Optimised Selection of Stroke Biomarker Based on Svm and Information Theory

Xiang WANG, Wei SHI, Xiao-Cui WANG, Tao WANG

School of Electronic and Information Engineering, Beihang University, Beijing, China
wxiang@buaa.edu.cn,littlestone_buaa@163.com,wangxc1992@buaa.edu.cn,wt860122@buaa.edu.cn

Abstract—With the development of molecular biology and gene-engineering technology, gene diagnosis has been an emerging approach for modern life sciences. Biological marker, recognized as the hot topic in the molecular and gene fields, has important values in early diagnosis, malignant tumor stage, treatment and therapeutic efficacy evaluation. So far, the researcher has not found any effective way to predict and distinguish different type of stroke. In this paper, we aim to optimize stroke biomarker and figure out effective stroke detection index based on SVM (support vector machine) and information theory. Through mutual information analysis and principal component analysis to complete the selection of biomarkers and then we use SVM to verify our model. According to the testing data of patients provided by Xuanwu Hospital, we explore the significant markers of the stroke through data analysis. Our model can predict stroke well. Then discuss the effects of each biomarker on the incidence of stroke.

1 Introduction

Nowadays, stroke has become one of the most prevalent and deadly diseases worldwide, both the occurrence rate and the mortality rate remain at a high level. For instance, approximately 6.9 million people had an ischemic stroke and 3.4 million people had a hemorrhagic stroke in 2013[1]. Between 1990 and 2010 the number of strokes which occurred each year decreased by approximately 10% in the developed world and increased by 10% in the developing world [2]. In 2013, stroke was the second most frequent cause of death after coronary artery disease, accounting for 6.4 million deaths (12% of the total) [3]. Because symptoms of lung cancer resemble those of other diseases such as phthisis or pulmonary infection, it is possible to misdiagnose clinically and even to lose the optimal opportunity for the treatment. So far, conventional detecting methods includes chest CT examination, Magnetic Resonance Angiography and Digital subtraction angiography, requiring high testing costs and expensive apparatuses. And the treatment of stoke are mostly surgical operation. All these methods rely on doctors' judgment based on experiments, and are time and money consuming to some extent. This current situation calls for new technologies to realize more efficient and accurate lung cancer detection, and the development of genetic circuit meets this need.

Marker detection has become a hit idea in achieving early-stage prediction and treatment of various diseases. It is an indicator of physiological and pathological state, produced by the illness itself or by the body in response to diseased tissue's growth. Compared to the traditional detection methods which usually have considerable experimental requirements and costs, markers can not only explore the pathogenesis at the molecular level, but also have unique advantages in accurate and sensitive evaluation of early, low levels damage, which provides an early warning of the disease [4]. Many markers such as, s100 and GFAP, have been found to play important roles in the diagnosis, treatment and prognosis of stroke [5-6]. Thus, taking advantage of the genetic circuits inside cells, we can achieve more sensitive and economical marker detection with good real-time performance.

Support Vector Machine (SVM), based on the statistical theory of VC dimension theory and Structural Risk Minimization Theory (SRM), is a specifically designed machine learning algorithm for small sample data processing. SVM shows many unique advantages in solving the problem of pattern recognition of small sample, nonlinear and high dimensional data, so SVM method has been widely used in many fields including pattern recognition, regression estimation, probability density function estimation, character recognition, speech recognition, text classification, signal processing and other fields [7].

In recent year, the biomarker testing of stroke has attracted more and more attention to help make the early diagnosis of stroke or make an evaluation of relent patients.

For acute stroke, the biologist cannot find out one single, effective biomarker to predict the condition, except use multiple markers to finish the analysis. In this paper, we focus on filtering and detecting the stroke markers so as to realize efficient and accuracy detection of stoke. The former part, based on the data provided by Xuanwu Hospital, is proposed to figure out the significant markers of the stroke through data analysis. The second part is provided to analyze the importance of each marker in stroke prediction or diagnosis.

2 Screening of biomarkers

Nowadays, in order to detect the acute ischemic stroke patients, we have to perform multiple biochemical examination such as routine blood test, threshold function test and blood fat and blood sugar examination. Plenty of biomarkers are in tests, however some of these biomarkers are useless and duplicate. At the same time, there will be some overlap of information between biological markers with similar biological function, increasing the test cost. In order to detect the acute stroke quickly and effectively, we first screen the biomarkers of acute stroke.

The experiment data set contains leukocyte, APOA and other biomarkers (24 species all in all). We conducted a total of 500 sets of experimental data processing, corresponding to 500 patients sample. We treat 3/4 as a training set, and the rest as validation set in order to verify our model.

2.1 Mutual information analysis

Because the mutual information does not need to presume the distribution type of data, it can effectively capture the nonlinear correlation of variables. The relationship between stroke markers and the prevalence of stroke is not simple linear. Therefore, we use mutual information as the correlation analysis method to screen the unrelated markers [9].

The mutual information of two discrete random variables X and Y can be defined by Eq.1:

$$I(X;Y) = H(X) - H(X|Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

Where $p(x,y)$ is the joint probability distribution function of X and Y, $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively. The mutual information satisfies $0 \le I(X;Y) \le H(X)$.The lower bound is reached if and only if X and Y are independent. The upper bound is achieved when Y fully determine X. Hence, the larger the mutual information, the more close the relation between X and Y is to a one-to-one relation. Treat multiple biomarkers as $x_1, x_2, x_3 \ldots$, and treat The sick(1) or not(0) as a Y variable, calculate the mutual information between X and Y, respectively, we can get the relationship between each biomarker and stroke, shown in table 1.

Through the results, we can see that Fib, Cholesterol, uric acid and several biomarkers of stroke has a high MI [10]. From the pathogenesis of stroke, these markers has a close relationship with acute stroke. The serum level of UA increasing slightly is relevant to the increasing cerebral ischemia pathological load. Fib can promote vascular smooth muscle cells, ultimately affect the blood viscosity. Cholesterol will make the blood pressure change and bring about acute stroke [11]. In order to detect the acute ischemic stroke patients accurately and efficiently, we take 0.02 as the threshold.

TABLE I. THE RESULT OF MUTUAL INFORMATION

Biomarker	***MI***	***Biomarker***	***MI***
platele	0.0241	High Density Lipoprotein	0.0137
red blood cell	0.0082	Low density lipoprotein	0.0372
leukocyte	0.0239	APOA	0.0157
blood sugar	0.0147	APOB	0.0270
Glycosylated Hemoglobin	0.0169	Fib	0.0582
HCY	0.0197	CRP	0.0207
creatinine	0.0183	anti-thyroglobulin antibody	0.0111
BUN	0.0236	TSH	0.0276
uric Acid	0.0707	T3	0.0245
triglyceride	0.0136	T4	0.0189
Cholesterol	0.0406	FT3	0.0294
		FT4	0.0110

2.2 Principal component analysis

The performance of the model is closely related to its structure, and the computational efficiency of the model is determined by the complexity of the structure [9]. There are complex relationships among the 12 high correlation biomarkers we screen out by mutual information. If all of them are adopted as inputs, it will directly increase the complexity of the model structure, and redundancy problems may be passed to the output of the model, which would cause serious influence on the detection performance.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [9]. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding

components [8]. The resulting vectors are uncorrelated orthogonal basis set.

High dimensional data are often transformed into lower dimensional data via the principal component analysis where coherent patterns can be detected more clearly. Thus it could decrease the complexity of the procedure and cut down the time consume.

The linear combination of the n variables $x_1, x_2 \ldots x_n$, can be composed of up to n integrated variables, listed in

$$\begin{cases} f_1 = a_{11}x_1 + a_{12}x_2 + \ldots a_{1n}x_n = \mathbf{a}_1^T \mathbf{x} \\ f_2 = a_{21}x_1 + a_{22}x_2 + \ldots a_{2n}x_n = \mathbf{a}_2^T \mathbf{x} \\ \ldots \\ f_n = a_{n1}x_1 + a_{n2}x_2 + \ldots a_{nn}x_n = \mathbf{a}_n^T \mathbf{x} \end{cases} \quad (2)$$

For each $i(i = 1, 2, \ldots n)$,it meets the conditions of standardization(3) and non-related conditions(4).

$$\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_i = \alpha_{i1}^2 + \alpha_{i1}^2 + \ldots \alpha_{in}^2 = 1 \quad (3)$$

$$\mathrm{cov}(f_i, f_j) = 0 \qquad j < i, j = 1, 2 \ldots n \quad (4)$$

Under the premise of satisfying these two conditions, the variance $\mathrm{var}(f_i)$ reaches the maximum, then the comprehensive variable f_i is the first i principal component of the overall x .Principal component analysis can achieve data compression, and achieve the purpose of revealing the intrinsic relationship between variables and statistical interpretation.

TABLE II. CUMULATIVE VARIANCE CONTRIBUTION OF EACH MARKER

Component number	*Eigen value*	*Cumulative variance contribution/%*
1	2.986	24.882
2	2.253	43.653
3	1.634	55.543
4	1.354	65.978
5	1.212	77.197
6	1.010	86.828
7	0.886	89.075
8	0.730	91.133
9	0.487	94.838
10	0.281	97.179
11	0.148	98.996
12	0.120	100.000

Formula (1) and (2) are used to measure the information ratio and variance contribution rate of the original independent variables contained in the principal components.

$$\gamma_i = \mathrm{var}(f_i) / \sum_{k=1}^{p} \mathrm{var}(f_k) = \lambda_i / \sum_{k=1}^{p} \lambda_k \quad (5)$$

$$\eta_m = \sum_{k=1}^{m} \gamma_k = \sum_{k=1}^{m} \lambda_k / \sum_{j=1}^{p} \lambda_j \quad (6)$$

The ETA _m determines the size should be to replace the original variables with the number of principal components, usually when the value of _m reached 70%~90%, while the rest of the main points can be omitted, thus completing the dimensionality reduction function. Table I shows the result of PCA.

2.3Total variance explained of PCA(principal component analysis)

The scree plot can be used to determine the best number of principal component. The abscissa plot in the diagram stands of principal components, and the vertical plot represents for the eigenvalue, so the steep part of the connection of principal component and eigenvalues would be the right number of principal component we take. In this study, the eigenvalue bigger than 1 is investigated and the scree plot and variance contribution ratio are both considered to determine the optimal principal component. From figure2, it can be seen that the eigenvalues of the first six principal components are larger, and the line is steep, so the first six principal components make the greatest contribution to the explanatory variables. From table 2, the first six principal components' eigenvalues are bigger than 1, and it is proper to extract six principal components and the cumulative variance contribution rate is 86.56%, which concentrates the most information of the acute cerebral stroke.

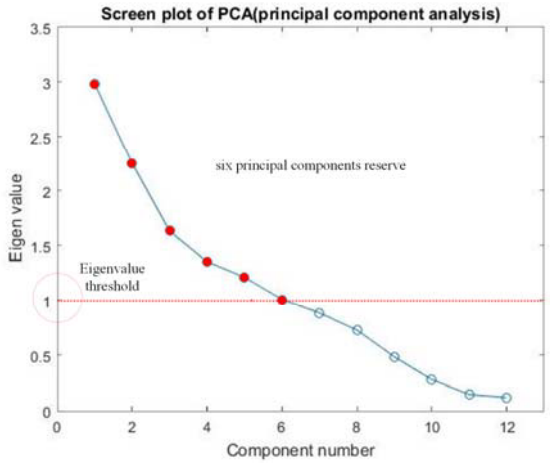

Figure 1. Screen plot of PCA (principal component analysis)

After the rotation of the load matrix of the principal component, the load factor is closer to 1 or -1, so the principal component can explain the named variable better. Table 3 shows that the first principal component f_1 mainly integrated the information of Cholesterol, low density lipoprotein and APOB. These three biomarkers are related to the body fat content, so f_1 is named as fat factor.

The second principal component f_2 mainly shows the information of CRP, and this protein value can be a good indication of acute stroke, so we name f_2 as protein factor. The third principal component f_3 mainly contains the information of T3, FT3. Both of these two are T3, but their forms are different, so f_3 is named as T3 factor. The fourth main component integrates two kinds of human metabolite, which characters the body's metabolic information, so we name it as the metabolic factor. The fifth principal component mainly reflects the content of TSH and Fib. The internal connection between the two is not clear, but they performs the same importance, so f_5 is named TSH-FIB factor. The sixth principal component mainly integrates the information of platelet and leukocyte, and the former is in the negative direction of the sixth principal component. The platelet is produced from a giant cell, and the giant cell and leukocyte are both belong to the human body cells, so we name f_6 as the cell factor.

TABLE III. ROTATED COMPONENT MATRIX OF PCA(PRINCIPAL COMPONENT ANALYSIS)

Biomarker	f_1	f_2	f_3	f_4	f_5	f_6
platele	0.170	0.237	0.420	-0.055	-0.077	-0.530
leukocyte	0.134	0.321	0.075	0.542	0.072	0.398
BUN	0.142	0.288	0.359	0.139	-0.400	-0.329
uric Acid	-0.118	-0.128	0.129	0.751	0.077	-0.081
Cholesterol	0.457	-0.304	0.171	0.029	0.030	0.152
Low density lipoprotein	0.482	-0.263	0.125	0.040	0.002	0.043
APOB	0.439	-0.252	0.138	-0.081	-0.061	0.223
Fib	0.127	0.356	0.320	-0.298	0.375	0.187
CRP	0.085	0.501	0.032	-0.017	0.290	0.254
TSH	0.178	-0.107	-0.190	0.119	0.726	-0.499
T3	-0.314	-0.252	0.512	-0.092	0.225	0.072
FT3	-0.363	-0.248	0.454	0.006	0.128	0.133

These six factors are set as the input of the next level of acute stroke screening process, and the results of the test will verify the rationality of the combination of selected biological markers.

3 Diagnosis and prediction of stroke

3.1 SVM(support vector machine) background

In order to extract enough information from the data to determine and predict the acute stroke, we use support vector machine to diagnose and predict. SVM (support vector machine) is a new kind of machine learning based on statistical learning theory, a technique for pattern recognition and classification in a variety of applications for its ability for detecting patterns in experimental databases [7]. First define the training and validation set in this paper.

The training set

$$S = \{(x_i, y_i), ...(x_l, y_l)\} \in (\mathrm{X} * \mathrm{Y})^l \quad (7)$$

The validation set

$$T = \{(x_{l+1}, y_{l+1}), ...(x_n, y_n)\} \in (\mathrm{X} * \mathrm{Y})^{n-l} \quad (8)$$

where $x_i \in X = R^n, y_i \in \mathrm{Y} = \{0,1\} (i=1,2,...n)$. n is the total number of stroke set, l is the number of training set. And we assume R^n represents the multidimensional space of the biomarker, x_i represents each of these biomarkers. y is a label to characterize the presence or absence of disease, 0 represents no stroke, 1 represents stroke. In the distinction between hemorrhagic stroke and ischemic stroke identification process, the result set Y is $\{2,3\}$, 2 represents hemorrhagic stroke, 3 represents ischemic stroke.

In this paper, RBF is adopted as kernel function in SVM to map the input variables into the high-dimensional Hilbert space by non-linear mapping. We assume

$$K(x_i, x'_j) = e^{\left\{\frac{-\|x_i, x'_j\|^2}{2\sigma^2}\right\}} \quad (9)$$

$$K(x_i, x'_j) = \phi(x_i) \cdot \phi(x'_j) \quad (10)$$

Thus, the corresponding classification function given in

$$f_{(x)} = \mathrm{sgn}[w.\phi_{(x)} + b] = \mathrm{sgn}[\sum_{j=1}^{l} x_i' \alpha_i K(\mathbf{x}_i, \mathbf{x}_j') + b] \quad (11)$$

Thus, nonlinear SVM is the optimization problem of the following equation.

$$\begin{cases} \min \phi_{(w)} = \frac{1}{2}(w.w) + c\sum_{j=1}^{l} \xi_i \\ s.t x_i'((w.\phi_{(x_i)}) + b) \geq 1 - \xi_i \geq 0, \xi_i \geq 0, i = 1, 2...l \end{cases} \quad (12)$$

3.2Combined detection

It is important to determine the exact location and extent of the injury before and after the treatment, which is important for deciding what treatment strategies to take. For example, before a proper treatment is performed, a physician must quickly determine the type of stroke, ischemic stroke or hemorrhagic stroke, because different types of stroke treatment strategies are very different. In order to obtain the complete diagnostic information, the need for the following tests to assess the risk of stroke or stroke: CT scan magnetic resonance (MRI), magnetic resonance angiography (MRA), transcranial Doppler (TCD), carotid artery ultrasound.

At present, most of the instruments are used to diagnose the type of stroke, which can be distinguished by the image or cell metabolism. These advanced detection means, but the high cost, while the popularity of these detection equipment is not high, the general hospital may not have such conditions, we can also achieve the goal of the traditional biochemical testing.

Although the clinical results of ischemic stroke and hemorrhagic stroke are similar, but the process is very complex, there are overlapping factors, there are more obvious factors. In modeling, the input variables are usually chosen as input variables of the model. However, many of the original input variables detected by modern instruments, there will be some factors and the results do not matter, if these factors into it, the model will not be accurate. Although some factors are related to the dependent variables, there is a redundant relationship between each other. Therefore, we need to filter out irrelevant factors and reduce the redundant variables.

3.2.1The detection of normal people and patients

Accurate and rapid determination of the exact location and size of the lesion is needed before treatment, and it is important to determine what treatment strategies are required. For example, a doctor must quickly determine the type of stroke before the appropriate treatment is performed, ischemic stroke or hemorrhagic stroke, because the different types of stroke treatment strategies are different. Thus we take steps to diagnose and predict stroke.

The first step is to distinguish patients from patients and normal people. Once the result is stroke, SVM is performed for further judgement of its type. The experiment result proves that SVM can achieve higher prediction results through the six factors of joint detection.

From Fig. 3 we can see that in the combined detection it can distinguish patients from normal people perfectly. And in the second step, it also has a high rate of distinguishing between ischemic stroke and hemorrhagic stroke. Thus combined detection has a good result of experiment.

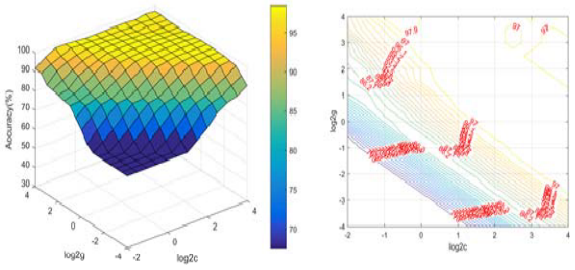

Figure 2. Learning process of the first step(between normal people and patients)

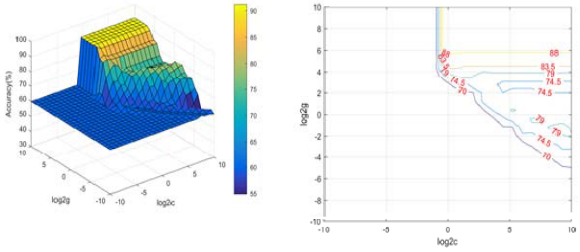

Figure 3. Learning process of the second step(between hemorrhagic and ischemic stroke)

TABLE IV. PREDICTION ACCURACY OF THE SVM

step	specificity	sensitivity	Correct index(specificity+ sensitivity-1)
The first step	100%	100%	100%
The second step	88.89%	92.59%	81.48%

3.2.2 Single biomarker detection

Then we analyze each marker's effect on the stroke based on the SVM model. For related biomarker such as: C-reactive protein, FIB, HCY, etc., the predicting accuracy rate is below 80%, compared with of joint detection of several kinds of markers, the detection accuracy is greatly reduced, which cannot meet the requirement of detection discrimination.

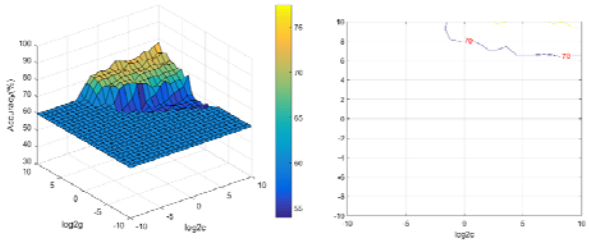

Figure 4. Learning process of HCY

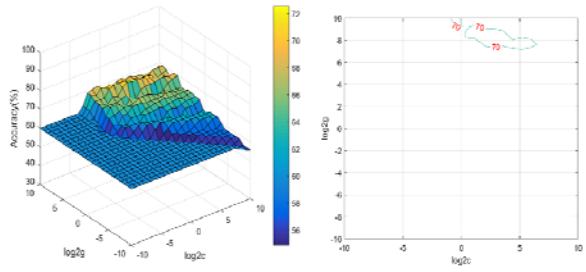

Figure 5. Learning process of FIB

According to simulation result, we can analyse each factor's effect on the onset of stroke. From figure 6-11, it can be concluded that HCY has the biggest influencing, it can achieve over 70% accuracy rate during the studying process. FIB and BUN also have a higher influence and the average rate of simulation perforation is 60 percent or more, followed by triglyceride and CRP. Blood sugar, general below 60% accuracy, is the lowest of the six. That is to say, Blood sugar could hardly distinguish between ischemic stroke and hemorrhagic stroke. It is a consistent biomarker of acute stroke and it can't be used to distinguish these two type of acute stroke.

From the forecasting result, we can also see that when only using single index detection, the correct rate of prediction low, which indicates that when using a single index, it can't be used to identify the different forms of stroke. The use of combined detection can greatly improve the prediction efficiency and accuracy.

TABLE V. SPECIFICITY OF THE BIOMARKER ON JUDGING STROKE

Biomarker	hemorrhagic stroke	ischemic stroke	Correct index(specificity+ sensitivity-1)
HCY	0%	100%	0%
FIB	27.78%	92.59%	20.37%
BUN	0%	100%	0%
triglyceride	0%	100%	0%
CRP	5.56%	96.30%	1.86%
sugar	5.56%	100%	5.56%

Conclusion

Due to the low cost and the development of Bioinformatics, machine learning draws an increasing attention to many countries and plays an important role in the diagnosis and prevention of various diseases. This paper demonstrates the biomarkers of stroke. First, six comprehensive and independent indexes are got by pre-treated data with the mutual information analysis and principal component analysis. A SVM-based approach is proposed to distinguish patients from patients and normal people. The validation results of the SVM method yields up to 85% classification accuracy with high sensitivity and specificity. Further, we analyse the importance of each marker in stroke prediction or diagnosis.

Acknowledgment

This research is supported by the National Science Foundation of China (Grant No. 81571142), the National Science Key Foundation of China (Grant No. 61232009), the National Science Foundation of China (Grant No. 60973106), Astronautic support Fund (Grant No. 374007), and National 863 Project of China (Grant No. 2011AA010404). Thanks to Xuanwu hospital for data support and assistance.

References

[1] Vos T, Barber R M, Bell B, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013[J]. Lancet, 2015, 386(9995): 743-800.

[2] Feigin V L, Forouzanfar M H, Krishnamurthi R, et al., "Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010". Lancet, 2014,383(9913): 245-254.

[3] Collaborators M C O D. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013[J]. Lancet, 2015, 385(9963):117-171.

[4] Jimeno A, Hidalgo M. Molecular. "biomarkers: their increasing role in the diagnosis, characterization, and therapy guidance in pancreatic cancer". Molecular Cancer Therapeutics, 2006, 5(4): 787-96.

[5] De Rubertis G, Davies S W. Genetic. "circuit design from an electronics perspective". Molecular, Cellular and Tissue Engineering, 2002. Proceedings of the IEEE-EMBS Special Topic Conference on IEEE, 2002: 145-146.

[6] Brophy J A, Voigt C A., "Principles of genetic circuit design. Nature Methods", 2014,11(5): 508-520.

[7] Wu Zhao,Hong Chen,Longfeng Xiang,Xiaomei Xie,Min Chen,Zuli Zhao,Qi Li."passive acoustic detection of diver based on SVM".Proceedings of 2016 IEEE, 2016: 623-628

[8] LI TianJiang, DU Qiang, "Abstract principal component analysis". science china, 2013, 56(12): 2783-2798

[9] Thomas M.Cover,Joy A.Thomas."elements of information theory(second edition)", A john wiley&sons, inc., 2006: 19-22

[10] LIN Ri-wu,CHEN Zao-shu,XU Yi-ye, "research on correlation of random blood clucose, CRP, FIB after the first acute stroke and atroke", modern preventive medicine,2012, 39(3): 759-760

[11] LI Zijian, "Recent advance in early diagnostic biomarkers of ischemic stroke", China science and technology information,Jan, 2014: 140-141