Stroke prediction model based on decision tree

YUHENG LIU, CHENXUAN ZHANG, XIAOYANG ZHENG, YUHAN LIU, JIANGPING HE
School of Artificial Intelligence, Liangjiang
Chongqing University of Technology
Chongqing, 401135, P.R.CHINA

Abstract: In this paper, the predictive model of stroke based on decision tree is implemented to predict the stroke probability of ten samples by using Python language. The dataset of stroke is collected and is preprocessed, then the Gini coefficients of each feature are calculated to select the division, and then the decision tree model is obtained. Finally, the stroke probability is predicted for ten samples. In addition, Naive Bayes model is applied to predict the stroke probability to compare with the decision tree method. The experimental results show that older people with high blood pressure, heart disease, habitual smoking are more possible to have stroke, with a prediction accuracy of 88% for decision tree method and 79% for Naive Bayes model, respectively.

Key-Words: Stroke prediction; Decision tree model; Naive Bayes model

Received: April 15, 2022. Revised: January 2, 2023. Accepted: February 3, 2023. Published: March 7, 2023.

1 Introduction

With the development and progress of society, people's requirements for physical health are getting higher and higher [1]. Stroke is an acute cerebrovascular disease and is a group of diseases that cause brain tissue damage due to the sudden rupture of blood vessels in the brain or the inability of blood to flow into the brain due to blood vessel blockage, which poses a great threat to people's health [2]. Therefore, it is very important to understand the connection between people's physical condition and the probability of incidence and take different precautions for different groups of people. In medical diagnosis, time series disease prediction of irreversible diseases is very important, and prediction of future disease development can help patients intervene in advance, which has great significance for the effective control of diseases. Because of this, machine learning algorithms are widely used in the field of medical forecasting. In this paper, the computational prediction of stroke probability using decision tree models is obtained by the Python language extension package.

2 Problem Formulation

2.1 Decision tree based on CART

The CART (Classification and Regression Tree) algorithm is done in two parts, namely the generation and pruning of the decision tree. We use the minimum Gini index to choose the best features for constructing a binary tree. The steps for constructing a CART decision tree are as follows [2]:

1) After calculating the Gini index for all the labels, the largest tag of the Gini index is selected as the separation feature for branching.
2) All features in this label are calculated by the Gini index, and the feature with the largest index is also selected as the segmentation node, and the above process is repeated until the Gini index reaches the optimal, or the branching stops when the threshold is reached.
3) Complete the construction of the decision tree.

For the classification problem, suppose that there is a K class, and the probability that the sample points belong to the Kth class is p_k, then the Gini index of

the probability distribution is defined as:

$$Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 \quad (1)$$

The principle of pruning is usually to minimize the loss function of the decision tree as a whole. Without restriction, decision trees tend to grow until the measure index is optimal or no residual features are available, then prone to overfitting. To this end, we prune by limiting the growth depth of the tree and the minimum number of samples of the current node before branching, ensuring the generalization of the model while trying to avoid overfitting[4][5].

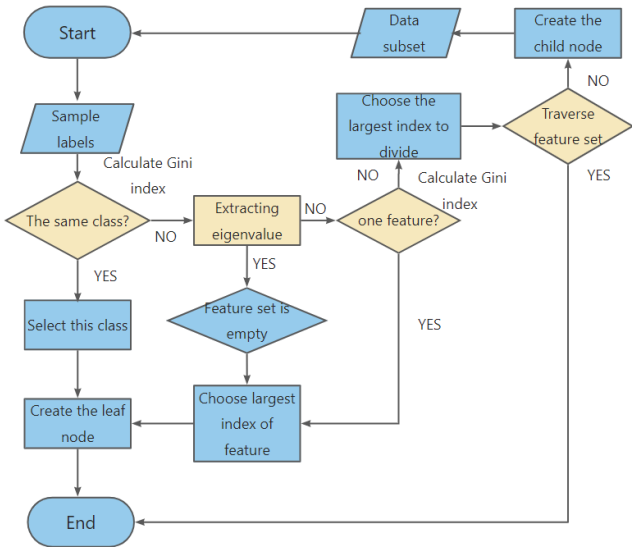

Fig.1 Decision tree flow chart

2.2 Naïve bayes method classification

The basic concept of the naïve Bayes method is a probability-based classification method that assumes independence from the dependent variable and is also a conditional model based on the Bayes theorem. Here's the classification process 0:

1) Calculate the prior probability, which is the proportion of each species as:

$$P(Y = c_k) = \frac{\sum_{i=1}^{N}(y_i = c_k)}{N}, k = 1,2,3, \ldots . K \quad (2)$$

2) Calculating the conditional probability, which is the conditional probability for each attribute in the training dataset:

$$P\left(X^{(j)} = a_{il} \middle| Y = c_k\right) = \frac{\sum_{i=1}^{N} I(x_i^{(j)} = a_{ji}, y_i = c_k)}{\sum_{i=1}^{N} I(y_i = c_k)} \quad (3)$$

$$j = 1,2,3,..,n, \quad l = 1,2,3 \ldots ., s_j, \quad k = 1,2,3, \ldots K$$

3) For the given sample $x_i = (x^{(1)}, x^{(2)}, \ldots x^{(i)})^T$, calculate the posterior probability:

$$P(Y = c_i) \prod_{j=1}^{n} p\left(X^{(j)} = x^{(j)} \middle| Y = c_k\right) \quad (4)$$

$$k = 1,2,3 \ldots K$$

4) The maximum posterior probability is determined, and the class of instance x is determined based on the value of the maximum posterior probability:

$$y = (Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k) \quad (5)$$

$$argmaxP = y$$

3 Problem Solution

3.1 Data preprocessing

We collected 5110 people's information containing age, BMI and a total of 10 other features as raw data. The raw data also include stroke or not, each column represents a factor and each row represents a sample. Remove the vacant and erroneous values, fill in the vacant values, convert the text data to numbers, and retain the rest of the data, thus converting the original data into a matrix of numbers, where stroke is represented by 1 and stroke is represented by 0. Before using a predictive model, we first divide the data matrix and divide 70% of the dataset into training sets and 30% into test sets based on experience to evaluate the accuracy of the model after training.

3.2 Predict model setting

When using the prediction model, because the classification goal of this dataset is not balanced, that is, the number of strokes accounts for a very small minority, we use the Smote algorithm[7] to oversample, artificially increase the number of strokes to make the data more balanced, and avoid the overfitting problem of decision trees and naive Bayes.

In the decision tree model, we set the maximum growth depth to 25 according to the maximum growth curve, and the minimum number of leaf node samples 3, which had a high degree of confidence.

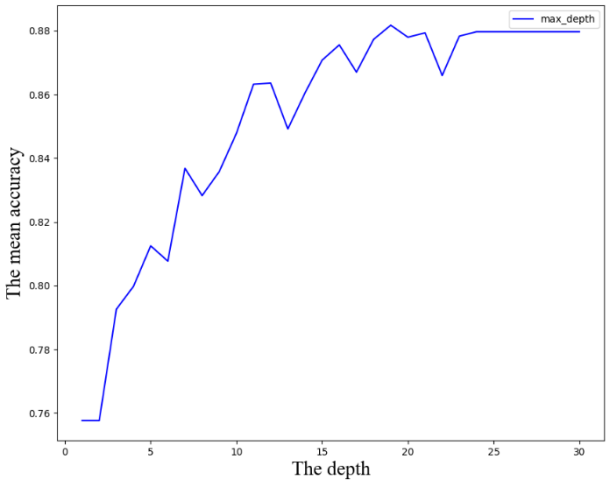

Fig.2 Decision tree growth curve

3.3 Predicting result analysis

Stroke prediction was performed on another 10 independent samples using the above two prediction models, the probability of the stroke are as shown in Table 1 and Table 2:

Table 1. Decision tree predicting result

	0	1	2	3	4
No stroke	0.8942	0.8963	0.8055	0.9716	0.7857
Stroke	0.1058	0.1037	0.1945	0.0284	0.2143
	5	6	7	8	9
No stroke	0..9951	0.8962	0.9951	0.9716	0.8942
Stroke	0.0049	0.1038	0.0049	0.0284	0.1058

Table 2. Naïve bayes predicting result

	0	1	2	3	4
No stroke	0.0967	0.9999	0.0699	0.9999	0.9976
Stroke	0.9033	0.0001	0.9301	0.0001	0.0023
	5	6	7	8	9
No stroke	0..9476	0.0418	0.9999	0.9959	0.5811
Stroke	0.05524	0.9582	0.0001	0.0041	0.4189

According to the Tables 1 and 2, and comparing the features of the 10 samples themselves, we found that people who were older, suffered from underlying diseases such as heart disease or hypertension, and had a greater probability of having a stroke, and had a lower correlation with their place of residence, whether they had a history of marriage and childbearing, and the type of work. By consulting the relevant medical literature[8], it is known that the population with the above characteristics does have a high probability of stroke, which can indicate that our model is reliable.

4 Conclusion

Based on the training of data on ten factors such as age, whether there is an underlying disease, and health status, this paper obtains a stroke prediction model, which can provide better medical evaluation for patients and provide diagnostic reference for doctors. According to this model, we can comprehensively consider many factors to predict stroke in order to achieve the purpose of early detection and early intervention. In addition, based on the physical data provided by the patient, doctors can evaluate based on this more reference model, which helps to discover new information, facilitate decision-making, prevent early, and develop more reasonable treatment intervention strategies.

References:

[1] McLaren, L., Braitstein, P., Buckeridge, D. *et al.* Correction to: Why public health matters today and tomorrow: the role of applied public health research. *Can J Public Health* **111**, 812–813 (2020). https://doi.org/10.17269/s41997-020-00398-z

[2] Santamaría A , Oliver A , Borrell M , et al. Higher risk of ischaemic stroke associated with factor XI levels in dyslipidaemic patients. *Int J Clin Pract.* 2007; **61**: 1819-1823

[3] Research on Heartbeat Classification Algorithm Based on CART Decision Tree, 2019 8th International Symposium on Next Generation Electronics (ISNE), 2019, pp. 1-3, doi: 10.1109/ISNE.2019.8896650.

[4] S. Shah and P. S. Sastry, "New algorithms for learning and pruning oblique decision trees," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 29, no. 4, pp. 494-505, Nov. 1999, doi: 10.1109/5326.798764.

[5] Thompson, D., Murray, G. & Whiteley, W. Prediction of recurrent stroke and myocardial infarction after stroke: a systematic review of clinical prediction models. *Trials* **14** (Suppl 1), O76 (2013). https://doi.org/10.1186/1745-6215-14-S1-O76

[6] The Abstract of Thesis Classifier by Using Naive Bayes Method, 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), 2021, pp. 312-315, doi: 10.1109/ICSECS52883.2021.00063.

[7] K. Cheng, C. Zhang, H. Yu, X. Yang, H. Zou and S. Gao, "Grouped SMOTE With Noise Filtering Mechanism for Classifying Imbalanced Data," in *IEEE Access*, vol. 7, pp. 170668-170681, 2019, doi: 10.1109/ACCESS.2019.2955086.

[8] (2008). Stroke. In: Encyclopedia of Genetics, Genomics, Proteomics and Informatics. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6754-9_16259

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.