Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/43/20230813

Breast cancer classification based on hybrid machine learning model

Zhe Lin

Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

3090534795@shu.edu.cn

Abstract. This study proposes a hybrid model that combines K-Means clustering and Random Forest classification as an approach for breast cancer classification. The objective is to exploit the advantages of unsupervised clustering and supervised classification techniques to enhance the accuracy and robustness of classification models. The dataset underwent preprocessing procedures encompassing the handling of missing values, feature normalization, and feature selection. Missing values were addressed through appropriate methods, and features were scaled and selected based on variance threshold or correlation analysis. Subsequently, K-Means clustering was applied to the preprocessed data to assign cluster labels to each sample. The study then proceeded to train a Random Forest classifier by incorporating both the cluster labels and the raw gene eigenvalues as mixed features. This integration of gene expression values and cluster labels provides supplementary information to the classifier, enabling the capture of more intricate patterns within the data. The Random Forest classifier was trained using optimized parameters determined through parameter tuning, including the number of trees, maximum depth, and minimum number of split samples. Extensive experiments and evaluations conducted in this study revealed that the hybrid model outperformed the standalone Random Forest classification. The incorporation of K-Means clustering facilitated the discovery of underlying data structures and patterns, ultimately enhancing the classifier's discriminatory ability. The hybrid model exhibited superior accuracy, precision, recall, and F1 scores, demonstrating its efficacy in accurately classifying breast cancer samples.

Keywords: Breast Cancer, K-Means Algorithm, Random Forest.

1. Introduction

Breast cancer is one of the most common malignant tumors among women and stands as a prominent cause of mortality among female populations [1]. Its profound impact on society is characterized by the substantial allocation of medical resources and expenses required for effective treatment and recuperation, placing considerable economic strain on patients and their families [2]. Therefore, early detection and early treatment of breast cancer are crucial [3]. Despite notable advancements in conventional screening and diagnostic techniques, their accuracy and efficiency remain constrained [4, 5]. In recent years, Artificial Intelligence (AI) has demonstrated tremendous potential in the healthcare field, providing novel direction to assist doctors in detecting and analyzing breast cancer.

AI leverages machine learning and deep learning algorithms to analyze and identify patterns in vast amounts of medical imaging data, facilitating precise detection and diagnosis of breast cancer by

DOI: 10.54254/2755-2721/43/20230813

medical practitioners. Google DeepMind is working with the UK's National Health Service (NHS) to develop an artificial intelligence system for early prediction of the risk of Acute Kidney Injury (AKI) [6]. The system analyzes patients' clinical data, including vital signs and laboratory results, to identify patients at risk of AKI in advance. This automated detection system improves screening accuracy, reduces the risk of missed and misdiagnosed diagnoses, and increases the likelihood of early disease detection.

In addition to breast cancer screening and diagnosis, AI holds promise for personalized treatment. IBM's Watson for Oncology is a medical decision support system that employs AI technology to provide doctors with personalized cancer treatment recommendations [7]. Based on a large amount of medical literature and patient data, the system assists doctors in analyzing cases, formulating treatment plans, and providing relevant clinical guidance. By analyzing a patient's genomic information and clinical data, AI can provide insights into treatment response and prognosis. This personalized approach helps optimize treatment plans, improve treatment outcomes and minimize unnecessary drug toxicity and side effects. Furthermore, artificial intelligence plays a vital role in breast cancer research. By analyzing large-scale breast cancer data, AI can help researchers discover new predictors, biomarkers, and potential therapeutic targets. These findings contribute to a deeper understanding of the pathophysiology of breast cancer and guide the development of new treatments and drugs.

Despite advances in AI research in breast cancer healthcare, several challenges and limitations remain. First, the development of artificial intelligence algorithms requires a large amount of high-quality data for training and verification. Furthermore, the interpretability and reliability of algorithms are key issues that are currently being addressed.

In this study, a breast cancer classification model using a combination of K-Means and Random Forest algorithms was implemented. The combination of K-Means clustering, and random forest classification is of great significance in breast cancer research. K-Means clustering can help preprocess data, select features, reduce dimensionality, and provide data analysis and interpretation, providing accurate input and preliminary labels for random forest classification models. But the application of the K-Means algorithm alone in breast cancer analysis may not yield sufficiently high accuracy owing to the inherent constraints of unsupervised learning. The random forest classification model uses the results of K-Means clustering to further classify and predict a single cluster on the basis of clustering, effectively improving the classification accuracy, improving the diagnostic accuracy of breast cancer and the accuracy of individualized treatment. This comprehensive approach can better understand the structure and characteristics of breast cancer data, provide more effective support, promote the progress of breast cancer research, and provide important guidance for clinical decision-making.

2. Method

2.1. Dataset Description and Preprocessing

In this study, the dataset is the Breast cancer gene expression called Curated Microarray Database (CuMiDa), sources from Kaggle [8]. The aim of CuMiDa is to offer homogeneous and state-of-the-art biological preprocessing of these datasets, together with numerous 3-fold cross validation benchmark results to propel machine learning studies focused on cancer research [9].

In the dataset, there are 151 samples, including 54,676 genes which contain a large number of redundant and noisy features, increasing computational complexity and affecting classification performance. In feature selection, this paper uses the F-classify function to evaluate and select features according to the scoring function, and returns the selected feature matrix X-selected. Standardization is one of the important steps in data preprocessing, which transforms feature data into a standard normal distribution with mean 0 and standard deviation 1 to eliminate dimensional differences between different features. In order to evaluate the accuracy of the K-Means algorithm, this paper converts the classification labels into numerical forms in advance. This study adopts 50% distribution of train split and test split.

DOI: 10.54254/2755-2721/43/20230813

2.2. Machine Learning Model

2.2.1. K-Means. K-Means is a commonly used unsupervised learning clustering algorithm, which is used to divide the samples in the data set into K different clusters [10]. The algorithm achieves clustering by minimizing the distance between the sample point and the center of the cluster to which it belongs, where the distance is usually measured by Euclidean distance or other similarity measures [11]. This study first uses the elbow rule to determine the range of K values, and then uses the algorithm to cluster.

2.2.2. Random Forest. Random Forest is an ensemble learning method based on an ensemble model constructed from decision trees. It builds multiple decision trees by randomly selecting subsets of features and samples [12], and leverages collective decision making for tasks such as classification, regression, and feature importance assessment.

In order to build an accurate random forest classification model and determine the optimal parameter configuration, first, this study set up a parameter network in the experiment, which contains the parameter combinations to be searched. This paper argues that these parameters are critical to building a model with superior performance. Next, a grid search using the GridSearchCV object was performed to find the best combination of parameters in the parameter network. By performing cross-validation on the training set, this article is able to evaluate the performance of different parameter combinations and choose the configuration with the best performance. Once the best combination of parameters was found, this study retrained the random forest classifier using these parameters. By applying an optimal parameter configuration, a better performing model is created. Finally, this study uses the trained model to make predictions on the reserved test set and evaluate the performance of the model.

2.2.3. K-Means Combined with Random Forest. Through K-Means clustering, the samples are assigned to different clusters, and the cluster labels are used as new features [13], so that the random forest can use clustering information for more accurate classification. In summary, on the basis of clustering, the random forest algorithm is further implemented. This method can comprehensively utilize the advantages of unsupervised clustering and supervised classification to improve the accuracy and stability of the breast cancer classification model.

2.3. Implementation Details

All models implemented are based on an open-source machine learning library called sklearn [14]. The K-Means algorithm uses the elbow rule to find the best k value, and calculates the silhouette coefficient and mutual information score corresponding to each K value to evaluate the clustering accuracy of different K values. The random forest model uses the parameter which has been set and uses accuracy as the main evaluation metric to measure how accurately the model classifies new samples. In addition, this study also considers other evaluation metrics, such as recall and F1-score, to obtain more comprehensive model evaluation results.

3. Results and Discussion

Based on the findings presented in Table 1, it is evident that the K-Means algorithm achieves the highest accuracy when the value of K is set to 5. However, it is important to note that breast cancer comprises six distinct types, indicating that relying solely on the K-Means algorithm is insufficient to meet the requirements of accurate cancer classification.

Table 1. K-Means clustering effect evaluation.

Value of K	Mutual Information Score	Silhouette Coefficient
4	0.725	0.219
5	0.854	0.190
6	0.749	0.195

DOI: 10.54254/2755-2721/43/20230813

Table 2. Hybrid Model Effect Evaluation (K=5).

Cluster	Accuracy	F1-Score
0	1.00	1.00
1	0.81	0.89
2	1.00	1.00
3	1.00	1.00
4	0.96	0.97

Table 3. Hybrid Model Effect Evaluation (K=6).

Cluster	Accuracy	F1-Score
0	1.00	1.00
1	0.95	0.97
2	1.00	1.00
3	0.94	0.97
4	0.94	0.97
5	1.00	1.00

Because of the supplement of the random forest model, the classification effect of the Hybrid model in this study meets the requirements of the data set. The model achieves the highest accuracy when it is divided into 6 categories, and the exact evaluation data are shown in Table 2 and Table 3 above.

This study demonstrates the potential of combining unsupervised clustering and supervised classification techniques in breast cancer classification. Mixed models provide powerful methods for identifying and differentiating different subtypes of breast cancer, facilitating individualized treatment and prognosis prediction.

However, further research and validation are needed to fully explore the potential of the K-Means and Random Forest hybrid model. Future research can explore different clustering algorithms, evaluate the impact of different feature selection methods, and verify the generalization ability of models on diverse datasets.

4. Conclusion

In summary, the integration of K-Means clustering and Random Forest classification in a hybrid model presents a promising strategy for breast cancer classification. The successful amalgamation of these distinct but complementary machine learning techniques has yielded a robust framework that exhibits notable improvements in both predictive accuracy and interpretability compared to conventional methodologies. The hybrid model facilitates more precise and insightful categorization of distinct cancer types by effectively leveraging the strengths of both unsupervised and supervised learning.

Building upon this research, future studies can further advance and optimize the approach. The research can delve into the integration of multimodal data sources, including genomics and clinical imaging data, to enhance the model's predictive capabilities and promote a more comprehensive understanding of the disease. Additionally, efforts must be made to validate the stability and applicability of the hybrid model in diverse patient cohorts and healthcare environments, ensuring its practical utility and widespread adoption.

In essence, this research underscores the transformative potential of machine learning in bolstering precision medicine approaches for breast cancer and, by extension, other complex diseases. By fostering a synergy between data-driven insights and clinical decision-making, this hybrid model offers a promising pathway for advancing personalized healthcare interventions, ultimately contributing to improved patient outcomes and a more comprehensive understanding of the intricate mechanisms underpinning cancer pathogenesis.

DOI: 10.54254/2755-2721/43/20230813

References

[1] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118 (2017).

[2] Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N., and Madabhushi, A.: Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. Scientific Reports, 3, 1564 (2013).

[3] Huang, P. S., Wei, C. Y., Chen, C. H., Wu, T. T., Chen, Y. Y., & Chen, C. C. An intelligent breast cancer detection system using support vector machines. Journal of Medical Systems, 35(5), 1165-1174 (2011).

[4] Cheng, J. Z., Ni, D., Chou, Y. H., Qin, J., Tiu, C. M., Chang, Y. C., ... & Wang, X. Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. Scientific Reports, p 6, 24454 (2016).

[5] Kazerooni, F., Lotfollahi, M., et al.: Deep neural network-based classification of breast cancer histopathological images using transfer learning. Computer Methods and Programs in Biomedicine, p 208, 106252 (2021).

[6] The National Health Service, https://www.nhs.uk/ (2023).

[7] IBM Watson for Oncology, https://www.ibm.com/watson/health/oncology-and-genomics/oncology/ (2023).

[8] Breast cancer gene expression - CuMiDa | Kaggle,[EB/OL], https://www.kaggle. com/datasets/brunogrisci/breast-cancer-gene-expression-cumida (2020).

[9] Feltes, B.C., Chandelier, E. B., Grisci, B. I., Dorn, M.: CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. Journal of Computational Biology, p 26 (4), pp 376-386 (2019).

[10] Lloyd, S. P.: Least squares quantization in PCM. IEEE Transactions on Information Theory, p 28(2), pp 129-137 (1982).

[11] MacQueen, J.: Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, p 1(14), pp 281-297 (1967).

[12] Breiman, L.: Random forests. Machine Learning, p 45(1), pp 5-32 (2001).

[13] Yu, Q., Chen, P., Lin, Z., et al.: Clustering Analysis for Silent Telecom Customers Based on K-means++, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020, 1: 1023-1027 (2023).

[14] Scikit-learn; Machine Learning in Python [EB/OL], https://scikit-learn.org/ (2023).