Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 2, No. 3, 2024

Fusion Logistic Regression: Advancing Towards Stroke Early Warning

Hongxi Liu *, Shichang Sun, Yiying Liu

College of Computer Science and Engineering, Dalian Minzu University, Dalian, China

* **Corresponding author**: Hongxi Liu (Email: liuhongxi123@163.com)

Abstract: In this study, we focus on developing a logistic regression-based binary classification model to identify potential hemorrhagic stroke patients accurately. The model utilizes personal history, medical records, onset and treatment information from a training set of 100 hemorrhagic stroke patients. The objective is to predict the extent of hematoma expansion, assessing whether patients face significant health risks. Compared to traditional methods for identifying hemorrhagic strokes, such as decision trees, support vector machines, random forests, and gradient boosting machines, our logistic regression model demonstrates significant advantages in performance metrics such as F1 score and recall. Moreover, it achieves an accuracy rate of 96% in testing, surpassing other comparative models. Additionally, the model provides each patient with precise disease probability predictions, aiding in early treatment, alleviating economic burdens, and substantially reducing the technical complexity for medical professionals during the diagnosis and treatment process.

Keywords: Logistic Regression, Hematoma Expansion, Hemorrhagic Stroke

1. Introduction

Stroke, also known as a cerebrovascular accident, is regarded as a highly challenging and difficult-to-treat global health issue due to its continuously rising incidence, potential for debilitating functional impairments, and significant fatality [1].

Hemorrhagic stroke, a subtype of stroke, typically results from the rupture of blood vessels within the brain, leading to the entry of blood into brain tissues without external trauma. Approximately 10% to 15% of all stroke cases fall into this category. Various factors, including cerebrovascular abnormalities and ruptured intracranial aneurysms, contribute to the occurrence of this condition [2]. Hemorrhagic stroke is an acute cerebrovascular disease with a mortality rate reaching up to 50% during the acute phase. Even for those who survive, the majority are left with neurological sequelae, imposing a substantial burden on both their health and family finances.

In the field of medical diagnostics, especially in the task of identifying hematoma expansion in stroke patients, the selection of an appropriate classification model is crucial. Traditionally, methods such as decision trees, support vector machines, and random forests have been commonly employed, each with limitations. For instance, decision trees are prone to overfitting on small datasets and are sensitive to minor fluctuations in data, resulting in considerable prediction variability. Support vector machines are complex regarding parameter and kernel function selection, lack adaptability to small datasets, and exhibit lower model interpretability than other methods. Although random forests are robust against overfitting, their performance on small datasets is suboptimal, and the model complexity diminishes interpretability.

In contrast, logistic regression stands out for its simplicity and robust interpretability, particularly well-suited for scenarios with limited training data, as in this study. Logistic regression efficiently handles small datasets and mitigates overfitting issues. Crucially, it provides clear probability scores, making the prediction results easy to understand and interpret,which is especially important for medical diagnostic scenarios. Physicians and researchers can better comprehend the model's predictions through these probability scores, enabling more accurate assessments and decisions.

2. Dataset Processing

2.1. Dataset

This dataset involving detailed information on 160 patients with hemorrhagic stroke. The core content of the dataset includes each patient's ID, age, gender, pre-bleeding mRS score, as well as medical histories such as hypertension, stroke, diabetes, atrial fibrillation, and coronary heart disease. Additionally, the dataset includes information on the overall volume of hemorrhage and its proportions in different brain regions, such as the anterior, middle, and posterior arteries of the left and right hemispheres, left and right brainstem/medulla oblongata, as well as the left and right cerebellum.

2.2. Dataset Processing

Handling Missing Values: In the analysis of 23 features for the first 100 patients (sub001-sub100), we utilized the Pandas library. During this process, it was observed that some patients had missing information regarding hematoma volume, location, serial number, and time. These omissions were not due to data errors, but occurred when patient information was not applicable in certain circumstances. Since these missing values do not impact the prediction of whether a hematoma occurred within 48 hours, we have decided not to address these missing data.

Handling Outliers: An error was identified in the 'Admission First Imaging Examination Serial Number' record for patient sub074 during data processing. The file displayed it as 20180719000630, but it should be 20180719000020 upon verification. This outlier pertains to the 'Follow-up 1 Serial Number' and has been corrected to ensure data accuracy and model robustness.

3. Construction of Logistic Regression Model

Logistic regression is a widely employed statistical method in machine learning and data analysis, particularly adept at handling binary classification problems[3]. It provides probability estimates regarding classification decisions, yields interpretable model outputs, and ensures effectiveness on small to medium-sized datasets. While logistic regression assumes a linear relationship between input features and output, limiting its application in dealing with complex or nonlinear relationships, regularization techniques enable effective risk reduction of overfitting. The practical utility and reliability of logistic regression as a classification tool have been demonstrated in various real-world applications.

Logistic regression models are known for their simplicity and high computational efficiency. As shown in Figure 1, they establish the relationship between input features and target output using a linear equation. The model transforms the linear output into probability values through the Sigmoid function, applying a probability threshold for the final classification. Logistic regression also incorporates a logarithmic loss function to measure the disparity between model predictions and actual data. Optimization algorithms such as gradient descent are utilized to find the optimal parameters. Additionally, regularization techniques enhance the model's generalization ability and prevent overfitting.

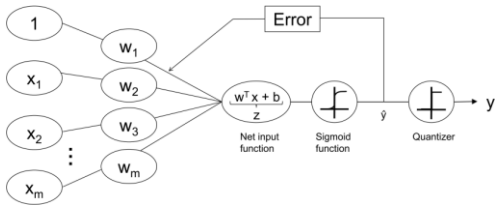

Figure 1. Logistic regression model diagram

This study represented the factors influencing stroke patients as features (x0, x1, x2,...,xn), and the score s of the patient's disease was obtained by using the matrix w_i for linear transformation:

$$s = \sum_{i=0}^{d} w_i x_i \quad (1)$$

Where a larger value of s indicates a higher likelihood of having a hemorrhagic stroke. To better control the range ofsand prevent it from becoming too large or too small, we used an activation function to compress its values within the range [0, 1]. This function is the logistic function and can be expressed as:

$$\theta(s) = \frac{1}{1 + e^{-s}} \quad (2)$$

So, the final form of the logistic regression Sigmoid activation function is represented as:

$$h(x) = \frac{1}{1 + e^{-w^T x}} \quad (3)$$

To improve model performance, it is expected to use a loss function for optimization, as shown below:

$$L(w) = -\frac{1}{m}\sum_{i=1}^{m} [y^{(i)} log(h(x^{(i)})) + (1 - y^{(i)}) log(1 - h(x^{(i)}))] \quad (4)$$

Where $L(w)$ is the loss function, $x^{(i)}$ represents a sample, and $h(x^{(i)})$ denotes the predicted probability. $y^{(i)}$ represents the true class label, taking two values, 0 or 1. This loss function assesses the difference between the model's predictions and the true labels. The loss is lower when the model's prediction is more consistent with the true label and vice versa. Generally, the objective in training a logistic regression model is to minimize this loss function. This is achieved through gradient descent, which automatically learns the weights w for more accurate classification performance.

4. Model Training, Resolution, and Analysis

Using the logistic regression model, we implemented a well-defined data partitioning strategy to estimate the risk of hematoma occurrence more accurately. In essence, we designated samples from sub001 to sub100 as the training set, primarily aiding in the training and parameter tuning of the model. Simultaneously, samples from sub101 to sub160 were set as the test set, allowing us to assess the model's predictive performance on new, unseen data. This partitioning approach ensures the robustness of the model on the known dataset while guaranteeing its adaptability to new data.

In this experiment, we relied on open-source machine learning frameworks within the Python environment for in-depth research. This integrated toolset enabled us to reduce the complexity of code writing, allowing more focus on fine-tuning parameters to ensure the model's results are accurate and stable. The hyperparameters of the logistic regression model are as follows:

Table 1.LogisticRegression() Model hyperparameter information

No.	Parameter	Default value	Parameter Description
1	penalty	12	Regularization type
2	tol	0.0001	Tolerance for stopping
3	C	1.0	Regularization strength
4	class_weight	None	Class weights
5	random_state	None	Random seed
6	solver	lbfgs	Optimization algorithm
7	max_iter	100	Maximum iterations
8	multi_class	auto	Multi_class strategy
9	warm_start	False	Warm start
10	n_jobs	None	Number of CPU cores

The regularization parameters, such as penalty and C, are primarily used to balance the stability and complexity of the model, preventing overfitting and enhancing its predictive ability on new data. Additionally, choosing the correct solver

is crucial and often depends on the characteristics of the data. Meanwhile, the random_state parameter ensures result consistency, and n_jobs allows us to adjust the usage of CPU cores, thereby influencing the model's computational speed. The judicious selection of these parameters helps us better establish and optimize the logistic regression model to identify patients with hematomas better.

In order to analyze the predictive performance of the model from different perspectives, we visualize the prediction results through graphics, as shown in the following figure:

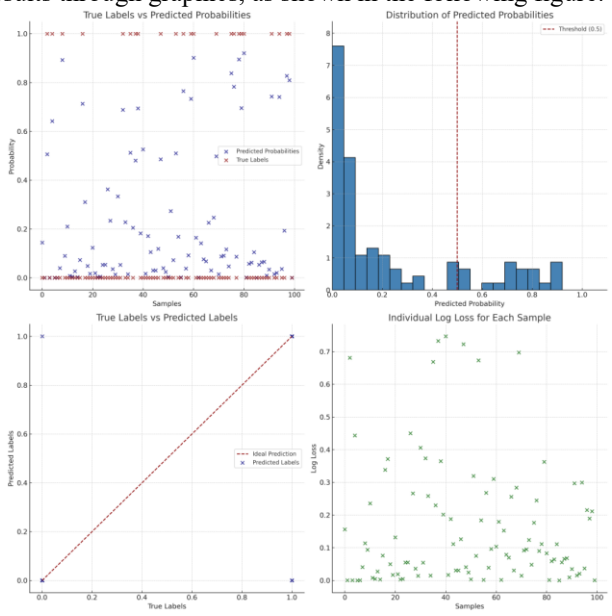

Figure 2. Visualization of classification model results

The figure (top left part) compares true labels and predicted probabilities. The red color represents label values of 0 or 1, while the blue markers indicate predicted classification probabilities. A distribution of blue markers closer to 0 or 1 signifies better model performance. The top right graph depicts the distribution of predicted probabilities, with the red line indicating the threshold used to determine the probability category of the predictions. The bottom right graph displays the Log loss for each sample, where lower loss values correspond to better model predictive performance.

To delve deeper into and quantify the predictive performance of the hematoma probability model, we employed two essential visualization tools: the confusion matrix and the ROC curve, as shown in the figure below. The confusion matrix provides an intuitive framework, showcasing the comparison between the predictions and actual occurrences of hematomas for 100 patients. Specifically, 76 and 20 indicate cases accurately predicted by the model, while 1 and 3 highlight instances where the model failed to identify occurrences correctly. The model achieves approximately 96% prediction accuracy, indicating its proficiency in handling the classification task.

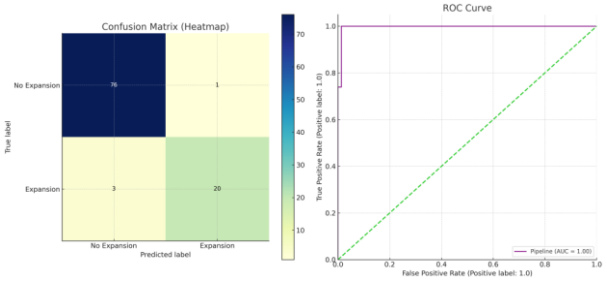

Figure 3. ROC curve and confusion matrix diagram

At the same time, the ROC curve further reveals the nuanced performance of the model. It rapidly trends towards the upper-left corner from its initial position in the graph, indicating a large AUC value. This further confirms the outstanding classification performance of the hematoma probability logistic regression model. The AUC provides a comprehensive metric to evaluate the model's discriminative ability between positive and negative classes. Through these two analysis methods, we can systematically comprehend the overall performance of the hematoma probability classification model and identify directions for subsequent enhancements and improvements.

5. Comparative Experiment and Model Optimization

5.1. Comparative Experiment

In order to validate the practicality and excellence of our model, we decided to adopt a comprehensive approach in this study by comparing it with several mainstream machine learning algorithms. This comparison aims to evaluate our model's uniqueness and potential limitations from multiple perspectives. Maintaining balance and fairness in the experiment is crucial when conducting such in-depth comparisons. Therefore, we implemented specific measures to ensure that each model operates in the same Python computing environment and employs uniform evaluation criteria to measure their respective performances. This approach helps eliminate any biases introduced by environmental variables or evaluation differences.

To comprehensively assess the performance of our model, we conducted a comparison with four commonly used models: Decision Trees, Support Vector Machines, Random Forests, and Gradient Boosting Machines. Each of these algorithms has its strengths and distinctive features. By contrasting our model with them, we aim to better understand our model's practical application potential and limitations. The experimental results for each model are visualized in Figure 4:

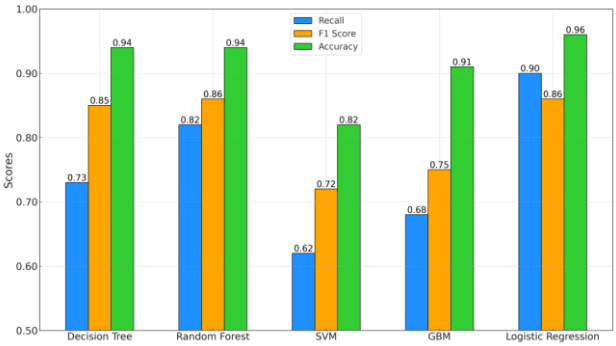

Figure 4. Model performance comparison

Clearly, logistic regression demonstrates superior accuracy and excels in recall and F1 score compared to other methods in predicting the probability of hematoma expansion. This conclusion highlights the unique advantages and robustness of logistic regression in addressing this classification task.

5.2. Model Optimization

By adjusting parameters to enhance model performance, we focused on fine-tuning the regularization coefficient 'C' in the logistic regression model. The figure below illustrates the results for different values of 'C':

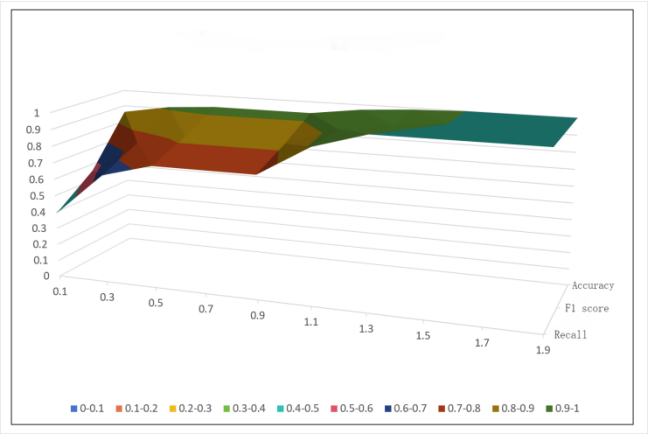

Figure 5.Influence of different regularization parameters C on model prediction optimization

In the figure, while keeping other conditions constant, we gradually varied the value of 'C' to find the optimal parameter configuration. Firstly, when 'C' is set to 0.1, the model's performance is relatively poor, with recall, F1 score, and accuracy being 0.39, 0.59, and 0.86, respectively. Next, as we incrementally increased the value of 'C', we observed a significant improvement in recall, F1 score, and accuracy when 'C' reached 0.5, with values of 0.73, 0.82, and 0.93, respectively. We continued fine-tuning the value of 'C', but found that the model's performance remained stable between 'C' values of 0.5 and 1.0, with slight variations in metrics. However, it is noteworthy that when 'C' reached 1.1, we observed a rapid increase in model performance indicators, which may suggest the onset of overfitting. As 'C' further increased, performance metrics continued to improve, but at 'C' = 1.5, the model performance seemed to have reached a perfect level, indicating potential overfitting on the training data. We suspect this phenomenon may be attributed to the relatively limited sample size and the high dimensionality of features in the dataset. Considering the risk of overfitting, we ultimately selected 'C' = 0.9 as the optimal model parameter setting.

6. Conclusion

This study aimed to identify potential patients with hemorrhagic stroke by constructing a binary classification model using logistic regression. The goal was to recognize all possible stroke patients early on, allowing for timely detection and avoiding delays in treatment. To ensure the accuracy of the model, we utilized gradient descent optimization. Simultaneously, we conducted L2 regularization analysis to prevent overfitting and fine-tuned the regularization coefficient for optimal performance. In the end, when the regularization coefficient was set to 0.9, the model exhibited the best performance and demonstrated robust resistance to overfitting. To emphasize the advantages and uniqueness of our model, we compared it with decision trees, support vector machines, random forests, and gradient boosting machines, evaluating using performance metrics such as F1 score and recall. The results showed that our model achieved a 96% accuracy, significantly outperforming the other models.

References

[1] Kisa A , Kisa S , Collaborators G S .Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019[J].The Lancet Neurology, 2021,26:795-820.

[2] Rui J M L ,Jaclyn T ,Yao Y A N , et al.Acceptance of disability in stroke: a systematic review[J].Annals of Physical and Rehabilitation Medicine,2024,67(2).

[3] Yicheng X ,Silong C ,Mengmeng Z , et al.The Prediction Models for High-Risk Population of Stroke Based on Logistic Regressive Analysis and Lightgbm Algorithm Separately.[J]. ranian journal of public health, 022,51(5):1999-1009.