Support Vector Machine for Stroke Risk Prediction

Yifan Feng*

Department of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, 201620, China

* Corresponding Author Email: 19057010@suibe.edu.cn

Abstract. Nowadays, there are many paralytic patients sent to the hospital, and it is impossible for hospitals to search out those particular patients at once because these patients share the same symptoms such as unconsciousness as those hypoglycemic patients. Nowadays, most patients need to conduct blood tests to ensure whether they have a stroke or not. However, this will take a lot of time and resources. The investigation aims at stroke prediction. By using machine learning, the model will be used to train and test data. Support vector machine (SVM) will be used in this project. By using a support vector classifier (SVC), the model will be trained to learn from data. Then it will react to another data set to find out if it is fitted. As a classification problem, the accuracy is 0.77. It shows that certain model performs well after training which reflects that the prediction is successful. What's more, its high recall which is 0.83 means that the model of stroke prediction can surely offer some help to patient classification to some extent because it can find most of the paralytic patients among all the samplers. Stroke prediction trained by the SVM model can help make the first division among all patients which can help save a lot of time and energy. However, since there is still a little deviation, it is still need to keep pace with modern medical technology to improve it.

Keywords: Machine learning; stroke prediction; support vector machine.

1. Introduction

It is known to all that stroke is deadly since it can cause irreversible damage to human beings by leading to poor blood flow. In the past, stroke used to affect those aged persons. Nevertheless, many young people are also suffering from a stroke for many different reasons: terrible life habits, imbalanced diets, lacking physical exercise. There are two kinds of stroke: ischemic stroke and hemorrhagic. The first one is caused by a lack of blood flow, while the latter one prevents parts of the brain from functioning correctly due to bleeding. There are many signs and symptoms of a stroke which can reflect someone may be affected by stroke, including an inability to move or feel on one side of the body, problems understanding, dizziness, and loss of vision to one side. These signs will remind those patients that they are in danger. However, these symptoms will only be visible when a stroke becomes serious. As a result, it is common for people to be unaware of suffering from a stroke, especially those people who are addicted to tobacco and junk food who will be more likely to suffer from a stroke. Stroke prediction will be extremely useful when the hospital needs to handle a great variety of patients, especially in the emergency department [1]. Nowadays, those patients will only receive medical care for their strokes only if they tell doctors to do so. With a method of stroke prediction, hospitals can order treatment for those patients who cannot truly their symptoms at once. After a precise prediction of the stroke, it will be easier for a hospital to deal with the patient who is potentially a stroke patient, thus he will be offered immediate medication to avoid further damage. As a result, many scholars have paid more attention to the field of stroke prediction. Nowadays, there have already been many ways of stroke prediction. For example, it has already been a mature method to evaluate stroke risk in patients with AF by assessing the value of the prediction [2]. What's more, there is a relationship between blood pressure and stroke [3]. Nevertheless, most of them use the simple feature to predict a stroke and this is far from enough since many sources will lead to this harmful disease.

In this project, machine learning will be used to better solve this problem, making it possible to receive a conclusion that can be determined by various complex features. By using machine learning, it will be possible to imitate the real situation to tell the difference between stroke-patient and non-

stroke ones [4]. Data from the stroke will be split into two parts: the data set used to train the model and the other one for testing the model. Before building the model, a suitable data set about stroke will be chosen on Kaggle.com. After importing the data set, data will be pre-processed to make them for further training and testing [5]. In this investigation, SVM will be used to build a model to train the data set, and then value the accuracy.

2. Method

2.1. Dataset

Data set of this experiment is chosen on Kaggle about stroke. By applying SVM to these data, prediction will be made to find out it is possible to search out the relationship between stroke and many factors. To train and test the accuracy and recall of the prediction model, 5110 data have been collected. For each person in this data set, each one has 12 indices, including 11 features and one label. Features are id, gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, bmi, smoking status. The label stroke means whether this person suffers from stroke. After pre-processing, half of the collection will be used as training part, while the other half will be used as testing. Some of the samples are shown in Table 1 and the distribution of people having stroke is demonstrated in Figure 1.

Table 1. Samples from dataset

	Id	Gender	Age	Hypertension	Heart disease	Ever married
1	9046	Male	67.0	0	1	Yes
2	51676	Female	61.0	0	0	Yes
3	60182	Male	80.0	0	1	Yes
4	1665	Female	49.0	0	0	Yes
5	56669	Male	79.0	1	0	Yes
	Work type	Residence type	Average glucose level	BMI	Smoking status	Stroke
1	Private	Urban	228.69	36.6	formerly	1
2	Self-employed	Rural	202.21	NaN	never smoked	1
3	Private	Rural	105.92	32.5	never smoked	1
4	Private	Urban	171.23	34.4	smokes	1
5	Self-employed	Rural	174.12	24.0	never smoked	1

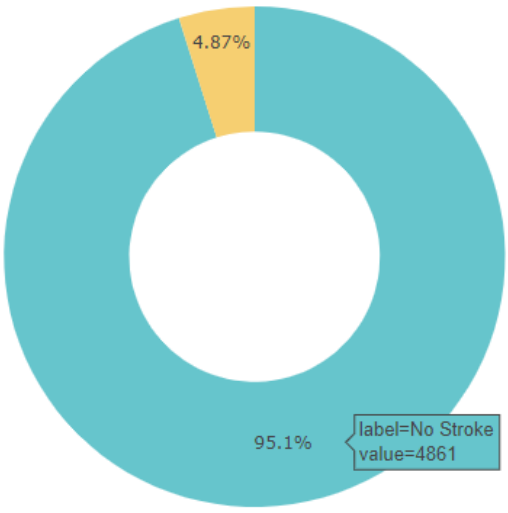

Fig. 1 Distribution of people having the stroke

2.2. Data-preprocessing

There are some flaws in this data set, which leads to bigger errors in the further modeling. As a result, preprocessing is a must before using these data. Before preprocessing, the dataset is required to be checked to find out if there is any missing value in each feature. Shown in Table 2.

Table 2. Proportion of Null values.

	Id	Gender	Age	Hypertension	Heart disease	Ever married
null (%)	0.00	0.00	0.00	0.00	0.00	0.00
	Work type	Residence type	Average glucose level	BMI	Smoking status	Stroke
null (%)	0.00	0.00	0.00	3.93	0.00	0.00

All features are good except for some missing in bmi, so it is required to make some improvement to make this data set neat.

Two methods were used to fill the missing: The first one is filling data randomly, while the second one is to fill the missing ones by using mean. Three distributions are compared in Figure 2.

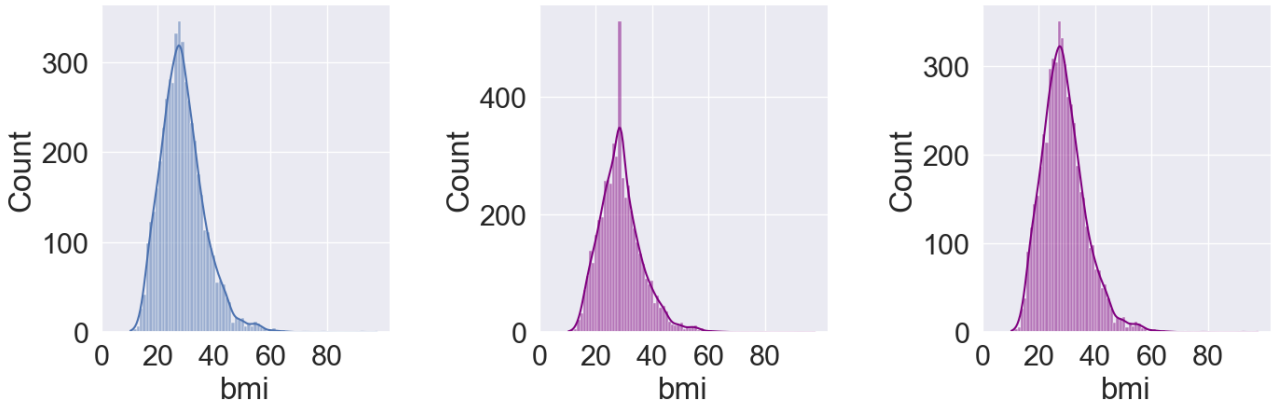

Fig. 2 BMI distribution. From left to right are the original distribution, the distribution of random filling, and the distribution of mean filling respectively.

It is easy to find out the first method is far more different from the original distribution graph than the second method which means that filling the missing data with the mean will help improve the accuracy of train model.

There is also an outliner in gender, an 'other' which may cause an error in the following encoding, so it is dropped. Label Encoder is used to transcribed those feature elements into numbers, so that they can be better scaled. Information is shown in Table 3.

Table 3. Encoded features

	Id	Gender	Age	Hypertension	Heart disease	Ever married	Work type	Residence type
0	9046	1	67	0	1	1	2	1
1	51676	0	61	0	0	1	3	0
2	31112	1	80	0	1	1	2	0
3	60192	0	49	0	0	1	2	1
4	1665	0	79	1	0	1	3	0

Before choosing the feature selection, use MinMaxScaler to put all feature into the same range. So the age, hypertension, heart disease, glucose level and ever married are kept, which have larger correlation with stroke according to the correlation gradient listed in Table 4.

Table 4. Feature correlation

	Id	Gender	Age	Hypertension	Heart disease	Ever married	Work type	Residence type
Id	1.000	0.002	0.003	0.004	-0.001	0.014	-0.015	-0.001
Gender	0.002	1.000	-0.028	0.021	0.086	-0.030	0.057	-0.006
Age	0.003	-0.028	1.000	0.276	0.264	0.679	-0.361	0.014
Hypertension	0.004	0.021	0.276	1.000	0.108	0.164	-0.052	-0.008
Heart disease	-0.001	0.086	0.264	0.108	1.000	0.115	-0.028	0.003
Ever married	0.014	-0.030	0.679	0.164	0.115	1.000	-0.353	0.006
Work type	-0.015	0.057	-0.361	-0.052	-0.028	-0.353	1.000	-0.007
Residence type	-0.001	-0.006	0.014	-0.008	0.003	0.006	-0.007	1.000

According to Figure 2, the imbalanced data need to be balanced by using random over-sampling and then split data into train and test. In List 5, data containing ‘stroke’ equals to that one of ’no stroke’. Finaly there are 4860 subjects for each category.

2.3. SVM

SVM is called Support Vector Machine, a supervised learning model which is connected with learning algorithms that analyzes data for classification [6]. It is one of the strongest prediction methods. The advantages are quite obvious. For example, it will be very efficient in higher dimensional space. A value small enough for λ will be chose to yield the hard-margin classifier to make data input to become linearly classifiable.

Computing the SVM classifier amounts to minimizing an expression of the form [7]:

$$\left[\frac{1}{n}\sum_{i=1}^{n} max(\,0,1-y_i(w^T x_i-b))\right]+\lambda\|w\|^2 \tag{1}$$

To solve problem mentioned above, a kernel function k which satisfies[8]:

$$k(x_i,x_j)=\phi(x_i)\cdot\phi(x_j) \tag{2}$$

is given. At the same time, the classification vector w needs to satisfy $w=\sum_{i=1}^{n} ciyi\phi(xi)$. In the form mentioned above, c_i are obtained by solving problem:

$$\text{Maximize } f(c_1 \dots c_n)=\sum_{i=1}^{n} c_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i\,(\phi(x_i)\cdot\phi(x_j))y_j c_j \tag{3}$$

subject to $\sum_{i=1}^{n} c_i y_i=0$, and $0\le ci\le\frac{1}{2n\lambda}$for all i.

To find some index i such that $0\le c_i\le\frac{1}{2n\lambda}$, the formula mentioned below also need to be worked out:

$$b=w^T\phi(x_i)-y_i=[\sum_{j=1}^{n} c_j y_j\phi(x_j)\cdot\phi(x_i)]-y_i=[\sum_{j=1}^{n} c_j y_j\,k(x_j,x_i)]-y_i \tag{4}$$

Finally,

$$z\mapsto sgn(\,w^T\phi(z)-b)=sgn(\,[\sum_{i=1}^{n} c_i y_i k(x_i,z)]-b) \tag{5}$$

Since the stroke prediction is actually a problem of classification, SVC is used to solve the problem [9]. Parameters used in SVC (C, kernel, degree, gamma) include: C =1.0, which is inversely proportional to the strength of the regularization; Degree = 3; Gamma = "scale".

Finally, the prediction values will be compared with test data to test whether they are different from each other. These data will be collected to further evaluate the model by using confusion matrix, precision and recall.

2.4. Evaluation Matrix

Precision and Recall. Precision means the proportion of the correct prediction in results which are meant to be positive. Recall stands for the proportion of the correct prediction in results which are turned out to be positive.

$$V_{precision} = \frac{TP}{TP+FP} \tag{6}$$

$$V_{\text{recall}} = \frac{TP}{TP+FN} \tag{7}$$

Specificity. Specificity means the proportion of the correct predictions in the negative.

$$V_{\text{spec}ificity} = \frac{TN}{TN+FP} \tag{8}$$

F1-score. F1-score is the average with power of Precision and Recall. It is ranged from 0 to 1.

$$V_{F1} = \frac{2 \cdot V_{precision} \cdot V_{recall}}{V_{precision} + V_{recall}} \tag{9}$$

3. Result

The performance is demonstrated in Table 5. The accuracy number is 0.77 which means that almost 8 people in ten can be truly predicted when they are sent to the hospital and they can be detected and ordered the true treatment. Precision of this model is 0.74, which means that 7 people will be turned out to have a stroke in the range of those people who are predicted to suffer from a stroke. At the same time, result of recall is 0.83,which equals to the fact that every 8 people can be searched out in all patients. Last but least, F1-score, the average with power of precision and recall, is 0.78.

The result of SVM model is totally good because all of its evaluation numbers are near and even above 0.75, which means that at least three quarters of the patients can be examined at a correct result and can receive suitable treatment. Among all these statistics, result of recall is the highest, reaches 0.83. Since stroke prediction is a classification problem, which aims to search out all stroke patients and order particular treatment. Under this circumstance, recall number plays the most important role in stroke prediction.

In the confusion matrix, shown in Figure 3, it is easy to find that except for recall and precision, there are also other important features. For instance, for those stroke patients who are not found by the system, the number of those people is only 161, in a total number of 1944. The rate is 8.29% which means that only a few patients will not receive immediate care and for most of patients, their physical health can almost be protected. In another aspect, the number of those healthy people who will be misunderstood as unhealthy is 290, at a rate of 14.92%. This reflects those hospitals will not waste too many medical resources on those people who do not need medical care at all and can make sure that more people in need can receive aids.

Table 5. Result of classification

Accuracy	0.77
Precision	0.74
Recall	0.83
F1_score	0.78

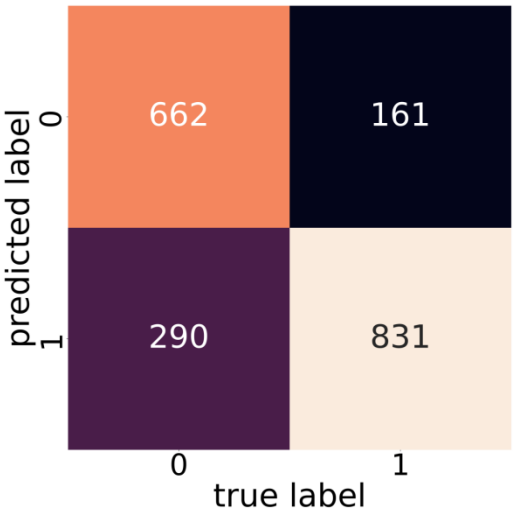

Fig. 3 Result measured by confusion matrix

4. Discussion

The result is nice because it has a good accuracy and recall number which shows that this prediction will work in a proper function in reality. Since the data set is formalized and trained by using SVM model, it is sure that this prediction will not be totally wrong. However, the accuracy and recall is less than 90 percent, mainly because for two reasons. Firstly, the data set is over-sampled to eliminate the imbalance of the original data. At the same time, the data set is enlarged which also make it more difficult to reach a high accuracy. Secondly, compared to neuron network, result will not be as good when the data is trained by SVC because data will be divided into different layer by using network, thus having a higher accuracy and recall number.

To improve the result, there are several ways. First and foremost, to improve this model, more information can be collected so that the data set will be large enough to avoid using over-sampling [10]. For instance, hospitals can build a data base to import and storage new data to enhance the accuracy of the model. What's more, it is also possible to improve performance of the model by using advanced skills, such as modifying the hypertext by using different ways or combine SVC with other methods such as KNN to work out the most suitable one [11].

There is some limitations of this prediction. To begin with, all features used in the prediction are fixed. However, it is a common knowledge that there are also other things which can cause stroke. Many patients may even inherit from his parents. In addition, patients can be correctly predicted only if their health statistics are exactly same as one of the data used in the model. It is known to all that even a little difference will lead to a huge distinction in the result of prediction. Last but not least, many patients will not be likely to share their own private data so that it will be tough to employ this prediction model in reality.

However, although there are still some problems and limitations, stroke prediction is still helpful to some extent.

5. Conclusion

To sum up, the accuracy of the model is 0.77, and the recall of that is 0.83. It is easy to find that the result is believable because of its excellent accuracy. Default hypertext makes it perform well but

it also accounts for the reason why it can't perform perfectly even if hypertext is changed to try to improve this model. Since stroke prediction is a classification problem, recall plays an important role in how many patients can be searched out accurately. To further improve stroke prediction, they are several methods. The most important one is to enlarge the data set by importing more real data to avoid using over-sampling. A decrease in accuracy and recall is unavoidable when over-sampling is exalted to eliminate the imbalance of the data set. As a result, there will be more samples for a model to get trained. What's more, it will also give a hand if stroke prediction can combine with estimations of other particular results which also lead to a stroke such as blood tests. Consequently, doctors can make decisions based on different examinations to make comprehensive predictions.

This prediction offers another perspective to test these patients and search out those people who suffer from a stroke to get an immediate cure. In the future, after the technology of stroke prediction is mature enough, it can help those hospitals tell the differences between paralytic patients and patients to suffer from other diseases without so many various tests which will waste a lot of time and medical resources. In another aspect, for these patients, it is almost impossible for everyone to know exactly about their physical conditions. With the help of stroke prediction, communication between doctors and patients will be more fluent.

In the future, the prediction model will be modified to improve its performance by adding more natural and real information into the train data set. Although accuracy and recall are not so high to make a perfect prediction at present, in the future this model will be perfected.

References

[1] Caulfeild S, Pak S, Yao N, et al. Stroke Prediction. Mechanical engineering and Automation, 2021, 011(006):180-182.

[2] Waterman A D, Shannon W, Boechler M, et al. Validation of Clinical Classification Schemes for Predicting Stroke. JAMA The Journal of the American Medical Association, 2001, 285(22):20-21.

[3] Mcmahon S, Richard P, Jeffrey C, et al. Blood Pressure, Stroke and Coronary Heart Disease. The Lancet, 1990, 335(8692):765-774.

[4] Khosla A, Cao Y, Lin C Y, et al. An integrated machine learning approach to stroke prediction. ACM International Conference on Knowledge Discovery & Data Mining. 2010, 183-192.

[5] Letham B, Rudin C, McCormick T H, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 2015, 9(3): 1350-1371.

[6] Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2011, 2(3): 1-27.

[7] Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression. Neural networks, 2004, 17(1): 113-126.

[8] Tsang I W, Kwok J T, Cheung P M, et al. Core vector machines: Fast SVM training on very large data sets. Journal of Machine Learning Research, 2005, 6(4):363-392.

[9] Wien M, Schwarz H, Oelbaum T. Performance analysis of SVC. IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(9): 1194-1203.

[10] Xiong Q, Zeng X. Application and improvement of SVM method in precipitation forecast. Meteor. Mon, 2008, 34(12): 90-95.

[11] Li C, Ding Y, Wen G. Application of SVMKNN combination improvement algorithm on patent text classification. Computer Engineering and Applications, 2006, 42(20): 193-212.