Prediction of Postpartum Hemorrhage Volume of Pregnant Women Based on GA-SVM Algorithm

Ren-Jun SHUAI[1], Yang HE[1] and Ping CHEN[2]

[1]*Nanjing Tech University, University, 211816, China*
[2]*Nanjing Health Information Center, 210003 China*

Abstract: To take most advantage of the medical data resources from maternal and child health information platform and to improve the medical level, the team bring up a method based on support vector machine (SVM) algorithm which is aimed at predicting blood flow and blood pressure within 2-24 hours after parturition. We cleaned up the extracted data, determine the linear correlation via Pearson correlation coefficient, and utilize the significance to test and justify the relevance of data. Also, genetic algorithm is used to optimize the parameters. Then, we filter out the data with strong correlation coefficient and make predictions through the SVM algorithm. Finally, we determine the effectiveness of the prediction by doing the comparison between predicted results and the real data. The experiments show that, SVM is valid and feasible for the prediction of postpartum hemorrhage and the blood pressure.

1. Introduction

Postpartum hemorrhage (PPH) refers to a serious childbirth complication where the amount of bleed flow of pregnant women within 24 hours after delivery is more than 500ml. Postpartum hemorrhage often happens suddenly and may lead to death if pregnant women cannot be treated on time. Postpartum hypertension is also a serious complication of maternal mortality. So, if blood pressure and the amount of blood flow can be predicted, proper measures can be taken to avoid death of pregnant women.

As early as 20 years ago, our country has developed a 'Postpartum hemorrhage prediction score table', which has played an important role in prediction of postpartum hemorrhage. Yan Jianying and Huang Kehua screened out high risk factors by analyzing 212 cases of postpartum hemorrhage and 424 healthy cases from Provincial Maternity and Child Care Center of Fujian province. They establish the high risk scoring system for postpartum hemorrhage based on 'Postpartum hemorrhage prediction score table'. By comparison, they found that scoring system provides more accurate prediction than scoring table[1]. However, because of development and changes of people's lifestyles, factors of postpartum hemorrhage increased as well as maternal complications. With enhancement of medical level and improvement of data analysis skills, all stages of maternal data can be used, through algorithm, to predict the amount of maternal postpartum hemorrhage, doctors can be able to take appropriate emergent measures to reduce the possibilities of maternal mortality. The study of postpartum blood pressure are mostly based on the productive process of blood pressure changes.

2. Algorithm Background

2.1 Pearson Correlation Coefficient

There exists certain connections among things and Pearson correlation coefficient is used to describe them through quantitative indicators.

When analyzing, we should pay attention to an important basic principle: correlation is not equal to cause and effect. Correlation analysis in part shows correlation between the two variables, but not a certain type of this correlation. In other words, correlation analysis doesn't reveal which is cause and which is result[2].

As to two variables x and y, after several groups of tests for multiple sets of data, we mark them as(x_i , y_i)(i =1,2,3$\cdots$ n),then the expression of correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

In the formula, $\bar{x}$, $\bar{y}$ respectively represents average value of x , y .The value range of r is(-1,1),and the positive and negative signs in front of r represent the directions of the association, the absolute value indicates the intensity of the association. The closer to 1, the higher the degree of linear correlation is. If $r = 1$ or -1, then there is a perfect linear correlation between x and y [3].

2.2 Significance test of r

The correlation coefficient r is calculated by the data, so it will be affected by the randomness and the number of samples. So, we need to examine the reliability of the sample which justifies the significance of the experiment.

First, the null hypothesis for the sample is H_O.Second, we calculate the statistics for the test. Usually we describe the test by using t-distribution, and the formula is:

$$t = |r|\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2) \tag{2}$$

Finally, with T-distribution table, we determine the linear relationship and find the critical value between variables according to the pre-determined significance level α and degrees of freedom $d_i = n - 2$ [4,5].

2.3 Support Vector Machine

SVM (Support Vector Machine) is a neural network model of small probability events proposed by Vapnik

The core idea of SVM is to maximize the classification interval. Taking the two classification problem as an example, we hope to find a super plane, which can successfully separate the sample points of the two categories, and ensures that the point-to-hyperplane spacing is maximized. Compared to other similar machine learning algorithms (eg, Perceptron, Logistic Regression), SVM has more generalization ability[6].

2.3.1 Basic Theory of SVM

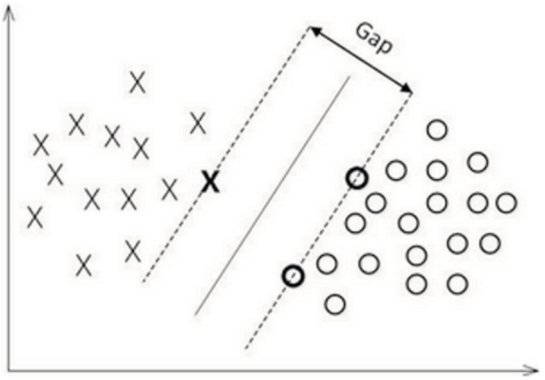

Figure 1. SVM algorithm idea

As shown in Figure 1, in the two-dimensional space, SVM algorithm is committed to find such a line, which can classify the data points and require data points to the distance of the straight line. The greater the distance is, the more improved accuracy will be. Therefore, we need to find such a plane that making this accuracy as high as possible. That is, we want to solve the optimization problem: $\max \gamma$.

In addition, some conditions need to be met:

$$y_i(w^T x_i + b) = \gamma_i \geq \gamma, i = 1,2.....n \tag{3}$$

Let $r = 1$, then this hyper plane can be expressed as:

$$\max \frac{1}{\|w\|} \tag{4}$$

$$s.t, y_i(w^T x_i + b) \geq 1, i = 1,2.....n \tag{5}$$

equivalent to:

$$\min \frac{1}{2}\|w\|^2, s.t, y_i(w^T x_i + b) \geq 1, i = 1,2.....n \tag{6}$$

According to the solving method of convex quadratic programming in optimization theory, we define the Lagrangian function:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1) \tag{7}$$

we also defined:

$$\theta(w) = \max_{\alpha_i \geq 0} L(w,b,\alpha) \tag{8}$$

When the above constraints are met, $\theta(w)$ is the same as $\frac{1}{2}\|w\|^2$, so the original problem is equivalent to:

$$\min_{w,b}\theta(w)=\min_{w,b}\max_{\alpha_i\geq 0}L(w,b,\alpha)=p^* \tag{9}$$

In order to facilitate the solution, the problem is transformed into a dual problem:

$$\max_{\alpha_i\geq 0}\min_{w,b}L(w,b,\alpha)=d^* \tag{10}$$

This optimization needs to satisfy the KKT condition:

$$1.\ h_j(x_*)=0, j=1,2.....p, g_k(x_*)\leq 0, k=1,2.....q \tag{11}$$

$$2.\ \nabla f(x_*)+\sum_{j=1}^{p}\lambda_j\nabla h_j(x_*)+\sum_{k=1}^{q}u_k\nabla g_k(x_*)=0,\ \ \lambda_j\neq o, u_k\geq 0, u_k g_k(x_*)=0 \tag{12}$$

To solve this equation, we need to seek the minimum of w and b .That is $\nabla_w L(w,b,\alpha),\nabla_b L(w,b,\alpha)=0$, then we get the results: $w=\sum_{i=1}^{n}\alpha_i x_i y_i$, $\sum_{i=1}^{n}\alpha_i y_i=0$ .Into the formula,we get:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j, s.t, \alpha_i\geq o, i=1,2.....n, \sum_{i=1}^{n}\alpha_i y_i=0 \tag{13}$$

For nonlinear functions, the SVM algorithm approach is to choose a kernel function K, which projecting data into higher dimensional space. In order to achieve this transformation which from high to low dimensional space conversion, we need to use kernel function $k(x,x_i)=\phi(x)\phi(x_i)$ .The common kernel functions are shown in Table 2, Corresponding linear separable equation (13) we will eventually become non-linear problems:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x,y), s.t., \alpha_i\geq 0, \sum\nolimits_{i=1}^{n}\alpha_i y_i=0, i=1,2.....n \tag{14}$$[7,8].

Table 2. Common kernel functions

Name	Expression
Linear kernel function	$K(x,y)=x^T y+c$
Polynomial kernel function	$K(x,y)=(rx^T y+c)^d$
Radial basis function (RBF) kernel function	$K(x,y)=\exp(-r\|x-y\|^2)$
Sigmoid kernel function	$K(x,y)=\tan r(a(x,y)+c)$

2.3.2 Regression prediction of SVM

In order to use the support vector machine for regression analysis, Cortes and others add the penalty term C to the objective function, and set the threshold to control the proportion of the error and convert the above equation into:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x,y),\ \ s.t.,0\leq\alpha\leq C, \sum\nolimits_{i=1}^{n}\alpha_i y_i=0, i=1,2.....n \tag{15}$$

$K(x, y)$ represents the kernel function type of the support vector machine from the low-dimensional to the high-dimensional space conversion, which describes prediction expression of SVM.

2.3.3 Realization of SVM Prediction Based on Genetic Algorithm Optimization

Genetic algorithm is an adaptive searching technique based on a selection and reproduction mechanism found in the natural evolution process, and it was pioneered by Holland in the 1970s.It has become very famous with its global searching, parallel computing, better robustness, and not needing differential information during evolution.

SVM algorithm needs to determine the three factors in dealing with the prediction of nonlinear function:1.Selection of kernel function;2.Parameters of the kernel function r;3.Penalties C. If the results are predicted without these factors, results will be ultimately affected. Genetic Algorithm is a method by using the evolution in biology sphere, that is, the natural selection of the biological world and the natural genetic mechanism of the randomized search method. The algorithm was proposed by Professor J. Holland of the United States in 1975. Its superiority lies in the use of probabilistic optimization method where it doesn't need to determine the rules, instead you can automatically get and guide the search space optimization. So, it is very suitable for large-scale parallel computing. It has the inherent heuristic stochastic search characteristic and is not easy to fall into the local optimum. The kernel function parameters and the penalty term are determined as follows:

Step 1: Coding, the kernel function type, the kernel function parameter and the penalty item in the SVM prediction algorithm are encoded by binary coding;

Step 2: Initialization, we need to select a group and extract a certain amount of individuals, then forms a population. The choice of individuals is to take a uniform distribution of methods rather than random methods, which ensures that the selected kernel function parameters and penalty terms are distributed evenly in the parameter space;

Step 3: Selection, crossover and mutation. Individual selection is based on the individual fitness of the population. The standard of individual fitness is the standard deviation of the prediction model. Crossover and mutation are used to process the population of the current generation, resulting in a new generation of population;

Step 4: Determine the best global convergence. If the selected individuals reach the expected threshold that the distance between the populations is less than the accepted value, the iteration process converges and the algorithm ends. Otherwise, it returns to step 3 and continues to cycle[9,10].

3.Data analysis

Data analysis is a process of gathering data, analyzing data, and making it become useful information. The main process of data analysis is:1.data collection;2.data storage;3.data extraction;4.data filtering;5.data analysis;6.data presentation;7.data applications.

The data of this experiment come from Nanjing Maternal and Child Health Care System. The subjects were collected from the Gulou Hospital from 2014 to 2015[11].

3.1 Data acquisition and storage

Data collection refers to the acquisition and collection of data. Doctors or pregnant women take the information of maternal and child into the system by using mobile Internet technology.

Data storage means that the data is stored in a certain format. Since Women and children's data are regarded as privacy, it must be stored on the database of personal sensitive privacy information (including patient privacy information related to the name, ID number, contact telephone, address, medical history and other data) for encryption. Nanjing Maternal and Child System uses M2-S5100 database security gateway to store and encrypt information of women and children.

3.2 Data extraction and data filtering

Data extraction traverses the entire database and extracts the relevant data from the cloud storage platform based on the data table name. Query the database, such as: AGE, PREGNANCIES (Pregnancy times), CGAGEW(Childbirth Week (Week)), CGAGED(Delivery Week (days)), MODELIVERY(Mode of delivery), THBPHIGH(2 hours postpartum blood pressure - high), THBPLOW(Two hours postpartum blood pressure - low), TWFBLEEDING(2-24 hours postpartum hemorrhage)...,a total of 8629 data is exported.

Data filtering is to delete, organize, and exclude useless data. Delete the entire data-free row, then we get the remaining 7833 data. We use K-means algorithm to select the most extreme clustering, and the data in the cluster is deleted by the whole row, remaining 5036 data. Finally, the CGAGEW, CGAGED integration components delivery days (DAY).

3.3 Experimental setup

We calculate the correlation coefficient and the significance level of TWFBLEEDING, THBPHIGH, THBPLOW,MODELIVERY,PREGNANCIES,DAY and two-tailed test was used for significance test.

We use SVM to predict the amount of blood loss and blood pressure. In the experiment, the original data were transformed into two sets of data samples with the class of subject. The first set of data is labeled the amount of bleeding, that is, each sample can be expressed as (xi, yi), xi is the input eigenvector, yi is bleeding. The second set of data is labeled with blood pressure, the same as the first set, xi is the input eigenvector, yi is the blood pressure value.

In addition, in order to test the prediction effect of the algorithm in this experiment, we use standard deviation to measure the performance of the forecast(the range of predicted average error).Otherwise,5- fold cross validation was used. The data was divided into 5 parts at random and the experiment was repeated 5 times. In each experiment, 4 training sets were selected and 1 were tested. The results are the average of the 5 experiments.

Genetic algorithm is used to optimize the experimental parameters to determine the kernel function, the kernel function parameters r and penalty items C .Repetitive use of SVM to predict maternal postpartum hemorrhage and blood pressure.

The experimental results were compared[12-16].

3.4 Data analysis

To determine the linear relationship between TWFBLEEDING and other data, the data were first analyzed for Pearson's coefficient and the significance test. The results in Table 3.

Table 3. Pearson correlation coefficient calculation results

		PREGNANCIES	MODELIVERY	DAY
TWFBLEEDING	Pearson	0.048**	0.131**	-0.048**
	Significance (two-tailed)	0.001	0.000	0.001
THBPHIGH	Pearson	0.001	0.027*	0.032*
	Significance (two-tailed)	0.923	0.045	0.023
THBPLOW	Pearson	-0.223	0.063**	0.029*
	Significance (two-tailed)	0.107	0.000	0.042

Notes:** Correlation is significant at 0.01 (double tail), * Correlation is significant at 0.05 (two tails).

Table 3 shows that TWFBLEEDING and PREGNANCIES, MODELIVERY is positively correlated and the DAY is a negative correlation, association is very significant, but very small. THBPHIGH and PREGNANCIES, MODELIVERY, DAY is positively correlated. THBPHIGH and PREGNANCIES is not significant. THBPHIGH and MODELIVERY, DAY association is significant, but very small. THBPLOW and PREGNANCIES is negative correlation, but MODELIVERY, DAY is positively correlated, THBPLOW and PREGNANCIES has no significant. The association of THBPLOW and MODELIVERY is very significant, but very small, THBPLOW and DAY association is significant, but very small.

Then, the postpartum hemorrhage and postpartum blood pressure were predicted by using the formula (14) and radial basis function (RBF) kernel function. TWFBLEEDING was predicted by PREGNANCIES, MODELIVERY, DAY according to the two-tailed significance test, THBPHIGH and THBPLOW were predicted by MODELIVERY and DAY. The prediction results are shown in Table 4.

Table 4. SVM algorithm prediction results (standard deviation)

	TWFBLEEDING	THBPHIGH	THBPLOW
SVM	49.2564	7.8239	7.2667
GA-SVM	30.1586	5.1257	5.3574

As can be seen from Table 4, the use of SVM and GA-SVM algorithm to predict TWFBLEEDING, THBPHIGH, THBPLOW is feasible, and the results are within the acceptable range. SVM prediction based on genetic algorithm has less error than traditional SVM prediction.

4 Concluding remarks

This study shows that the Pearson correlation coefficient and significance test were used to eliminate the factors which may influence the blood loss and blood pressure. Finally, the SVM prediction model based on the genetic algorithm was used to predict the hemorrhage and blood pressure. The results of this study show that SVM algorithm for maternal postpartum hemorrhage and blood pressure prediction is a viable method for improving the survival rate of pregnant women and it has a certain practical significance. In this paper, the data used in the prediction of the results is not a lot, and the object of predict is relatively simple, we remove some special circumstances, so the future can use more data, a better algorithm for postpartum hemorrhage and blood pressure prediction.

References

1. YAN Jian-ying, HUANG Ke-hua, LIU Qing-min, HUANG Xiao-yan, XU Rong-li. Study on risk factors of high risk postpartum hemorrhage and its clinical value [J].Chinese Journal of Practical Gynecology and Obstetrics, 2014,10: 791-797.
2. David de la Mata-Moya; María Pilar Jarabo-Amores; Jaime Martín de Nicolás; Manuel Rosa-Zurera. Approximating the Neyman–Pearson detector with 2C-SVMs. Application to radar detection[J].Signal Processing.2017:364-375.
3. Hogan, Anna1;Sellar, Sam2;Lingard, Bob2.Commercialising comparison: Pearson puts the TLC in soft capitalism[J].Journal of Education Policy.2016,Vol.31(No.3):243-258.
4. Zhao, Sen1;Shojaie, Ali1.A significance test for graph-constrained estimation[J]. Biometrics.2016,Vol.72(No.2):484-493.
5. Adam Claridge-Chang1, 2, 3,;Pryseley N Assam4, 5,.Estimation statistics should replace significance testing.[J].Nat Methods.2016,Vol.13(No.2):108-109.
6. Milad Jajarmizadeh, Elham Kakaei Lafdani, Sobri Harun,Azadeh Ahmadi. Application of SVM and SWAT models for monthly streamflow prediction, a case study in South of Iran[J]. KSCE Journal of Civil Engineering,2015,191:.

7. David de la Mata-Moya; María Pilar Jarabo-Amores; Jaime Martín de Nicolás; Manuel Rosa-Zurera.Approximating the Neyman–Pearson detector with 2C-SVMs. Application to radar detection[J].Signal Processing.2017:364-375.
8. Abbas M. Abd;Suhad M. Abd.Modelling the strength of lightweight foamed concrete using support vector machine (SVM)[J].Case Studies in Construction Materials.2017:8-15.
9. Aslahi-Shahri, B.1;Rahmani, R.2(r.r.rahmani@ieee.org); Chizari, M.3;Maralani, A.4;Eslami, M.5;Golkar, M.5;Ebrahimi, A.1.A hybrid method consisting of GA and SVM for intrusion detection system[J].Neural Computing and Applications.2016,Vol.27(No.6):1669-1676.
10. Nadeem,Mohammad1(mail.mdnadeem@gmail.com);Banka,Haider1;Venugopal,R.2.SVM-Based Predictive Modelling of Wet Pelletization Using Experimental and GA-Based Synthetic Data[J].Arabian Journal for Science and Engineering.2016,Vol.41(No.3): 1053-1065.
11. Francesco Di Tria;Ezio Lefons;Filippo Tangorra.Cost-benefit analysis of data warehouse design methodologies[J].Information Systems.2017: 47-62.
12. Yan-Qing Zhang;Nian-Sheng Tang.Bayesian local influence analysis of general estimating equations with nonignorable missing data[J].Computational Statistics and Data Analysis.2017: 184-200.
13. Shirong Denga;Kin-yat Liub;Xingqiu Zhaobc.Semiparametric regression analysis of multivariate longitudinal data with informative observation times[J].Computational Statistics and Data Analysis.2017: 120-130.
14. V. Bhanu Prasada; Supriya Mallicka; Ashish Dutt Upadhyayb; G.K. Ratha. Systematic review and individual patient data analysis of pediatric head and neck squamous cell carcinoma: An analysis of 217 cases[J].International Journal of Pediatric Otorhinolaryngology.2017: 75-81.
15. G.S. Vyasa(Assistant Professor);K.N. Jhab(Associate Professor).Benchmarking green building attributes to achieve cost effectiveness using a data envelopment analysis[J].Sustainable Cities and Society.2017: 127-134.
16. Shu-yi Guo;Qi Si.Mechanical hydraulic characteristic analysis scheme based on lightweight crowd data in mobile embedded devices[J].EURASIP Journal on Embedded Systems.2017,Vol.2017(No.1)