Application of Principal Component Analysis in the Diagnostic Classification of Breast Cancer

Kunye Luo

Civil Engineering, Guangdong University of Technology, Guangdong, 523083, China
Sebastian.young@my.sheltonstate.edu

Abstract. In recent years, there has been a steady increase in the incidence of breast cancer, positioning it as the leading form of malignant tumors among women. Consequently, leveraging artificial intelligence (AI) technology to accurately classify and diagnose breast cancer has emerged as a crucial field of study within machine learning. This investigation demonstrates the implementation of the Principal Component Analysis (PCA) technique in the diagnostic classification of breast cancer, offering novel perspectives and methodologies for the development of breast cancer classifier models. The experimental design incorporated a control group, wherein the original breast cancer diagnosis dataset was subjected to dimensionality reduction using the PCA method. Subsequently, random forest classification and diagnosis were employed, and the resultant accuracies were compared across different groups. The experimental outcomes indicate that a decrease in the dimensionality achieved through the PCA method correspondingly leads to a decline in classification diagnosis accuracy. Overall, the accuracy of classification diagnosis post-PCA dimension reduction is suboptimal, considerably inferior to the control group utilizing random forest classification. This disparity can be attributed to the excessive original data dimensionality within this experiment, resulting in substantial information loss during PCA dimension reduction. Therefore, considerable potential exists for enhancing the utilization of the PCA preprocessing method in this domain, necessitating further improvements.

Keywords: Machine Learning, PCA, Breast Cancer

1 Introduction

Breast cancer represents a significant burden among female malignant tumors. Due to the change of lifestyle, the incidence of breast cancer is increasing year by year, and it has become the first place of female malignant tumors. It is the second leading cause of cancer death in women, and seriously endangers the health and life safety of women. Every year, more than one million women die of breast cancer, and early detection, early prevention and early treatment have become urgent problems to be solved. Therefore, accurate diagnosis of breast cancer and its specific type holds paramount significance in effectively managing the disease and reducing mortality rates. With the rapid development of artificial intelligence, the use of computer-

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_72

assisted diagnosis of breast cancer has become a common means, computer-assisted diagnosis has made a lot of achievements in medical diagnosis, which is also considered in this study.

In the 1950s, Artificial Intelligence (AI) gained extensive attention and investment, and entered the development stage. In this period, the focus of the research is on how to solve the logical problem of machine intelligence, how to make machine learning and so on. Since 2015, AI has received unprecedented extensive attention and development. Key technologies such as deep learning, neural networks and natural language processing have made breakthroughs, and the application fields of AI are expanding. Autonomous driving, smart home and intelligent medical care are all fields with great development potential [1-6]. Among them, the development of intelligent medical field is very rapid, and AI is widely used in clinical disease diagnosis and disease treatment. In terms of disease diagnosis, data preprocessing technology and classification technology are essential, PCA is one kind of data preprocessing technology. Principal component analysis (PCA) is a commonly used technique in data preprocessing. It can reduce the data, make the data retain important information while reducing the dimension, and provide help for the subsequent medical diagnosis classification.

In the diagnostic classification of breast cancer, the diagnostic classification of breast cancer poses significant challenges, primarily pertaining to the preprocessing of sample data and the development of an effective data classifier. The dataset used in this study comes from Kaggle, it is called Dataset GSE45827 on breast cancer gene expression from CuMiDa, contained 151 samples,54676 genes and 6 classes. In the study of this paper, the study will use PCA method to do the preprocessing for the data of 151 samples in this dataset, and make classifiers by random forest, an unsupervised machine learning classification method. The primary focus of this paper revolves around assessing the impact of using the PCA method for data preprocessing on the diagnostic accuracy of the breast cancer classifier. Moreover, the study also aims to analyze the underlying causes that contribute to any observed changes in diagnostic accuracy resulting from data dimension reduction using the PCA method.

2 Method

2.1 Data Introduction

The dataset used for this experiment was obtained from Kaggle [7]. This dataset contains 151 samples with 54,676 genes and six diagnostic classifications related to breast cancer, most of data are about gene expression, it is mainly used for diagnostic classification in machine learning.

2.2 PCA Preprocessing

In order to complete the study smoothly, the study was compiled using PyCharm 2023.1.1. In terms of data preprocessing, the experiment used the PCA method. In

machine learning, principal component analysis is the common data preprocessing method, that is PCA [8-10]. Principal component analysis is a method for feature extraction. It is a commonly used unsupervised learning method, which uses orthogonal transformation to transform the observed data represented by linearly relevant variables into a few data represented by linearly irrelevant variables, and the linear irrelevant variables are called the dominant component. Principal component analysis is mainly used to discover the basic structure in the data, namely the relationship between the variables in the data. It is a favorable tool for data analysis, and it is also used for the pre-processing techniques of other machine learning methods. What the PCA does is to eliminate the redundant data, leaving only the most useful ones, in other words, to reduce the dimension of the data. From a mathematical point of view, for an n-dimensional column vector x, an m * n-dimensional matrix A, and an m-dimensional column vector b can be constructed, so that Y = Axb, the resulting Y is the m-dimensional column vector, the original X can be replaced with Y, so as to achieve dimension reduction. In this study, PCA will be used to reduce the breast cancer diagnosis data in various dimensions, and compare the accuracy of these data by using random forest classification method. After adjusting the compilation language and compilation environment, Importing the library needed for the institute including pandas numpy and sklearn. Then load the dataset and check the distribution and null values in the 'type' column. For the smooth conduct of the study, scale the features is necessary. Next, conduct PCA to do the data preprocessing. In this process, the proportion of variance occupied by each principal component needs to be calculated. After fitting the data set and calculating the variance contribution of the principal components and the cumulative variance contribution, the minimum number of principal components required to retain 95% information quantity is calculated. After completing these steps, you can obtain the dataset after dimensionality reduction and found the dimension was reduced from 5,4675 to 126, that means the study retained 126 most important attributes in the dataset. In this experiment, the study assumes that the most important factor affecting the final diagnostic accuracy is the dimension after dimension reduction, so the research needs to set the dimension as the control variable of this experiment, in order to make the experimental results of this experiment more intuitive and show the effect of the data dimension after dimension reduction on the accuracy, the study also reduces the data with dimensions of 252, 504 and 1008,the steps are similar to the above.

2.3 Machine Learning Models

Random Forest. After completing these steps, the study will use Random Forest to do the classified diagnosis using the preprocessed data. Random forest is a common classification algorithm that belongs to unsupervised learning in machine learning and is a classifier. A decision tree is constructed through the training set, so that its classification can be predicted for the new data. The principle is that the classification decision tree without pruning is constructed by the classification regression tree algorithm, and i is a random vector of independent and equally distributed data. Given the input vector x, each decision tree classifier votes to determine the optimal classification outcome. Previous studies of the same type showed that this method has

good classification accuracy in the classification diagnosis of breast cancer. The Random Forest method to determine whether breast cancer is divided into four steps: data preprocessing, extraction of feature vector, training model and model prediction. In detail, the research split the dataset into a training set and a test set then conducted Random Forest. In this step, the study needs to initialize the Random Forest classifier object and define the hyperparameter network. The next thing is to perform a grid search and output the best parameter combination and score, then predict on the test set and output the accuracy score on the test set. The purpose of this step is to intuitively see the accuracy score on the training set and the accuracy score on the test set. The last thing to be done is to evaluate the accuracy and print the confusion matrix, the study also did some visual treatments such as heat maps and PCA information plot which make the results more intuitive and easier to understand. After obtaining the accuracy of each group of dimension reduction data using random forest classification for diagnosis, the study also needs to set up a group of preprocessed diagnosis data, directly use random forest for classification and diagnosis, and compare the accuracy of direct diagnosis with the accuracy of diagnosis used the data after preprocessing. In this way, the experiment can compare the accuracy of the diagnostic classification of each group of data. By observing the accuracy score, the study can analyze whether the accuracy increases or decreases after the dimension change, and the effect of PCA method in breast cancer diagnosis. Through these comparisons, the experiment is also able to accurately analyze the important factors affecting the accuracy of the preprocessing diagnosis, this is significant for the future direction of improvement in this experiment.

3 Results and Discussion

Through the above method, the accuracy of five groups of models can be obtained, as shown in the graph. Obviously, by observing the accuracy data of each group of models obtained by the experiment, the study can find that the random forest group model without preprocessing has the highest diagnostic accuracy of 1.0000. After the PCA dimension reduction treatment, the model with dimension 126 had the lowest diagnostic accuracy of 0.7826,the model with dimension 252 was the second lowest. In general, the diagnostic accuracy of the random forest model was the highest, reaching 1.0. After the PCA pretreatment model, the higher the retained dimension, the higher the accuracy of the later diagnosis, but the overall accuracy is still not good.

This means that in this study, the pretreatment method of PCA was not effective in the diagnostic classification of breast cancer, and the lower the dimension of data reduction, the lower the accuracy of diagnosis was achieved. After analysis, the study believes that there are several main reasons for the poor diagnostic accuracy of PCA model. In this experiment, the main reason for the low accuracy of diagnostic classification is that the data dimension is too high, the data dimension is 54675 in the original dataset, if the data dimension is too high, using PCA to do the dimension reduction may lose a lot of information that lead to the original data cannot be well

reflect, in addition, if the data dimension is too high, may lead to overfitting in the process of dimension reduction. The second reason may be the uneven data distribution, and if the sample data is not distributed, PCA may not capture the main features of the data well, resulting in low classification accuracy. The third reason may be the data noise. If there is noise or outliers in the data, PCA may be affected, resulting in the reduced data can not well reflect the characteristics of the original data, thus leading to low classification accuracy. This experiment shows that the application of PCA method in the classification and diagnosis of breast cancer is not ideal, but there is still room for development in the application of medical classification and diagnosis. Therefore, in the process of pretreatment with PCA method, the above several factors affecting the accuracy need to be considered.

4 Conclusion

This study delves into the application of PCA as a data preprocessing method within the context of breast cancer classification and diagnosis. The investigation entails a comprehensive analysis of the intricate determinants that influence diagnostic accuracy within this framework. Although the empirical findings of this research illuminate the integration of PCA into breast cancer diagnosis classifiers, revealing its practical utility, it is evident that further advancements are requisite to attain an optimal diagnostic precision. Notably, the current experimentation underscores the salient impact of original sample data dimensionality on the ultimate diagnostic accuracy, thereby highlighting the imperative of meticulous consideration in practical implementation.

As with any scientific endeavor, this study is not devoid of limitations. A conspicuous limitation pertains to the absence of a control group characterized by varying dimensions, a facet that could provide a more robust assessment of PCA's influence on diagnostic outcomes. Furthermore, the experimental setup and procedural rigor of the equipment warrant heightened stringency to ensure the validity and generalizability of the results. These limitations thus beckon for rectification and refinement in subsequent research endeavors.

In its entirety, this study endeavors to forge a pathway toward enhanced breast cancer classification and diagnosis methodologies. By shedding light on the pragmatic integration of PCA within diagnostic classifiers, this work contributes to the ever-evolving discourse surrounding breast cancer research. It is envisaged that this research not only kindles novel avenues of exploration in breast cancer classification and diagnosis but also augments the broader domain of medical diagnosis through its empirical analysis of the PCA method's viability. As a consequence, this study aspires to be a catalyst for future investigations that will unravel the full potential of PCA in advancing medical diagnostic paradigms.

References

1. Yurtsever, E., Lambert, J., Carballo, A., et al.: A survey of autonomous driving: Common practices and emerging technologies. IEEE access, 8: 58443-58469 (2020).
2. Mao, J., Shi, S., Wang, X., et al.: 3D object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision, 1-55 (2023).
3. Kulurkar, P., kumar, Dixit, C., Bharathi, V. C., et al. AI based elderly fall prediction system using wearable sensors: A smart home-care technology with IOT. Measurement: Sensors, 25: 100614 (2023).
4. Malik, I., et al.: IoT-Enabled Smart Homes: Architecture, Challenges, and Issues. Revolutionizing Industrial Automation Through the Convergence of Artificial Intelligence and the Internet of Things, 160-176 (2023).
5. Qiu, Y., Wang, J., Jin, Z., et al.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72: 103323 (2022).
6. Steimann, F.: On the use and usefulness of fuzzy sets in medical AI. Artificial intelligence in medicine, 21(1-3): 131-137 (2001).
7. Kaggle.: Breast cancer gene expression cumida https://www.kaggle.com/brunogrisci/breast-cancer-gene-expression-cumida (2020)
8. Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G., et al.: PCA in studying coordination and variability: a tutorial. Clinical biomechanics, 19(4): 415-428 (2004).
9. Kurita, T.: Principal component analysis (PCA). Computer Vision: A Reference Guide, 1-4 (2019).
10. Maćkiewicz, A., Ratajczak, W.: Principal components analysis (PCA). Computers & Geosciences, 19(3): 303-342 (1993).