Prediction model of stroke recurrence based on support vector machine

Wenbing Chang, Yinglai Liu, Xingxing Xu and Shenghan Zhou

School of Reliability and System Engineering, Beihang University, Beijing 100191, China

Email: zhoush@buaa.edu.cn

Abstract. Stroke is a disease characterized by high morbidity, high disability and high mortality. In recent years, with the progress of society and the improvement of people's living standards, the cure and survival rate of stroke is gradually increasing, but at the same time, cases of recurrent stroke have also increased. Among the patients who have experienced recurrent stroke, it often leads to higher mortality and disability than those first-ever stroke patients. Therefore, it is important to improve the risk awareness of patients and medical workers, and to make scientific and accurate predictions on whether patients relapse, which is of great significance for the prevention and treatment of stroke. The purpose of this paper is to analyze the recurrence of stroke patients by analyzing the medical data of stroke patients, and to establish and develop a stroke prediction model based on Support Vector Machine (SVM). This model provides a basis for recurrence warning of stroke patients and rational allocation of medical resources.

1. Introduction

The 2010 Global Disease Burden Study (GBD 2010) showed that stroke was the third cause affecting disability-adjusted life years and had become a global health problem [1]. Stroke is one of the most lethal neurological diseases. Data show that stroke is the second leading cause of death in people over 60 years old, and the fifth cause of death in people aged 15 to 59 years [2]. More than two-thirds of stroke deaths occur in developing countries [3]. In China, stroke has become the leading cause of death. For this reason, many scholars have made research on stroke from various perspectives in order to make breakthroughs in etiology, diagnosis and treatment, so as to reduce the disability and death caused by stroke.

Recurrence of stroke refers to the phenomenon of recurrence of stroke patients. According to modern medical statistics, the fatality rate of recurrence of stroke is much higher than that of the first onset of the patient. At present, the medical community generally believes that prevention is the best way to prevent the recurrence of stroke. General medical research believes that the recurrence factors of stroke patients can be divided into two categories: non-intervention factors and intervention factors. Various influential factors act on the blood vessel tissue of patients, eventually causing the occurrence or recurrence of stroke, thereby endangering the life safety of patients. In order to obtain the important factors affecting stroke, many scholars have done a lot of research [4][5].

In 1995, Cortes and Vapnik first proposed a machine learning method based on statistical learning theory, VC dimension theory and structural risk minimization principle - Support Vector Machines (SVM) [6]. Subsequently, Vapnik and others carried out in-depth research on SVM, further
consolidating the theoretical basis. SVM takes into account both empirical risk and confidence interval to minimize the actual risk and has excellent performance in model generalization. Many scholars have carried out relevant research. Burges studied the application of support vector machine in pattern recognition and intrusion detection [7]. SVM is used in cancer prediction, pattern recognition and regression prediction [8][9]. Sun uses Lib SVM package to realize the optimal selection of parameters [10].

The increase of input variables will increase the computational complexity of the support vector machine model and reduce the prediction accuracy of the model. The setting of model parameters has an important influence on the prediction ability of the model, many studies have optimized the support vector machine model from the perspective of feature selection and parameter optimization. For feature selection, Simulated Annealing Approach and Recursive Feature Elimination can be used; for parameter optimization, genetic algorithm, grid-search algorithm and gradient-based adaptability can be used [11].

2. Method

2.1. Support vector machine classification

SVM was originally used to solve the problem of binary classification. The core of this method is to obtain a hyperplane, so that the hyperplane has the largest separation distance when dividing two kinds of data. Suppose a linear separable data set \( \{x_i, y_i\} (i = 1, 2, \cdots, N) \), \( x_i \in \mathbb{R}^d \) is the input variable and \( y_i \in \{+1, -1\} \) is the corresponding classification variable. Under linear separable conditions, assume that the data satisfies:

\[
y_i (w \cdot x_i + b) \geq +1
\]

(1) 

\( w \) is a row vector with the same dimension as the input variable \( x_i \), \( b \) is a scalar

Under the condition of linear inseparability, the relaxation variable \( e_i \geq 0 (i = 1, 2, \cdots, N) \) is introduced to indicate the deviation degree of classification errors.

\[
y_i (w \cdot x_i + b) + e_i \geq +1
\]

(2) 

At this point, the optimal hyperplane needs to be satisfied:

\[
\min_{w, b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} e_i
\]

(3) 

\( C \) is the model penalty parameters.

In order to solve the non-linear separable problem, the non-linear problem can be transformed into a linear problem in higher dimensional space by using the kernel function \( K(x_i, x_j) \). The kernel function used in this paper is the radial basis function.

\[
K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)
\]

(4) 

2.2. Support vector machine regression

Suppose there are data sets \( \{x_i, y_i\} (i = 1, 2, \cdots, N) \), \( x_i \in \mathbb{R}^d \) is vector, \( y_i \in \mathbb{R} \) is scalar, \( x_i \) is input variable, \( y_i \) is output variable, \( x_j \) and \( y_j \) are numerical variables, \( N \) is data sample size. For linear SVM regression, its purpose is to build a linear regression model:

\[
f(x) = w \cdot x + b
\]

(5) 

\( f(x) \) is the prediction result, \( w \) is the coefficient vector, \( x \) is the input variable and \( b \) is the intercept term.
For non-linear SVM regression, input variables are transformed into high-dimensional linear space by mapping, and a linear regression model is established in the high-dimensional space to satisfy the sample data.

\[ f(x) = w \cdot \Phi(x) + b \]  \hspace{1cm} (6)

\( \Phi(x) \) is a non-linear mapping that transforms \( x \) into high-dimensional feature space.

Considering the accuracy \( \varepsilon (\varepsilon > 0) \), the linear relationship between the following \( f(x) \) and the true value \( y_i \) is established. Then considering the data other than the accuracy \( \varepsilon (\varepsilon > 0) \), the relaxation variable \( \theta_i (\theta_i > 0) \) and \( \theta_i^* (\theta_i^* > 0) \) are added. Finally, considering the constraints, the SVM regression is equivalent to the following optimization problems.

\[
\min \left[ C \sum_{i=1}^{N} (\theta_i + \theta_i^*) + \frac{1}{2} \| w \|^2 \right] \\
\left\{ \begin{array}{l}
y_i - w \cdot \Phi(x) - b \leq \varepsilon + \theta_i \\
w \cdot \Phi(x) + b - y_i \leq \varepsilon + \theta_i^* \\
\varepsilon, \theta_i, \theta_i^* \geq 0, i = 1, 2, \cdots, N
\end{array} \right. \hspace{1cm} (7)
\]

Kernel function \( K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \) is also needed to map input variables into high-dimensional spaces.

### 2.3. Model framework

The main methods and steps of this paper are shown in Figure 1. Firstly, the original data are pretreated and the missing values are filled by the nearest neighbor average method. Because the number of original variables is relatively large, there is inevitably information redundancy, so factor analysis method is used to leave irrelevant variables.

Because each patient has a label, so some samples are selected as test sets, and others are the training sets. Then, the recurrence of stroke is predicted by SVM classification method and SVM regression method.

In this paper, two problems are considered, one is the parameter of SVM, the other is the feature selection. For parameter selection, that is, to select the optimal penalty parameter \( C \) and the parameter \( \gamma \) of the kernel function. There are three parameter optimization methods used in this paper, they are Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Grid Search Algorithm (GSA). By comparing the three methods, the parameters that make the final accuracy the highest are selected. For feature selection, this paper uses Recursive Feature Elimination (RFE) method. The idea of this method is to construct the model repeatedly and select the best feature combination. Put the selected features aside and repeat the process on the remaining features until all the features are traversed.

Finally, the classification and regression models are evaluated by the evaluation index constructed. Table 1 represents a record of the predicted results of the model.

**Table 1.** Record of model prediction results.

| Record indicator | Real situation | Predicted results |
|------------------|----------------|------------------|
| TP               | recurrence     | recurrence        |
| FP               | no recurrence  | recurrence        |
| TN               | no recurrence  | no recurrence     |
| FN               | recurrence     | no recurrence     |

\[
RR = \frac{TP}{TP + TN} \hspace{1cm} (8)
\]
The Second International Conference on Physics, Mathematics and Statistics
IOP Conf. Series: Journal of Physics: Conf. Series 1324 (2019) 012095
doi:10.1088/1742-6596/1324/1/012095

\[
NR = \frac{TN}{TN + FP}, \quad (9)
\]

\[
ACC = \frac{TP + TN}{TP + TN + FN + FP}, \quad (10)
\]

\[RR\] denotes recall rate, \(NR\) represents the prediction accuracy of no recurrence, and \(ACC\) denotes accurate rate.

\[\]

\[
\begin{align*}
\text{Raw Data} & \quad \downarrow \\
\text{Data Preprocessing} & \quad \downarrow \\
\text{Stroke-related Variables} & \quad \downarrow \\
\text{Factor Analysis} & \\
\text{SVM Classification Model} & \quad \downarrow \\
\text{SVM Regression Model} & \\
\text{Parameter Selection} & \quad \downarrow \\
\text{Feature Extraction} & \\
\text{Comparison of Accuracy Between SVM Classification and Regression Methods}
\end{align*}
\]

Figure 1. Method flow chart.

2.4. Data preprocessing
The data of this study were collected from 198 case records of a hospital in Beijing. 19 medical health variables were selected as input variables, including age, height, weight, sex, heart rate, blood pressure difference, drinking history, smoking history, glycosylated hemoglobin, blood sugar, total cholesterol, triglyceride, high density lipoprotein cholesterol, low density lipoprotein cholesterol, carotid plaque, lipoprotein-alpha, uric acid, fibrinogen, and homocysteine.

After slicing the original sample according to age, the final number of samples was 117. Because the patient is sliced with age, and age is also a variable of choice, so finally there are 18 variables remained. Next, the dimensionality of 18 variables is reduced, and the variables with correlation are removed, and four variables without correlation are generated as shown in Table 2. Among them, 1 represents recurrence and 0 represents no recurrence. The number of stroke recurrences is 60, and the number of non-recurrences is 57.

To explore the correlation between health variables, 18 health variables were analyzed. Using t-test, the original hypothesis is \(H_0\), assuming that there is correlation between variables. The P-Value values of heart rate, total cholesterol, low density lipoprotein cholesterol and homocysteine are less than 0.05,
therefore, it rejects the original hypothesis, these four variables are significantly related to stroke recurrence.

Table 2. Results of slicing medical health data.

| Patient number | Age | Heart rate | Blood pressure | Homocysteine | State |
|----------------|-----|------------|----------------|--------------|-------|
| 1              | 77  | 80         | 70             | 13.85        | 1     |
| 2              | 78  | 80         | 65             | 8.61         | 1     |
| 51             | 59  | 80         | 40             | 9.67         | 1     |
| 52             | 61  | 76         | 63             | 11.93        | 1     |
| 53             | 62  | 76         | 61             | 8.01         | 1     |
| 89             | 80  | 78         | 60             | 1.07         | 0     |
| 117            | 49  | 80         | 60             | 9.28         | 0     |

Summary

| Male | 58 | Recurrence | 60 |
|------|----|-------------|----|
| Female | 59 | No recurrence | 57 |

Because strongly correlated input variables will bring adverse effects such as model redundancy and prediction accuracy decline, if there is strong linear correlation between individual medical health variables, it is necessary to reduce the dimension of medical health variables. The results of KMO test for HR, TG, LDLC and HCY were 0.376 and less than 0.5, which indicated that there was no correlation between medical health variables and it was not suitable for factor analysis. Table 3 denotes the full name of the variable

Table 3. Description of main attributes.

| Attribute | Attribute description |
|-----------|-----------------------|
| HR        | heart rate            |
| TG        | total cholesterol     |
| LDLC      | Low density lipoprotein cholesterol |
| HCY       | Homocysteine          |

3. Process and result

3.1. Empirical study on recurrence prediction of stroke patients based on SVM classification

The four variables mentioned above are selected as input variables of SVM model, and the predicted results of different patients are taken as output variables. During the experiment, two-thirds of the samples are taken as training set and the rest as test set.

All four variables are introduced into the SVM classification model to optimize the parameters of the initial model. The range set of penalty parameter $C$ is $\{2^{-10}, 2^{-9}, 2^{-8}, \ldots, 2^{19}, 2^{10}\}$, and the range set of the kernel parameter $\gamma$ is $\{2^{-10}, 2^{-9}, 2^{-8}, \ldots, 2^{8}, 2^{9}, 2^{10}\}$. Through comparative analysis, the combination value of parameters ($C = 0.76$, $\gamma = 0.0051$) is the best choice, and the SVM classification model can achieve the highest prediction accuracy under the combination value of parameters. The results of comparative analysis after parameter optimization are shown in the Figure 2, 3 and 4.
According to Table 4, the accuracy of the model calculated by the grid search algorithm is the highest.

| Optimization method                  | Parameter combination $(C, \gamma)$ | Accuracy  |
|--------------------------------------|-------------------------------------|-----------|
| Particle swarm optimization           | $(0.95, 615.1768)$                   | 65.56%    |
| genetic algorithm                     | $(0.64, 83.2765)$                    | 65.56%    |
| Grid search algorithm                 | $(0.76, 0.0051)$                     | 67.78%    |

Figure 2. Grid search algorithm parameter optimization diagram.

Figure 3. GA parameter optimization.

Regarding feature selection, this paper uses RFE method to select the optimal model input variables. Because the optimal number of input variables in SVM classification model is unknown, all possible combinations of input variables need to be tested one by one. For example, when the number of input variables of the SVM classification model is set to 3, one of the four influential factors needs
to be excluded, so there are four combination schemes. The best combination for each case is shown in Table 5.

![PSO algorithm parameter optimization](image)

**Figure 4.** PSO algorithm parameter optimization.

In order to determine the accuracy of the SVM algorithm, the model is validated by ten-fold cross-validation method. Finally, the average value of ten cross-validations is 95.21%, which proves that the algorithm is more accurate.

| Number of input variables | Best combination          | RR    | NR    | ACC  |
|---------------------------|---------------------------|-------|-------|------|
| 4                         | HR, TG, LDLC, HCY         | 70.12%| 79.10%| 76.41%|
| 3                         | HR, TG, LDLC              | 71.83%| 81.22%| 76.00%|
| 2                         | TG, LDLC                  | 62.39%| 75.74%| 70.04%|
| 1                         | TG                        | 39.49%| 80.86%| 63.53%|

**Table 5.** Results of SVM classification model with different combinations of input variables.

![SVM classification prediction accuracy change graph](image)

**Figure 5.** SVM classification prediction accuracy change graph.

According to the table, the contribution degree of key influential factors to recurrence prediction of stroke patients is: TG > LDLC > HR > HCY. Introducing TG, LDLC and HR one by one will gradually improve the measurability of recurrence prediction of stroke patients, while continuing to introduce HCY will reduce the measurability.
Figure 5 shows that when the number of input variables of SVM classification prediction model is 3, the maximum predictable value of recurrence prediction of stroke patients will be obtained, and HR, TG and LDLC are the optimal combination of input variables of the model.

3.2. Empirical study on recurrence prediction of stroke patients based on SVM regression

All four health factors are incorporated into the SVM regression model, and the parameters of the initial model are optimized. The GSA algorithm, GA and PSO algorithm are used to optimize the algorithm. Through comparative analysis, the combination value of parameters ($C = 29, \gamma = 1$) is the best choice. The results of comparative analysis before and after parameter optimization are shown in Table 6.

Table 6. Parameter optimization results of SVM regression model.

| Evaluating indicator | Initial model | GSA | GA | PSO |
|----------------------|---------------|-----|----|-----|
| RR                   | 56.34%        | 62.33% | 60.85% | 63.35% |
| NR                   | 81.77%        | 79.81% | 78.48% | 76.68% |
| ACC                  | 70.93%        | 72.84% | 71.29% | 71.42% |

Similar to the SVM classification model, the REF method is used to select the optimal model input variables. Because the optimal number of input variables in SVM regression model is unknown, all possible combinations of input variables need to be tested one by one. The comparison results are shown in Table 7.

The relationship between the number of input variables and prediction accuracy of SVM regression prediction model is shown in Figure 6.

When the number of input variables of SVM regression prediction model is 4, the measurable recurrence prediction of stroke patients will get the maximum value. HR, TG, LDLC, HCY are the optimal combination of input variables.

Table 7. Results of SVM regression model with different combinations of input variables.

| Number of input variables | Best combination       | RR     | NR     | ACC    |
|---------------------------|------------------------|--------|--------|--------|
| 4                         | HR, TG, LDLC, HCY      | 66.28% | 81.78% | 75.19% |
| 3                         | HR, TG, LDLC           | 62.29% | 83.03% | 74.23% |
| 2                         | TG, LDLC               | 57.94% | 77.40% | 68.97% |
| 1                         | TG                     | 35.25% | 82.72% | 62.82% |

Figure 6. SVM regression prediction accuracy change graph.
3.3. Comparative analysis of models
According to the Table 8, the prediction accuracy of the SVM classification method is higher. From the perspective of model input, three medical health factors such as HR, TG and LDLC are used as model input combinations to establish a predictive model of stroke recurrence based on support vector machine classification, which can achieve 75.67% prediction measurability.

Table 8: Comparison of predictive models of stroke recurrence in patients.

| Predictive model  | Factor combination | ACC   |
|-------------------|--------------------|-------|
| SVM classification| HR, TG, LDLC       | 75.67%|
| SVM regression    | HR, TG, LDLC, HCY  | 73.31%|

4. Conclusions
Although there are many factors influencing the recurrence of stroke patients, but the analysis shows that only a few key medical health variables have an important contribution to the recurrence prediction of stroke patients, focusing on these key points. Medical health variables and rational use can achieve higher predictive measures of recurrence of stroke patients, so as to take early measures to avoid life threats caused by stroke recurrence. In this paper, four unrelated variables are finally obtained from the original eighteen features by correlation analysis, which can be used as the input of SVM. Finally, these variables are combined by RFE method reasonably and effectively to obtain more accurate prediction results. In the medical field, for the diagnosis of a disease, we often need to use a large number of detection indicators, which are related in many cases, so there is information redundancy, how to reasonably eliminate the unimportant variables is crucial.

In this paper, correlation analysis and RFE are used in feature selection, which is a relatively basic method. SVM is a relatively basic classification algorithm, there are many other more accurate classification algorithms, in the next study can also use other classification algorithms to predict.

References
[1] Murray C J L and Lopez A D 2013 Measuring the Global Burden of Disease New England Journal of Medicine 369 448-457
[2] Johnston S C, Mendis S and Mathers C D 2009 Global variation in stroke burden and mortality: estimates from monitoring, surveillance and modelling The Lancet Neurol 8 345-354
[3] Zhang W, Fan Y W, Gao J and Zhang L M 2014 Stroke Epidemiological Investigation with Literature Review Chinese Journal of Clinical Neurosciences 22 699-703
[4] Ois A, Gomis M and Rodriguez C A 2008 Factors associated with a high risk of recurrence in patients with transient ischemic attack or minor stroke Stroke 39 1717-1721
[5] Liu G, Yun Y Z and Neurology D O 2018 Study on the secondary prevention of ischemic stroke Shanghai Medical & Pharmaceutical Journal 39 22-24+55
[6] Ding S F, Qi B J and Tan H Y 2011 An Overview on Theory and Algorithm of Support Vector Machines Journal of University of Electronic Science & Technology of China 40 2-10
[7] Burges C J C 1998 A Tutorial on Support Vector Machines for Pattern Recognition Data Mining and Knowledge Discovery 2 121-167
[8] Huang M W, Chen C W and Lin W C 2017 SVM and SVM Ensembles in Breast Cancer Prediction Plos One 12 e0161501
[9] Sloin A and Burshtein D 2008 Support Vector Machine Training for Improved Hidden Markov Modeling IEEE Transactions on Signal Processing 56 172-188
[10] Sun J P and She J 2013 Coal-rock image feature extraction and recognition based on support vector machine Journal of China Coal Society 38 508-512
[11] Yang S, Huang Q and Li L 2009 An Integrated Scheme for Feature Selection and Parameter Setting in the Support Vector Machine Modeling and Its Application to the Prediction of Pharmacokinetic Properties of Drugs Artificial Intelligence in Medicine 46 155-163