Integrated Logistic Regression–XGBoost and Bayesian Network Models for Disease Prediction

Yihao Zhu[1,a], Ruixin Zhang[1,b], Junfeng Li[1,c]

[1]School of Mathematics and Physics, Nanyang Institute of Technology, Nanyang, China

[a]1798779567@qq.com, [b]1612691954@qq.com, [c]1836737813@qq.com

Keywords: K-S test; Spline function; Logistic regression - XGBoost integration; Bayesian network; Disease prediction; Comorbidity analysis

Abstract: Cardiovascular disease, stroke and cirrhosis are diseases that pose a major health threat worldwide. This study carries out data-driven disease risk prediction and association analysis based on three disease datasets. In the first step, the stroke, heart disease and cirrhosis datasets were systematically preprocessed, including outlier processing based on the K-S test, spline function interpolation of missing values, and standardization and visual analysis. Through correlation analysis and chi-square test, the key influencing factors such as age, ST-segment depression, and albumin were identified. In the second step, a logistic regression-XGBoost integrated model was constructed to predict the prevalence probability of three types of diseases, and the model performance was evaluated through accuracy, AUC-ROC and other indicators, among which the accuracy of logistic regression for heart disease prediction reached 68%, and XGBoost's performance in the multi-classification task of liver cirrhosis needs to be improved. The results showed that the probability of heart disease-cirrhosis complications reached 83.33% in the high-risk group and 25.75% in the high-risk group, revealing a strong correlation between diseases. This study provides data support and method reference for multi-disease risk prediction and collaborative prevention and control.

1. Introduction

In the context of global public health governance, health damage caused by non-communicable diseases continues to rise, among which cardiovascular disease, stroke and cirrhosis constitute the three key diseases that threaten the quality of human life[1][2]. At present, cubes covering the above three diseases have gradually accumulated, providing basic support for analyzing the occurrence of diseases and mining influencing factors from the data level[3][4].

This study aims to explore the risk mechanism of the three types of diseases through big data analysis technology, construct an accurate prediction model, and reveal their comorbidity characteristics[5]. In this paper, systematic preprocessing and statistical analysis of the three disease datasets were carried out to identify the key influencing factors[6]. Then, a logistic regression-XGBoost integrated model is constructed for disease prediction, and the model performance is evaluated[7]. The Bayesian network model is further used to explore the probability and association pattern of concurrency between diseases[8]. Relevant research can provide a scientific basis for early disease early warning, risk stratification and comprehensive prevention and control, and has important theoretical and practical value[9].

In recent years, many scholars have conducted extensive research on disease prediction and association analysis[10]. For example, Chen and Guestrin proposed the XGBoost algorithm, which excels in structured data prediction[11][12]. Koller and Friedman system expounded the application of Bayesian network in probabilistic reasoning[13][14]. The Scikit-learn library developed by Pedregosa et al. provides efficient implementation for models such as logistic regression[15][16]. In addition, D'Agostino et al. established the Framingham risk scoring model in terms of cardiovascular disease risk prediction[17][18]. In the prognosis assessment of liver cirrhosis, the Model for End-Stage Liver Disease scoring system is widely used. These studies provide an

  DOI: 10.25236/iwmecs.2025.020

important methodological basis for this paper[19][20].

2. Model creation, solution and discussion

2.1. Model establishment

2.1.1. Data preprocessing and feature extraction

Firstly, the datasets of the three diseases were systematically cleaned and preprocessed. The K-S test method was used to detect the outliers, and the statistics were:

$$D = \max_x | F_n(x) - F(x) | \tag{1}$$

Among them, $F_n(x)$ is the empirical distribution function, and $F(x)$ is the theoretical distribution function.

Impute missing values using a spline function:

$$f(x) = \sum_{i=1}^{n} a_i \phi(\| x - x_i \|) + b \tag{2}$$

Perform Z-score normalization:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

The categorical variables are encoded with unique heat. Finally, the key influencing factors of various diseases were identified by Pearson correlation coefficient and chi-square test.

2.1.2. Logistic regression-XGBoost ensemble prediction model

Logistic regression maps linear combinations to disease probabilities through the Sigmoid function:

$$P(y = 1 | x) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{4}$$

Its parameters are solved by maximum likelihood estimation.

XGBoost is a gradient lifting tree ensemble model with a predicted output of:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{5}$$

The objective function includes the loss function and the regular term:

$$\text{Obj} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{6}$$

Optimize the tree structure through Taylor expansion and the greedy algorithm. Finally, integrate the results of logistic regression and XGBoost to enhance prediction robustness.

2.1.3. Bayesian network comorbidity probability model

Bayesian networks are a probabilistic graph model used to represent dependencies between variables. Based on the common characteristics of preprocessing (such as age, gender, blood pressure, blood glucose, etc.), the network structure is constructed, and the node relationship is quantified by the conditional probability table. For the probability of disease complication, the chain rule is used to calculate:

$$P(A, B, C) = P(A) \cdot P(B | A) \cdot P(C | A, B) \tag{7}$$

Among them, A, B, C represents the disease status of stroke, heart disease, and liver cirrhosis

respectively. The posterior probability of the concurrent occurrence of any two or three diseases is calculated by inference algorithms.

2.2. Model Solution and Results

2.2.1. Step 1 model solution results

Table 1Three key disease factors affect the results and chi-square values

Disease	Key numerical variables (correlation coefficients)	Key categorical variables (significant chi-square test)
Stroke	Age (0.2453), History of heart disease (0.1349)	Marital status, type of job, smoking status
Heart disease	ST segment depression (0.4081), maximum heart rate (-0.4004)	ST-segment slope, type of chest pain, exercise angina
Cirrhosis of the liver	Days of disease duration (-0.36), albumin (-0.30)	Ascites, hepatomegaly, spider nevus, edema

The results of Table 1 show that there are significant differences in the key influencing factors of different types of diseases. For stroke, age (r=0.245) and history of heart disease (r=0.135) were the main numerical risk factors, and their occurrence was significantly correlated with social behavioral factors such as marital status, job type, and smoking status. The identification of heart disease is highly dependent on ECG indicators, in which ST-segment depression shows a strong positive correlation (r=0.408), while maximal heart rate shows a strong negative correlation (r=-0.400). In addition, classification features such as ST-segment slope, chest pain type, and exercise angina are key clinical discriminants. In cirrhosis, the number of days (r=-0.36) and albumin level (r=-0.30) are important laboratory and course indicators, while signs such as ascites, hepatomegaly, spider nevus, and edema are highly significant clinical classification manifestations.

2.2.2. Step 2 model solution results

The performance of the logistic regression and XGBoost ensemble model on the test set is as follows:

Table 2 Disease prediction model performance

Disease	Model	Accuracy	AUC-ROC
Heart disease	Logistic Regression	0.68	0.69
Stroke	Logistic Regression	0.94	0.84
Cirrhosis of the liver	XGBoost	0.39	0.43–0.69

The specific results are shown in Table 2. In the prediction of heart disease, the logistic regression model achieved an accuracy of 0.68 and an AUC-ROC value of 0.69, indicating that the model has a certain discriminant ability, but its performance is still at a moderate level. In contrast, logistic regression performed well in the stroke prediction task, with an accuracy of 0.94 and an AUC-ROC of 0.84, indicating that the selected features (such as age, cardiac history, and social behavior variables) could better distinguish the risk of stroke, and the modeling task was relatively clear.

2.2.3. Step 3 model solution results

The Bayesian network model calculates the probability of disease comorbidity as follows:

Table 3 The probability of complications of the three diseases

Concurrency type	Probability of high-risk group (%)	Probability of medium-risk group (%)	Probability of low-risk group (%)
Stroke - heart disease	19.78	0.56	0
Stroke - cirrhosis	28.98	1.38	0
Heart disease - cirrhosis	83.33	48.55	32.26
Three diseases are complicated	25.75	0.72	0

According to the results of Table 3, the probability of heart disease-liver cirrhosis in the high-risk group was as high as 83.33%, which was significantly higher than that of other complication types, suggesting that the two diseases may have a strong common basis in terms of pathological mechanism, risk factors or lifestyle. The probability of stroke-cirrhosis is 28.98%, which is also at a high level. It is worth noting that even in the medium-risk group, the probability of heart disease-cirrhosis complications is still 48.55%, and the probability of complications in the low-risk group is 32.26%, indicating that this complication combination is more common in different risk groups and should be used as the focus of clinical monitoring and health management.

2.3. Results and discussion

Through systematic data analysis and modeling, the key influencing factors, predictive performance and comorbidity patterns of the three types of diseases were revealed. In terms of influencing factors, age, ST-segment depression, albumin and other indicators have strong predictive power, which is consistent with clinical cognition. In predictive modeling, logistic regression is stable and suitable as a baseline model. XGBoost needs to further optimize features and parameters in complex multi-classification tasks. In the comorbidity analysis, Bayesian networks clearly demonstrate the probabilistic dependence between diseases, especially the high complication probability of heart disease and liver cirrhosis, suggesting that joint screening and management need to be strengthened in clinical practice. Overall, the combination of the integrated model and the probability graph model provides an interpretable and scalable analysis framework for multi-disease risk prediction.

3. Conclusion

This study focuses on three major non-communicable diseases, namely cardiovascular disease, stroke and cirrhosis, from data preprocessing, key factor identification, disease prediction to comorbidity probability analysis. In the first step, the key influencing factors such as age, ST segment characteristics, and albumin were clarified through the K-S test, spline interpolation and statistical test, which provided an index basis for risk identification. In the second step, a logistic regression-XGBoost ensemble model is constructed, which achieves good results in the prediction of heart disease and stroke, and verifies the effectiveness of ensemble learning in disease prediction. In the multi-classification task of liver cirrhosis, the model performance indicates that more features and structural optimization need to be introduced. The third step is to construct a comorbidity probability model based on Bayesian network, which quantitatively reveals the strength of the association between diseases, especially the comorbidity probability of heart disease and liver cirrhosis in high-risk groups exceeds 80%, providing data support for joint prevention and control.

The main contributions of this study are: proposing a disease analysis process that combines traditional statistical methods and machine learning models; The systematic exploration from single disease prediction to multi-disease association has been realized. It provides data-based risk stratification and prevention and control recommendations for public health decision-making. However, there are still certain limitations in the research, such as a single data source, limited sample size, and uneven model performance in complex multi-classification scenarios. In the future, we can consider introducing more heterogeneous data sources, combining deep learning models to improve prediction accuracy, expanding the association analysis to more disease types, and promoting the pilot application of models in clinical practice.

Acknowledgements

Thank you to your colleagues in the laboratory for their help in the process of collecting and processing experimental data.

References

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[2] Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.

[3] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[4] D'Agostino, R. B., et al. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation, 117(6), 743–753.

[5] Kamath, P. S., et al. (2001). A model to predict survival in patients with end-stage liver disease. Hepatology, 33(2), 464–470.

[6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

[7] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. Annals of statistics, 28(2), 337–407.

[8] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

[9] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[10] Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

[12] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[13] Liu, Y., et al. (2020). A deep learning system for differential diagnosis of skin diseases. Nature medicine, 26(6), 900–908.

[14] Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

[15] Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), 18.

[16] Weng, S. F., et al. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS one, 12(4), e0174944.

[17] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920–1930.

[18] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine, 375(13), 1216.

[19] Goldstein, B. A., et al. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 24(1), 198–208.

[20] Shickel, B., et al. (2018). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), 1589–1604.