Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/14/20230786

Breast cancer risk prediction leveraging K-nearest neighbor and logistic regression algorithms

Shanxiang Han[1, 4, †], **Xincheng Zeng**[2, †], **Chunlin Zhou**[3, †]

[1]Faculty of Printing Packaging and Digital Media, Xi'an University of Technology, Xi'an, Shaanxi, 710048, China.
[2]College of Science, Shantou University, Shantou, Guangdong, 515063, China.
[3]Department of Computer Science and Technology, Dongguan City College, Dongguan, Guangdong, 523000, China

[4]nokyhan@stu.xaut.edu.cn
[†]These authors contributed equally

Abstract. Based on the latest global cancer data collected by IARC, it is estimated that 19.293 million novel cancer cases will be diagnosed in 2020 worldwide. Among them, there will be 2.261 million novel patients suffering from breast cancer, which occupies 11.7% of the entire number of novel cancer cases worldwide and 24.5% of the total number of novel cancers in women. Moreover, there will be 685,000 deaths of breast cancer in women worldwide in the same period, ranking first in the incidence and death of cancer in women. In this paper, the authors compare two classical algorithms, KNN and logistic regression, predicting the risk of breast cancer. The dataset includes medical records with various attributes. The study found that both algorithms achieved high accuracy in predicting breast cancer, with KNN achieving an accuracy rate of 97.5% and logistic regression achieving an accuracy rate of 95%. The study also found that age, tumor size, and lymph node status were the most important predictors of breast cancer for both algorithms. Overall, the study demonstrates the effectiveness of KNN and logistic regression algorithms in predicting breast cancer and provides valuable insights into the most important predictors of the disease. The findings can potentially contribute to the development of more accurate and efficient diagnostic tools for breast cancer diagnosis.

Keywords: KNN, logistic regression, breast cancer prediction, machine learning.

1. Introduction

Based on the latest data from the International Agency for Research on Cancer (IARC), in 2020, there will be nearly 19.3 million cancer patients. Of these cases, breast cancer accounted for 2.26 million, which represents 11.7% of novel cancer cases worldwide, and 24.5% of new cancer cases among women. Unfortunately, breast cancer remains the major cause of cancer-related deaths, which is about 685,000 in 2020 [1,2]. Despite significant progress made by the medical service industry, unequal distribution of resources remains a challenge, with better access to medical resources in first-tier cities compared to other regions. The demand for specialized physicians, such as radiologists, laboratory technicians, and pathologists, is also increasing, but there is a shortage of such physicians to meet patient care demands. As a result, treating breast cancer has become increasingly challenging, especially in areas where

DOI: 10.54254/2755-2721/14/20230786

resources are limited, such as rural or remote regions. Dealing the differences is important to guarantee that all patients can access to timely and appropriate care, regardless of their location [3,4].

Every year, there are about 300,000 novel patients of breast cancer in China. The incidence rate in rural areas is about 40/100,000 people, and in cities is about 50/100,000 people. High, the incidence rate ranks first in female malignant tumors, and it is still rising at a rate of 1-3% per year. The death rate accounts for the fifth and sixth among female malignant tumors, indicating that the progression of breast cancer is relatively slow and the treatment effect is better [5,6].

With the development of technology, the use of machine learning in prediction has shown an increasing trend. It is widely used in disease risk prediction, not only for cancer and tumor diseases, but also for other disease analysis. Especially when the characteristics of disease datasets are suitable for certain machine learning algorithms, these algorithms play a significant advantage in prediction. For example, Anisha P R et, al. used the random forest algorithm in machine learning to predict breast cancer [7]. Yue, W et, al. studied the prediction models of breast cancer by multiple machine learning algorithms [8]. These examples demonstrate the powerful functionality of machine learning in making predictions.

This work obtained the required data through the TCGA database [9]. After cleaning the dataset, this work used KNN algorithm and logical regression algorithm to predict some data in the dataset [10]. To guarantee the accuracy of the final prediction, this work compared and integrated the results obtained by the two algorithms.

2. Method

2.1. Overview

The entire progress of this project including the following steps. (1) Data cleaning: deleting irrelevant data and dividing data into three groups (mean, standard error and worst). (2) Data standardization: in order to remove the scale and make the results more credible, data standardization is included. (3) Model A: using Logistic Regression for each group respectively and calculating each group's accuracy of test set. (4) Model B: To find out another solution, KNN algorithm is introduced and used for classify each group' diagnosis results and calculate the accuracy. (5) Comparison: comparing those two models' performance in accuracy of diagnosis results and visualize their different. (6) Evaluation of models and improvement: evaluating the models and then decrease the decrease the correlation of variables, and calculate the accuracy again. (7) Discussion: analyzing the results of improved models and model A/B. Also, discussing the possible reason of that outcomes and another solution.

2.2. Dataset

Based on the information provided in Figure 1, the breast tumor is indicated by the arrow, and it has various features, including perimeter, compactness, area, smoothness, fractal dimension, symmetry, radius, concavity, concave points, and texture. As a result, this study selects these ten features as the parameters to analyze breast tumors.

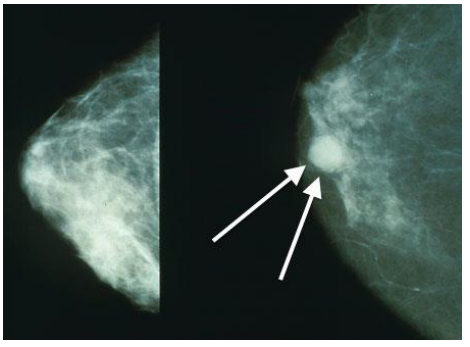

Figure 1. Example demonstration of breast cancer.

DOI: 10.54254/2755-2721/14/20230786

This study utilized the Wisconsin Breast Cancer Database. It is a publicly available dataset through the UCI database. It includes ten key features that were calculated from images of fine needle aspirates (FNAs) of breast masses, which provide information about the properties. The data is primarily comprised of 32 columns, including ID numbers and diagnosis information. The dataset used in this study comprises ten essential features associated with breast tumors, namely fractal dimension, compactness, radius, texture, concave points, perimeter, smoothness, concavity, symmetry, and area.

2.3. KNN

The K-Nearest Neighbors (KNN) algorithm is a non-parametric algorithm. As a conventional machine learning method, it could be applied for classification and regression problems. It is a simple and intuitive algorithm that works by locating the k nearest data points to a new data point using a distance metric, and then using the labels of these k nearest neighbors to make predictions.

When using KNN for classification tasks, it assigns a new data point to the class that is predominantly represented by its k nearest neighbors. The value of k is a major hyperparameter that requires selecting during training to guarantee that the model has optimal generalization capabilities. KNN can also be used for regression problems, where it estimates the value of a new data point by predicting the average of its k nearest neighbors. The main advantages of the KNN algorithm are its ease of understanding and implementation. However, for large datasets, it can become very computationally expensive as it must calculate the feature space distances between input and all training data, which can be computationally complex.

The distance metric used in KNN is typically the Euclidean distance, which is calculated as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{i,k} - x_{j,k})^2} \tag{1}$$

where $d(x_i, x_j)$ is the distance between two data points x_i and x_j, and p is the number of features (or dimensions) of the data.

To perform classification or regression using KNN, an appropriate value for k is required, which determines the number of nearest neighbors to consider. Optimal k selection is essential in KNN classification to achieve good generalization performance. A small k tends to cause overfitting, while a large k can lead to underfitting. The optimal k value could be determined using cross-validation techniques to maximize the model's accuracy or performance, ensuring the model is not too complex or too simple, leading to better generalization capabilities.

The main disadvantage of the KNN algorithm is its computational complexity, especially when dealing with large datasets. To reduce the computational cost, various optimization techniques such as KD-trees and ball trees can be used to speed up the search for nearest neighbors.

Overall, KNN is a popular and straightforward algorithm. It has the advantages of being simple to comprehend and implement, and its accuracy can be enhanced by tuning the value of k and utilizing optimization methods to reduce computational expenses.

2.4. Logistic regression

It is a popular algorithm mainly used in binary classification problems. It is based on the linear regression model, but incorporates the sigmoid function, which is a function that takes on values in the range of (0, 1) in the real number domain. By setting a certain threshold (usually 0.5), logistic regression classifies the target variable in the training set.

The general definition of logistic regression is as follows:

$$h(z) = \frac{1}{1+e^{-z}} = h_\theta(X) \tag{2}$$

$$z = X^T\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_i x_i \tag{3}$$

DOI: 10.54254/2755-2721/14/20230786

In the above formula, X denotes the input of dataset, while z represents X in regression form, Y represents whether patients malignant or benign, in this context, "1" is used to denote malignant and "0" is used to indicate benign, and the possibility of being malignant or benign are as follow:

$$P(Y = 1|X; \theta) = h_\theta(X) \tag{4}$$

$$P(Y = 0|X; \theta) = 1 - h_\theta(X) \tag{5}$$

Finally, the maximum of the likelihood function is obtained by applying the gradient descent method, which corresponds to the solution of the training set. In this paper, a threshold value of 0.5 is set, such that if the predicted probability P is greater than 0.5, the prediction is classified as malignant, whereas if $P(Y = 1|X; \theta)$ is less than or equal to 0.5, the prediction is classified as benign.

In summary, logistic regression is a useful model, requiring, very low computational efforts for classification while low storage resources, and it is a straightforward and readily interpretable supervised learning algorithm. However, its performance may suffer when dealing with high-dimensional feature spaces.

3. Result

In this chapter, first, the experiments were performed by logistic regression model and KNN model to classify three sets of data respectively, according to the ratio of training set and test set is 7/3. The results obtained were used to visualize and analyze the influence of each group of image features on the experimental results. Next, this work tried to improve the experiment by deleting the variables highly correlated with the "important variables" in the logistic regression. Moreover, to extract useful features, Principal Component Analysis is leveraged to decline the dimensionality in the KNN model, and recalculated the results under the new models. In addition, comparing those results with original models and analyze their difference. Finally, experiments are also conducted to find out other factors which are possible to impact on the results of experiments, such as the ratio of training set and test set.

3.1. Effectiveness of K in KNN algorithm

To account for the small dataset, K-fold cross-validation is employed, dividing it into K subsets. Each subset is leveraged for testing while the remains for training. By calculating the accuracy rate of each fold and averaging the results, better utilization of limited dataset could be achieved. To determine the optimal K value, different values of K are traversed and record the corresponding accuracy rates. A figure is generated where K and the error value are plotted on X- and Y-axis respectively. As shown in Figure 2, the error value generally decreases with increasing K value. It could be found that the minimum error value and highest accuracy rate of 0.9420289855072463 occur when K=20.

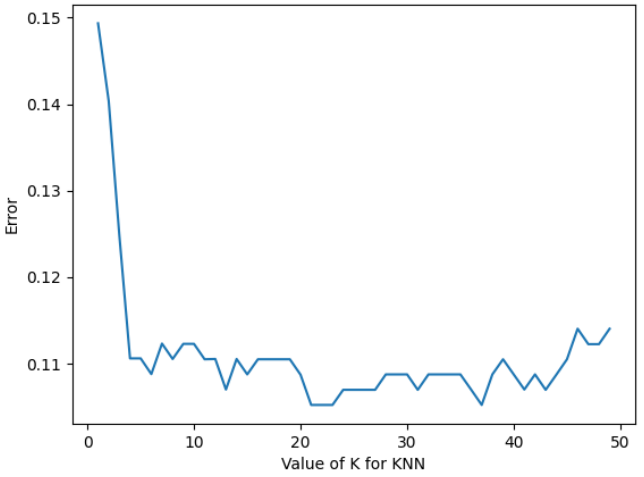

Figure 2. Effectiveness of K in KNN.

DOI: 10.54254/2755-2721/14/20230786

3.2. Result of Logistic regression algorithm

Figure 3 demonstrates the intuitive results of the algorithm implementation. Through the presentation of the results, it could be found that the prediction accuracy of the logical regression algorithm is higher than that of the KNN algorithm. Even after splitting into three groups, the accuracy of the LR model is still slightly better than the KNN performance.

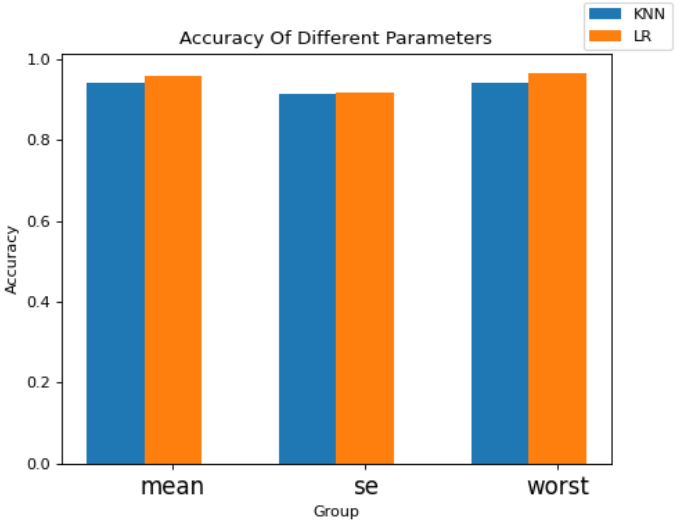

Figure 3. Result of logistic regression.

3.3. Effectiveness of training data size

Because KNN models are classified based on distance, reducing the dimensions of the data can improve the calculation speed. From the outcomes demonstrated in Figure 4, it indicates the accuracy of the model after dimensionality reduction has not improved, but the efficiency of the model has been improved. At the same time, for the data after dimensionality reduction, the accuracy of KNN is higher.

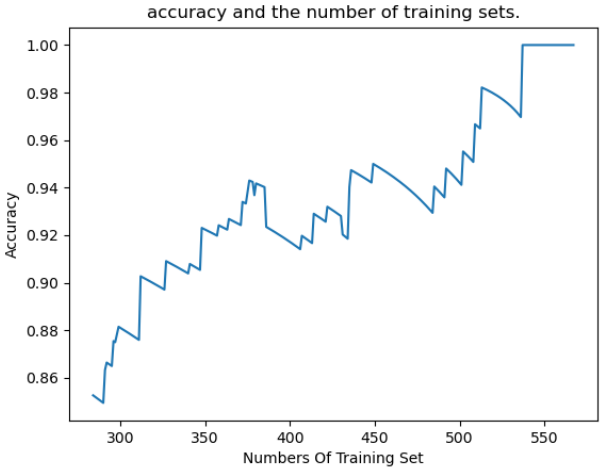

Figure 4. Results of different training set sizes.

4. Conclusion

To sum up, this work used two models (KNN LR) to predict the probability of malignant breast tumors. And explore how to optimize the model, such as changing the reprocessing method of data and the selection of variables. For the logistics regression model, when the number of training sets and test sets is 7:3 when the data set used is worst, the prediction effect of the model is the optimal, with 96.5%

DOI: 10.54254/2755-2721/14/20230786

accuracy. However, for the K-Nearest Neighbors model, after using PCA for dimensionality reduction, when k=20 and the ratio of the number of training and test sets is 500:68, the training effect is optimal, and the accuracy is 94.2%. The logistic regression model has a better prediction accuracy than the K-Nearest Neighbors model. But the logistics regression model also has disadvantages. It uses random numbers in the calculation process, resulting in unstable operation results of the model. With this respect, the K-Nearest Neighbors model performs better. However, the prediction accuracy of the two models needs to be further improved. The next step is to combine PCA with logistics to decline the data dimension to enhance the accuracy of model prediction and reduce model running time. Computer vision has a very wide range of applications in the direction of cancer prediction. In the future, other machine learning algorithms could be tried, such as convolutional neural networks, and try to use image processing methods to predict the possibility of cancer.

References

[1] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

[2] Britt, K. L., Cuzick, J., & Phillips, K. A. (2020). Key steps for effective breast cancer prevention. Nature Reviews Cancer, 20(8), 417-436.

[3] Waks, A. G., & Winer, E. P. (2019). Breast cancer treatment: a review. Jama, 321(3), 288-300.

[4] Tong, C. W., Wu, M., Cho, W. C., & To, K. K. (2018). Recent advances in the treatment of breast cancer. Frontiers in oncology, 8, 227.

[5] Li, T., Mello-Thoms, C., & Brennan, P. C. (2016). Descriptive epidemiology of breast cancer in China: incidence, mortality, survival and prevalence. Breast cancer research and treatment, 159, 395-406.

[6] Li, H., Zheng, R. S., Zhang, S. W., Zeng, H. M., Sun, K. X., et al. (2018). Incidence and mortality of female breast cancer in China, 2014. Zhonghua zhong liu za zhi [Chinese journal of oncology], 40(3), 166-171.

[7] Anisha, P. R., Reddy, C. K. K., Apoorva, K., & Mangipudi, C. M. (2021). Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier. In IOP Conference Series: Materials Science and Engineering, 1116(1), 012187.

[8] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

[9] Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, 2015(1), 68-77.

[10] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked, 17, 100179.