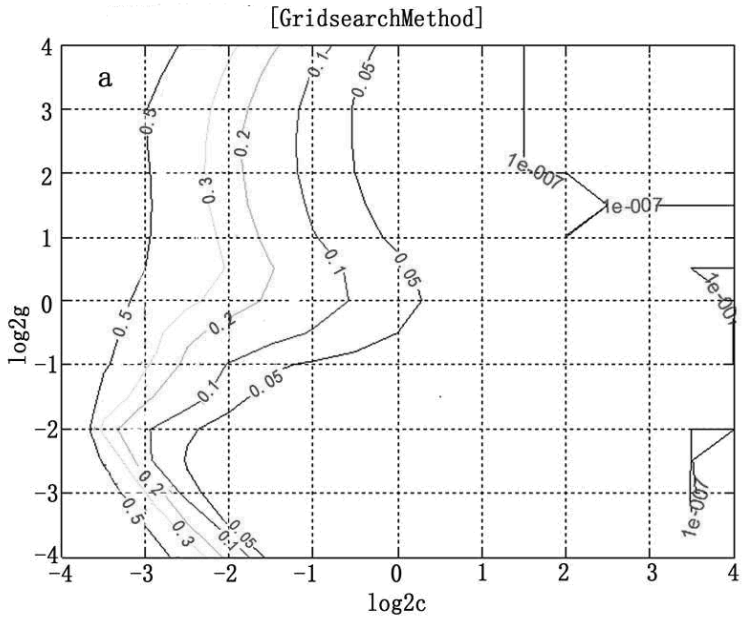

Figure 2. The SVR parameter selection results.

3.1.3. Training and regression forecast. The model is set up by using the number of hypertension inpatient from January, 2010 to November, 2012 and the meteorological data samples of the previous 5 days, training set and test set are constructed according to the proportion of 95% and 5%. Prediction set is constructed with a total of 70 samples from October 1, 2012 to December 9, 2012, i.e. the optimal regression model is adopted to forecast the number of daily hypertension inpatient from October 1, 2012 to October 9, 2012. In order to simplify forecast verification, the average number of daily hypertension inpatient is calculated, plus or minus 20% of average as the medium, then decreasing and increasing by 40%, the number of patients forecast is divided into 5 grades[32]. This is the published medical meteorological forecast results to the public by Media, which usually include 5 grades: few hypertension patients, a few hypertension patients, an increasing number of hypertension patients, many hypertension patients, a significant increasing number of patients as showed in Table 1 . By using a successive approximation optimization parameters and training repetitively, the parameters of forecast model of hypertension inpatient in Lanzhou are finally determined, among which c, g and w are determined in 3.1.2. A forecast model is established by using SVR to predict the number of hypertension inpatient in Lanzhou. The results are shown in figure 3. The classification prediction accuracy is up to 97.1429%, much higher than that of statistical methods and artificial neural network. This study uses meteorological factor as predictors to build a model for forecast, the correlation coefficient between forecast value and real value of hypertension inpatient is 0.95, which passes the test of significance; besides, the forecast value fluctuates significantly, which reflects that hypertension inpatient assembles in third and fourth grade with the fluctuation of time. It is showed that the forecast model set up by meteorological factors obtains a good forecast effect, the correlation coefficient of forecast value and true value is 98.016%, which passes the test of significance of α=0.01. According to the forecast grades of the patients' number, the accuracy rate is 97.1429% for the forecast value and true value of hypertension inpatient being in the same grade, the mean square error values is 0.0026389, it is a satisfactory result. It shows that the forecast model of hypertension disease can use meteorological factors as dependent variables for predication. By comparing the forecast value and true value, we can see the forecast value can reflect the variation trend of true value, but can't predict the peak time in second and fifth grade, which demonstrates that non-meteorological factors also have important effect on the number of hypertension inpatient such as holidays, festivals and the immeasurable factors of pathology (work pressure, dietary habit etc.), these factors can directly cause cardio-cerebrovascular diseases.

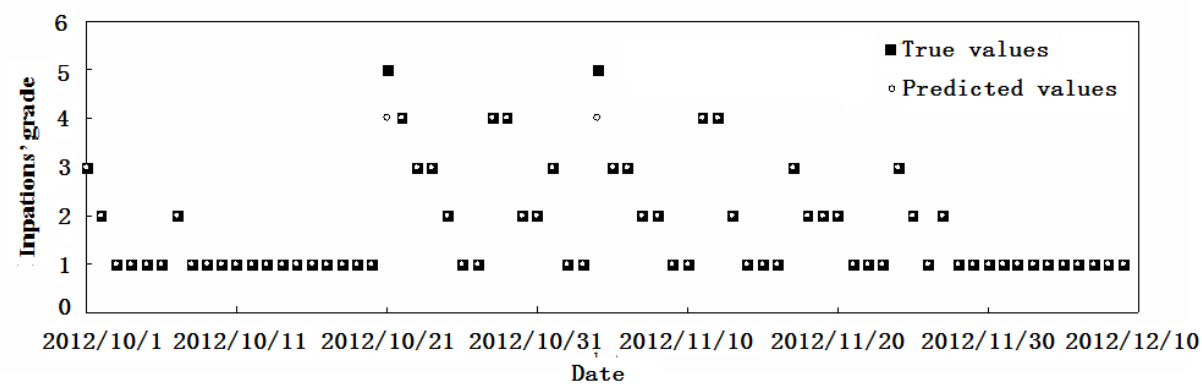

Figure 3. Daily forecast result comparison chart of hypertension inpatient by SVR method

Table 1. Daily hypertension inpatient forecast by grade.

Grade	Published level	Inpatient
1	few hypertension patients	<10
2	A few hypertension patients	[10,14)
3	an increasing number of hypertension patients	[14,22]
4	Many hypertension patients	(22,26]
5	a significant increasing number of hypertension patients	>26

5. Conclusions

This paper uses meteorological factors as dependent variables for forecast, screens relevant parameters by stepwise regression of relevant parameters and constructs a forecast model of hypertension disease in Lanzhou based in SVR. The results are as follows:

(1) The hypertension inpatient of Lanzhou presents obvious monthly change, which is related to meteorological factors. The meteorological factors influencing the hypertension inpatient include temperature, pressure, visibility, wind speed and dew point temperature of the previous five days, besides, the number of inpatients will also be influenced by the weather conditions' change, especially by dust strong convective weather.

(2) The number of hypertension inpatient is divided into 5 grades. The forecast model with this method displays a high accuracy for the hypertension disease in the same grade and good actual forecast results, which proves the practical business application value.

(3) It must be clear that meteorological condition is just one of the factors for hypertension disease, other factors may be contained in the periodic variation and trend. From the weekly variation of the number, it is showed that the number of hypertension inpatient is also constrained by social or economic factors like holidays and festivals, clinic specialists and so on. Therefore, it is necessary to consider these factors and design a more accurate forecast model.

Acknowledgements

This work was partially supported by grants from the China Postdoctoral Science Foundation (2016M600827), Specific Subjects of National Scientific Data Sharing Platform for Population and Health (2016NCMIZX09), National Natural Science Foundation of China(71472079), Key Project of China Ministry of Education for Philosophy and Social Science(16JZD023). Also, the authors would like to thank anonymous reviewers and editors for their useful comments which have helped to improve the quality of this manuscript.

References

[1] Goncalves, F. L. T., Braun, S., 2007. Influences of the weather and air pollution on cardiovascular disease in the metropolitan area of Sao Paulo. Environ Res. 104(2), 275-281.

[2] Liu, F., Zhang, J. L., Lu C., 2004. Review of Researches on Relationship of Meteorological Factors and Cardiovascular Diseases in China. Meteorological Science and Technology (in Chinese). 32(6), 425-428.
[3] Danet,S., Richard, F., 1999. Unhealthy effects of atmospheric temperature and pressure on the occurrence of myocardial infarction and coronary deaths. A 10-year survey: the Lille-World Health Organization MONICA project. 100(1), E1-E7.
[4] Yang, X. W., Ye, D. X., 2003. Medical Meteorological Research on Brain-Heart Vascular Syndrome in China. Meteorological Science and Technology (in Chinese). 31(6), 376-380.
[5] Chen, G. H., Zhang, Y. H., Song, G. X., 2007. Is diurnal temperature range a risk factor for acute stroke death. Int J Cardiol. 116(3), 408-409.
[6] Chen, Z. H., Yang, H. Q., Zhang, H. Y., Wang, Z. C., 2001. Research on Meteorological Forecast about Sick Rate of Diseases in Respiratory Tract and Heart and Brain Blood Vessels in Wuhan, china. Journal of Hubei college of TCM (in Chinese).3(2),15-18.
[7] Vapnik V. The Nautre of Statistical Learning Theory (2nd edition). New York:Springer-Verlag 1998.
[8] Smola, A. J., Scholkopf B., 1998. A Tutorial on Support Vector Regression. Neuro COLT2 Technical Report Series NC2-TR-1998-030 (http://www.neurocolt.com).
[9] Lin, C. J., 2001. Formulations of support vector machines: a note from an optimization point of view. Neural Computation. 13(2), 307-317.