Prediction Model of Stroke Based on Few-Shot Learning

Xujun Wu* and Guoxin Wang
Yantai Vocational College, Yantai 264670, Shandong, China

Keywords: Stroke, Machine Learning, Few-Shot Learning, Decision Tree, Bagging Algorithm.

Abstract: Stroke has the characteristics of high morbidity, high disability rate and high mortality, and has become the first cause of death in China; timely screening of specific populations and prediction through prediction models are of great significance for disease risk control. In this paper, we use machine learning algorithm to intelligently process the stroke screening data, and reasonably expand the small sample data. On this basis, we use decision tree and Bagging algorithm to establish a stroke prediction model, and discuss the modeling process and model parameter selection in detail. The test results show that the prediction model based on the extended data set runs well on the test data set, and this method provides a reference for Few-shot learning modeling.

1 INTRODUCTION

"Stroke" is also known as "cerebrovascular accident" (CVA). It is an acute cerebrovascular disease, which is a group of diseases caused by brain tissue damage due to sudden rupture of cerebral blood vessels or blood blockage that prevents blood from flowing into the brain. The survey shows that stroke in urban and rural areas has become the first cause of death in China, and also the primary cause of disability for Chinese adults. Stroke is characterized by high morbidity, mortality and disability. There are many etiological factors of stroke, among which vascular risk factors are considered to be the most important. Other important factors include bad living habits, gender, age, race, etc. At present, there are more than 10 million new strokes worldwide every year, which brings serious burden to society and families. Due to the lack of effective treatment, effective stroke prevention measures are urgently needed. According to the 2013 Global Burden of Disease (GBD) study, more than 90% of strokes are caused by adjustable risk factors, and more than 75% of stroke patients can reduce their occurrence by controlling metabolic and behavioral risk factors. Daily living habits are the most controllable factor and the most effective prevention strategy to prevent cerebrovascular diseases, such as smoking control, alcohol control, salt control, keeping good habits of work and rest, which belong to the primary prevention means; Secondary prevention measures include drug intervention, smoking cessation, blood sugar control and other measures.

The establishment of stroke prediction model is of great significance for the prevention of stroke. Stroke prediction can avoid unnecessary waste of medical resources. In view of this, many recurrence risk models of ischemic cerebrovascular disease have been established internationally. Early prediction models mainly use risk factors as the main evaluation content, which can effectively stratify the risk of ischemic cerebrovascular disease, but their reliability and validity are limited. In recent years, with the rapid development of neuroimaging and the popularization of MRI and vascular imaging, more and more studies have pointed out that imaging markers have become important markers of stroke recurrence.

The development of big data and artificial intelligence has provided opportunities for the establishment of data based prediction models. The machine learning method is used to analyze and mine the data of stroke patients, extract some typical features and laws of patients, and establish an artificial intelligence model to predict, so as to provide methods and means for early prevention of stroke.

Wu, X. and Wang, G.
Prediction Model of Stroke Based on Few-Shot Learning.
DOI: 10.5220/0012042100003620
In *Proceedings of the 4th International Conference on Economic Management and Model Engineering (ICEMME 2022)*, pages 694-698
ISBN: 978-989-758-636-1

2 DATA PROCESSING AND SAMPLE EXPANSION

As early as 2016, China has formulated data standards for the registration of patients with various diseases. The standard document WS-375 specifies the metadata attributes and data element attributes of the basic data set of stroke registration reports, so as to provide them to disease prevention and control institutions, medical institutions providing relevant services and relevant health administrative departments for relevant business data collection, transmission, storage, etc. Due to the late start of medical data accumulation in China, the early data records are few and non-standard, and there is still no stroke data set of appropriate size. In this case, machine learning method based on small sample learning becomes an optional method for building prediction models.

2.1 Data Sources

The data in this paper mainly comes from two parts:

1. The original data of 300 cases obtained by the local medical institution from the screening of the target population in the community has more than 200 effective attributes of the screening data. In the 300 cases, 3 cases have been diagnosed, and others do not have the characteristics of stroke.

2. The data set of 300 confirmed patients provided by the local hospital over the years, including obvious stroke symptoms, and some cases have died. These registration forms are organized according to the WS-375 document standard.

Because there are so many attribute fields for these data, and many fields are missing data, these data can be used accurately only after they are cleaned, formatted, and normalized.

2.2 Data Processing

Data processing mainly deals with incomplete values, incorrect values and attribute types of data to make the data conform to the format required by the model. These data are processed using the following rules:

1. If all data of an attribute is missing, the attribute will be treated as invalid and deleted.

2. For categorical variables, if they are ordered categorical variables, the corresponding values will be assigned directly; for unordered category variables, use one-hot coding.

3. Some attributes with some missing data need to be handled as appropriate. If only a few instances of an attribute have data (less than 10%), and there are no stroke patients in the instances with data, delete the attribute; On the contrary, the same kind of mean value interpolation is used for processing.

4. For out of range or incorrect data, if it is numerical data, select the value closest to it. For nominal attributes, select different negative integers to distinguish.

In addition to the above processing rules, some attribute data are also standardized and normalized。

2.3 Establish Training Data and Test Data Set

In the screening data, there are only 3 confirmed cases. If machine learning algorithm is used, it will be difficult to learn the typical characteristics of patients, so that a reasonable prediction model cannot be built. If the accumulated 300 confirmed cases are combined into the screening data set, such simple combination will distort the entire learning process. Therefore, this paper randomly confuses the confirmed data with the screening data to establish a more reasonable data set, which has 400 instances, 30% of which are confirmed cases, so that the established data set can establish a more reasonable and accurate prediction model.

After data attribute processing, the available attributes are reduced from more than 300 to less than 100, and the data set size is about 400 records, including more than 130 patient records. See Table 1 for data attributes.

Table 1: Category attribute quantity.

Category	Attribute quantity
demographic characteristics	5
Personal History	21
family history	8
use medication history	3
history of treatment	20
health check indicators	20
chemical examination indicators	10
declaration of health	2
final diagnosis	2
other diagnosis	8

Figure 1 shows the age attribute of confirmed cases of stroke for the first time. It can be seen intuitively from Figure 1 that 55-80 years old are the age group with high incidence of stroke.

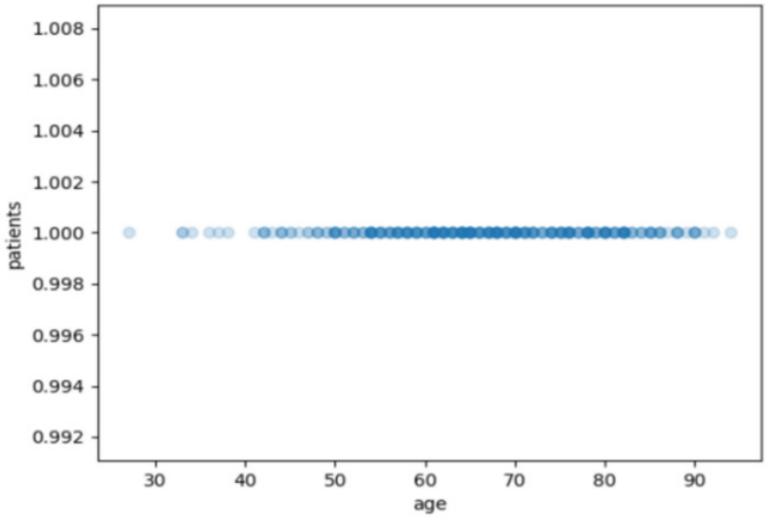

Figure 1: Age distribution of confirmed cases.

2.4 Classify Confirmed Cases Using K-Means Algorithm

In order to merge some confirmed case datasets with screening datasets, representative cases need to be selected. The laboratory indicators in Table 1 are the inspection indicators of the confirmed personnel when they are hospitalized. When screening the target population, these indicators do not exist. Therefore, the confirmed population is divided into several groups by K-means clustering calculation according to 10 laboratory indicators. Then, one instance is randomly selected from each group and put into the screening data set. Finally, it is combined into a new data set with reasonable distribution of positive and negative examples, establishment of prediction model for stroke.

After repeated trials, the family with the highest classification score is 3, followed by 9. Because of the consideration of selecting confirmed cases to participate in modeling, the confirmed patients are finally divided into 9 categories. From the nine categories of confirmed cases, 35% of each category was randomly selected to join the prediction data set. Classification of confirmed cases is shown in Figure 2.

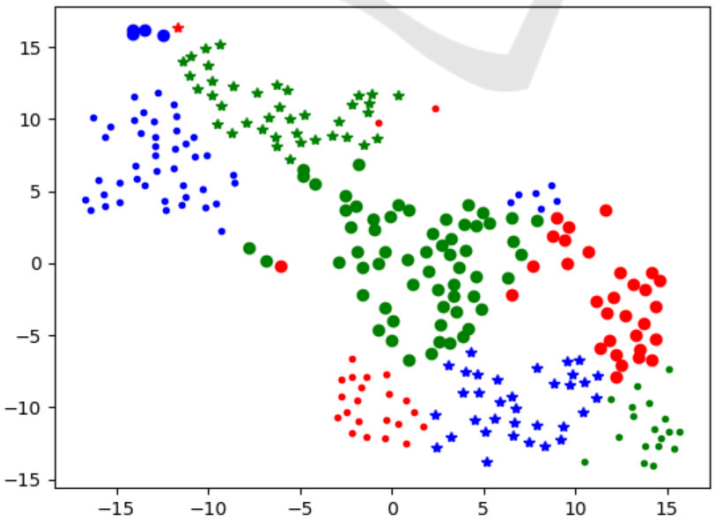

Figure 2: Classification of confirmed cases using K-means algorithm.

3 ESTABLISH PREDICTION MODEL

The data size of the collated dataset is about 400 instances, and its size is still very small. You can use the Few-shot learning method to establish a prediction model. Small sample learning usually uses PAC algorithm, but when the data characteristics are obvious, the traditional integration method can also achieve good results. Because most attribute types in the current data set are nominal attributes, it is suitable to use the integration method to establish a prediction model. A model with good performance can be obtained by training the model with sample data and verifying the model with test data。

The integration method is composed of two layers of algorithms. First, we need to use a basic learner of a single machine learning algorithm, and then integrate this algorithm into an integrated method. Generate many models by integrating algorithms, and finally select the model output with the best performance.

3.1 Decision Tree

Decision tree learning is one of the most widely used inductive learning algorithms, which is robust to noisy data. In this paper, decision tree learning is used as the basic learner of the integration method. The key of decision tree training lies in the selection of segmentation points and the depth of the decision tree. Generally, after the depth of the decision tree reaches a certain value, an optimal value is obtained, and then the error will slowly increase when the depth of the decision tree is increased. However, the lowest value is not necessarily the best depth of the integrated algorithm, and the best classification effect can only be obtained by comparison. Figure 3 shows the influence of decision tree depth on error in the current dataset.

Obviously, when the depth of the tree is 5, the minimum value is obtained. When integrating algorithms, we choose the tree depth of 4, 5 and 6 for learning, and finally select the best one.

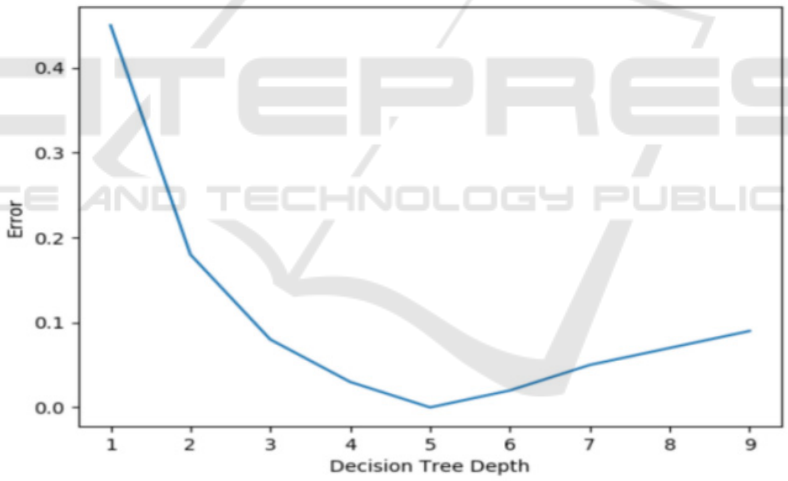

Figure 3: Influence of decision tree depth on error.

3.2 Integrated Algorithm Bagging

Bagging is a kind of integration algorithm, also called bootstrap integration algorithm. It starts with a base learner. There can be many algorithms as base learners, and decision trees are the most commonly used learners in Bagging. The advantages of the integrated algorithm over a single decision tree are that it can make the prediction curve of a single decision tree algorithm more smooth and closer to the real value.

We use 2/3 of the data set as the training set, and the remaining 1/3 as the test set. When the tree depth is 4, 5, and 6 respectively, we use Bagging for integrated learning, cross validation for 4 times, and finally select the best model as the final result. Figure 4 shows the ROC curve predicted by stroke patients with different model numbers when the tree depth is 6 (when the tree depth is 4, 5 and 6, the effect is not significantly different, and 6 is taken at last). In Figure 4, when the number of models>7 (corresponding to the horizontal axis point 0.35), the difference of identification error is small.

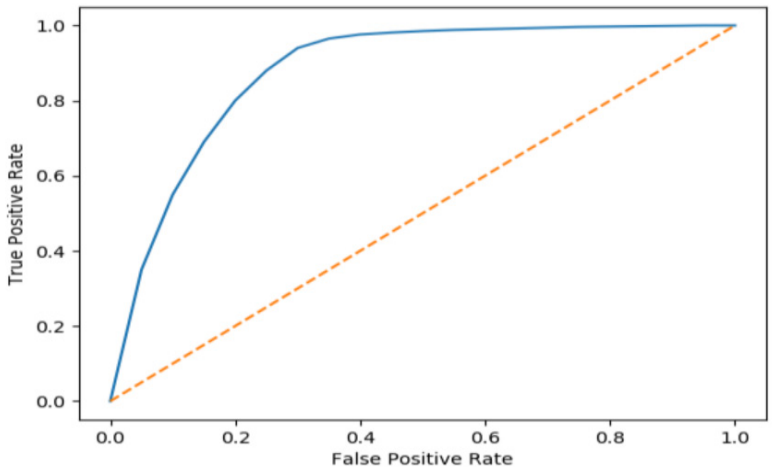

Figure 4: ROC curve.

4 CONCLUSION

From the perspective of the explosive factors of the new generation of AI technology, the success can be attributed to three key factors: powerful computing resources, complex neural networks, and large-scale data sets. Many practices also prove that sufficient data is more important to machine learning than the algorithm itself. Machine prediction is often based on a large amount of data. How to predict when the number of samples is insufficient is a direction worth discussing. When the data in a dataset can summarize the characteristics of all elements of a target, a model that meets the requirements can be obtained through small sample learning. In this paper, we use the confirmed cases of stroke to expand the screening data and build an effective stroke screening dataset. The expansion of the data set is to first classify the confirmed cases through the k-means algorithm, and then select the representative confirmed cases to add to the screening data set. After the data set is built, the prediction model can be built through machine learning algorithm. In this paper, instead of using various algorithms based on small sample learning, we use integrated algorithms to build a prediction model and choose the best model parameters. The test results show that the model also has a good effect on the test set.

The significance of this paper lies in how to effectively use relevant data sets for data expansion when the data samples are small and the positive and negative examples of the examples are incomplete. The data collection work of many public health systems has been carried out for a short time, and insufficient data is a common problem at present. Therefore, it is necessary for the application of artificial intelligence in the medical and health field to explore the use of machine learning for prediction modeling by expanding data sets. Next, we will study the optimization of prediction models to explore how to obtain prediction models with better performance and more dynamic scalability.

REFERENCES

GUO, Zhiheng., LIU, Qingping., LIU, Fang., WANG, Chengwu. & RUAN, Xuling (2021). Early Prediction Model of Stroke Based on Machine Learning Algorithm. Computer & Digital Engineering 2021, 49(11): 2180-2183.

PENG, Chen. Influencing factors and prediction of prognosis of patients with is chemic stroke based on machine learning [D]. Nanchang: Nanchang University: 2021.

WANG, Fang. ZHANG, Xueying. HU, Fengyun & LI, Fenglian. Ensemble Method Classifies EEG from Stroke Patients [J]. Computer Engineering and Applications, 2021, 57(24),276-282

WU, Maolin. The study of prediction model of ischemic stroke recurrence based on machine learning [D]. Shanxi: Shanxi Medical University:2022.

WU X.ZHU B. FU L. Prevalence, incidience, and mortality of stroke in the Chinese island populations: a systematic review[J]. PLoS One,2013,8(11):67-69.