Stroke Prediction Based on Support Vector Machine

Hanqing Zhang*

Naval Architecture and Ocean Engineering, Department of Shipping and Ocean Engineering, Shanghai Maritime University, Shanghai, 201306, China

*Corresponding author: 201910413047@stu.shmtu.edu.cn

Abstract. Stroke needs to be solved as soon as possible because it has made so many people die or become disabled around the world. Therefore, the prediction of stroke is of great importance. In this paper, in order to verify the feasibility of stroke prediction by machine learning, SVM is proposed to predict the stroke. We construct the SVM model to map features about patients' relevant information to stroke. We use the real dataset to predict the stroke and compare its result with the results of some other models. As a result, we find that SVM can predict the stroke effectively and its result is superior to other's. Hence, stroke prediction based on SVM can be applied in real life.

Keywords: Support Vector Machine (SVM), stroke prediction, Machine Learning.

1. Introduction

Stroke is responsible for most disability and death around the world [1]. The fatality rate of stroke is very high. To be precise, from 2000 to 2019, stroke is responsible for about 11% of total deaths, which is the second main cause of death [2]. Therefore, detecting and diagnosing stroke are very important. Besides, with the development of Machine Learning, many methods and models have been applied to the medical field. Hence, carrying out stroke diagnosis based on Machine Learning is of great importance for theoretical research and practical application in the future.

Currently, the diagnosis and prediction of stroke mainly depend on experts in the medical field, who determine whether the patient has a stroke according to their professional knowledge and abundant work experience. However, this method is not stable. First, it is quite time-consuming to cultivate such an expert in the medical field. Second, the result of the diagnosis might be influenced by a doctor's physical and psychological condition. Therefore, an auxiliary diagnostic system should be established with the help of Machine Learning. Until now, aiming at the prediction of stroke, scholars at home and abroad have conducted a lot of research and many models are applied to the prediction of stroke, such as deep neural network (DNN)[3], K Neighbors Classifier[3], and XGBoost[4]. Although the above models have obtained good results, they still have some drawbacks. For example, DNN is too complex, the accuracy of K Neighbors Classifier is low, and XGBoost requires too much data. Here, the paper proposes the prediction of stroke on the basis of Support Vector Machine (SVM) in order to solve the problems above. Besides, the SVM model is compared to other machine-learning models in terms of stroke prediction. The factors that might cause stroke are identified, the model to predict stroke is established, and the dataset with 5110 data is employed to predict stroke.

The introduction of the methodology is conducted in the next part. And the discussion of experiments and results is conducted in part 3. In the fourth section, the conclusion is put forward. In the last section, the references used in this paper are listed.

2. Methodology

In this section, the principle of SVM and the application of SVM in stroke prediction are introduced.

2.1. The principle of SVM

Support vector machine (SVM) can be utilized to do binary classification. It originated in statistical learning theory (SLT) developed by Vapnik in 1995[5]. It drives from Vapnik-Chervonenkis theory and the structural risk minimization principle. It can be used to tackle the issue of binary classification by constructing the hyper-plane to separate positive class and negative class. SVM can do well in generalizing the issue and thus reaching the goal in statistical learning. SVM performs better when it comes to not over-generalization while other models, such as neural networks, might easily end up with over-generalization. However, in the classification process, the value cannot be either too large or too small. If the parameters are too low in complexity, the performance of classification is low and vice versa [6].

2.1.1 Optimal Hyper Plane (OHP)

Whether the dataset is linearly separable or linearly inseparable, the aim of SVM is to construct an OHP to separate the patterns. So OHP is the core concept of SVM.

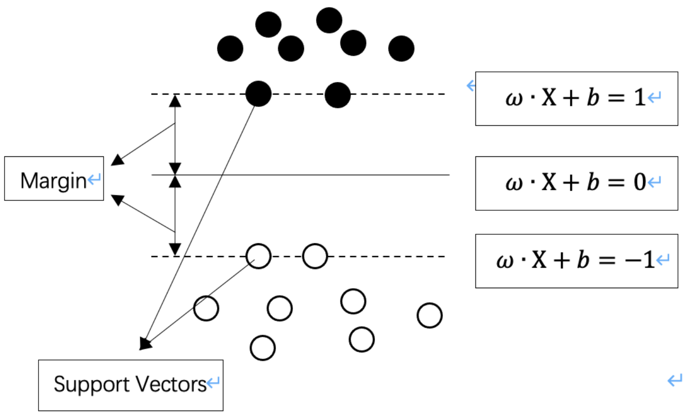

Fig. 1 Optimal Hyper Plane

Figure 1 shown above is the basic idea of hyper-plane. It describes how it looks like when two different patterns are separated by a hyper-plane, in the two-dimensional plane. There are three different lines in Figure 1. The line $wx+b=0$ is known as the marginal line which is the hyper-plane in the two-dimensional plane. The lines $wx+b=1$ and $wx+b=0$ are the lines on the either side of the marginal line. The perpendicular distance between the marginal line and the lines $wx+b=1$ and $wx+b=0$ is known as margin. Support vectors refer to those data points located nearest the hyper-plane.

The optimal classification in the SVM refers to that two different patterns are separated correctly and maximize the margin. If $||w||$ represents the norm of x, the margin is $2/||w||$. Therefore, if we want to maximize the margin, $||w||$ should be minimized.

2.1.2 Linear support vector machine

According to the discussion above, the aim of SVM is to find the linear function: $g(x)=wx+b$, which can separate two different patterns correctly. In order to maximize the margin, the optimization objective function below is established.

$$\min_{\omega,b} \frac{1}{2}\omega^T \cdot \omega \tag{1}$$

$$s.t. \quad y_i(\omega^T \cdot x_i + b) \geq 1, i = 1,2,\dots,n$$

Because the objective function above is quadratic programming, we need to construct a Lagrange function [7].

$$L = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\omega^T \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i \tag{2}$$

Here, α_i and β_i are Lagrange multipliers.

To solve the objective function, we need to compute the partial derivatives of *w, b* and ξ_i and let them be equal to zero. Then, this issue can be converted into duel problem.

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_i y_j\alpha_i\alpha_j\left(x_i\cdot x_j\right)-\sum_{i=1}^{n}\alpha_i \tag{3}$$

$$s.t.\quad \sum_{i=1}^{n}\alpha_i y_i=0$$

$$0\le\alpha_i\le C,\quad i=1,\dots,n$$

The solution of the optimization problem above is α_i^*. When $0<\alpha_i^*<C$, α_i^* is the normal support vector. Here, the Karush-Kuhn-Tucker (KTT) Conditions are shown below:

$$\alpha_i^*\left[y_i(\omega^*\cdot x_i)+b^*-1+\xi_i^*\right]=0 \tag{4}$$

$$\beta_i^*\cdot\xi_i^*=0$$

The solution $b^*$is shown below:

$$b^*=y_i-\omega^*\cdot x_i=y_i-\sum_{i=1}^{n}y_i\,\alpha_i^*\left(x_i\cdot x_j\right) \tag{5}$$

And the decision function is shown below:

$$f(x)=sgn\left(\sum_{i=1}^{n}y_i\,\alpha_i^*(x_i\cdot x)+b^*\right) \tag{6}$$

Here, b^* is the threshold value in the issue.

Hence, to identify vector *x* belongs to which classification, we just need to use the decision function *f(x)* above to predict the classification of *x*.

2.1.3 Non-linear support vector machine

For those linear inseparable patterns, SVM maps them into a new space usually a higher dimension space in order to make these patterns become linearly separable in a higher dimension space [8]. This process is shown in Figure 2 below.

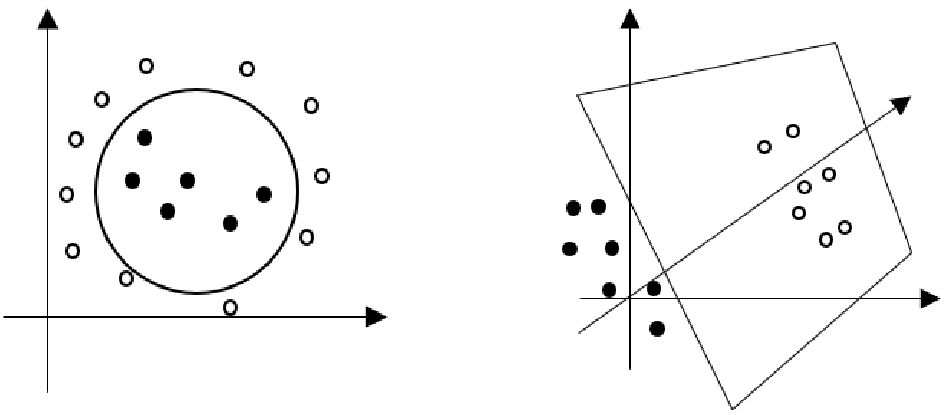

Fig. 2 Hyper-Plane in high-dimensional space

Under the circumstances, the patterns are linearly separable and we can use SVM to find the optimal hyper-plane. However, to solve the problem of the curse of dimensionality caused by data mapping, SVM introduces the concept of kernel function *K(xi,xj)*.

$$K\left(x_i,x_j\right)=\Phi(x_i)\cdot\Phi\left(x_j\right) \tag{7}$$

However, quadratic programming resulting from replacing inner product operation with kernel function is still a convex optimization problem. So it can be converted in to a dual optimization problem below:

$$\max_{\alpha} Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i x_j) \tag{8}$$

$$s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \le \alpha_i \le C, i = 1,2,\dots,n.$$

And the decision function is:

$$f(x) = sgn\left(\sum_{i=1}^{n} \alpha_i^* y_i \, K(x_i \cdot x) + b^*\right) \tag{9}$$

Hence, to identify vector *x* belongs to which classification, we just need to use the decision function *f(x)* above to predict the classification of *x*.

2.2. The application of SVM in stroke prediction

In this section, I will introduce the process of using SVM to predict stroke.

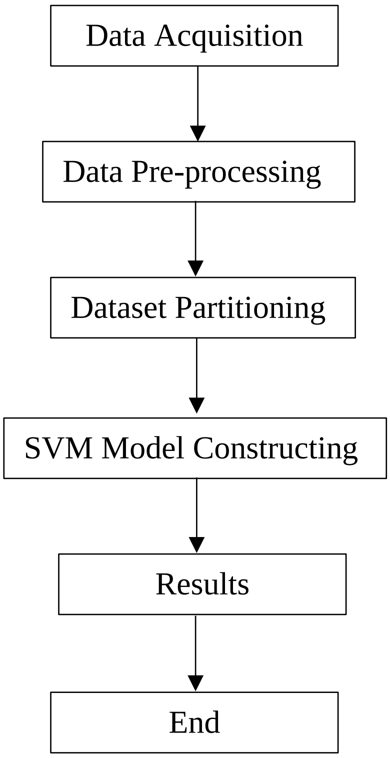

Fig. 3 The process of application of SVM in stroke prediction

In step 1, the data that will be used later are acquired. In step 2, we pre-process the data, including data cleaning, feature selection, and feature extraction, etc. In step 3, the dataset is divided into two different sets responsible for training and testing respectively. In step 4, a SVM model is constructed. In step 5, we can get the result of the prediction.

3. Experiment and discussion

In the section above, we have introduced the methodology. In this section, we use the real data and conduct the experiment to verify the feasibility of the methodology.

3.1. Data Description

The dataset comes from Kaggle[9]. It is a numeric dataset. In this dataset, the input parameters, such as age, smoking status, and various diseases, etc., are utilized to predict whether a patient might get a stroke. Each patient's relevant information is provided in each row of this dataset. Table 1 below describes some basic information about the dataset.

Table 1. Basic information about the dataset

Item name	Description
Sample size	5110
The number of features	12
Label	1: stroke. 0: no stroke

3.2. Design of experiment

In this section, the contents and objective of the experiment, evaluation indices and the setting of the experiment are introduced.

3.2.1 The contents and objective of the experiment

(1) To verify the effectiveness of SVM, SVM is applied to the dataset of stroke prediction.

(2) To verify the superiority of SVM, some other methods of machine learning are used to do contrast experiments.

3.2.2 Evaluation metrics

There are some evaluation metrics to show the ability of different models. The names and functions of these evaluation metrics are shown in Table 2.

Table 2. Evaluation metrics

Evaluation metrics	Function
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$\frac{2\times Precision\times Recall}{Precision+Recall}$

3.2.3 The setting of the experiment

The kernel function of SVM used in this article is the Gaussian kernel function [10], which is shown below.

$$K(x_i, x_j) = e^{-\frac{\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}} \tag{10}$$

Two different sets make up the dataset. They are the training set and the test set. In this experiment, the former one is responsible for 75% of the dataset and the latter one makes up 25% of the dataset.

3.3. Experiment results and analyses

(1) SVM can predict whether patients get the stroke according to his or her relevant information in the dataset. The performance metrics of SVM is shown in Table 3 below.

Table 3. The performance metrics of SVM

Metric	Score
Accuracy	0.792
Precision	0.712
Recall	0.912
F1-score	0.8

From Table 3, it is clear that the accuracy of the prediction is 0.792, the precision of the prediction is 0.712, the recall of the prediction is 0.912 and the F1-score of the prediction is 0.8.

(2) Some models including Logistic Regression, KNN, Decision Tree, Naïve Bayes, SVM and Random Forest are used to do the same prediction, and we can get the performance metrics of these models in order to verify the superiority of SVM. Table 4 below shows the performance metrics of these models.

Table 4. The performance metrics of some models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.752	0.710	0.772	0.739
KNN	0.76	0.737	0.736	0.737
Decision Tree	0.728	0.677	0.772	0.721
Naïve Bayes	0.768	0.733	0.772	0.7521
SVM	0.792	0.712	0.912	0.8
Random Forest	0.76	0.696	0.842	0.762

From Table 4 above, even though the Precision of SVM ranks third, other performance metrics of SVM are all first. And we can clearly discover this relation from Figure 4 below.

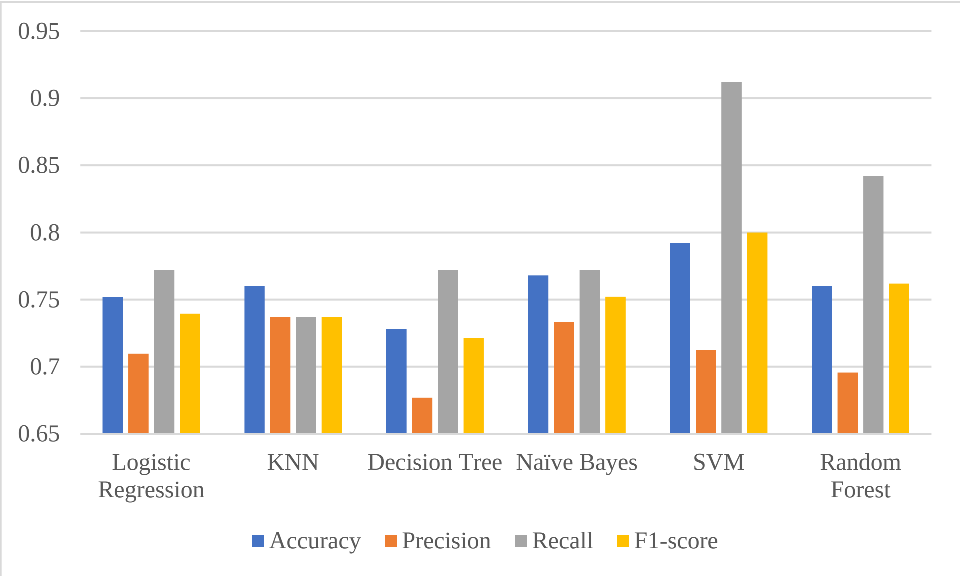

Fig. 4 The performance metrics of some models

From Figure 4 above, we can easily see that when it comes to Accuracy, SVM performs best, when it comes to Precision, KNN performs best, when it comes to Recall, SVM performs best, and when it comes to F1-score, SVM performs best.

From the results above, we can conclude that SVM can be used to predict the stroke and SVM performs best among all the models above.

4. Conclusion

This paper discusses whether machine learning can be applied to stroke prediction. To solve the issue of stroke prediction, the method of SVM is proposed. In this paper, the SVM model whose kernel function is the Gaussian kernel function is constructed to map eleven different features to stroke. And two experiments are conducted in this paper in order to verify the proposal above. The result shows that when it comes to stroke prediction, the Accuracy of SVM is 0.792, the Precision of SVM is 0.712, the Recall of SVM is 0.912, and the F1-score of SVM is 0.8. Clearly, the performance metrics of SVM are high, especially Recall. Hence, it can be concluded that SVM is able to predict the stroke effectively and performs better than other methods of machine learning. And stroke prediction based on SVM can be applied to the issue in real life.

Reference

[1] Pastore D, Pacifici F, Capuani B, et al. Sex-genetic interaction in the risk for cerebrovascular disease[J]. Current medicinal chemistry, 2017, 24(24): 2687-2699.

[2] The top 10 causes of death, 9 December 2020, 28 September 2022, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[3] Cheon S, Kim J, Lim J. The use of deep learning to predict stroke patient mortality[J]. International journal of environmental research and public health, 2019, 16(11): 1876.

[4] Jiang Chao;Chen Tian-Ge;Du Xin;Li Xiang;He Liu;Lai Yi-Wei;Xia Shi-Jun;Liu Rong;Hu Yi-Ying;Li Ying-Xue;Jiang Chen-Xi;Liu Nian;Tang Ri-Bo;Bai Rong;Sang Cai-Hua;Long De-Yong;Xie Guo-Tong;Dong Jian-Zeng;Ma Chang-Sheng.A simple and easily implemented risk model to predict 1-year ischemic stroke and systemic embolism in Chinese patients with atrial fibrillation[J].Chinese Medical Journal,2021,134(19):2293-2298.

[5] Vn V. The nature of statistical learning theory[J]. NY: Springer-Verlag, 1995.

[6] Pradhan A. Support vector machine-a survey[J]. International Journal of Emerging. Technology and Advanced Engineering, 2012, 2(8): 82-85.

[7] Durgesh K S, Lekha B. Data classification using support vector machine[J]. Journal of theoretical and applied information technology, 2010, 12(1): 1-7.

[8] Peng Lin, Research on Diabetes Prediction Model Based on SVM, Jilin University, 1 May 2022.

[9] Fedesoriano, Stroke Prediction Dataset, 28 September 2022, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[10] Keerthi S S, Lin C J. Asymptotic behaviors of support vector machines with Gaussian. kernel[J]. Neural computation, 2003, 15(7): 1667-1689.