Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/17/20230936

Effectiveness of principal component analysis in breast cancer diagnosis: An empirical and visualization analysis

Kangjie Lu

International College, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

2021214943@stu.cqupt.edu.cn

Abstract. With the development of computer science technology, it is increasingly common to apply computer science technology to the medical field, and one of the typical examples is to predict the probability of cancer cells recurrence in cancer patients. However, there is still a lack of unified explanation for the impact of the principal component analysis (PCA) dimensionality reduction method used in prediction on the final experimental results. Therefore, the research topic of this paper is the effect of PCA dimensionality reduction method on the accuracy of cancer prediction. The methodology in this paper is as follows: First, patient data were collected, based on a usable machine learning repositor. It has 569 instances, which contains both malignant and benign tumors. Second, exploratory data analysis is performed on the data, detecting outliers and discovering existing correlated data. Then, the comparative method is used to explore the effectiveness of PCA dimensionality reduction method on the prediction accuracy by observing and analyzing tables and images. It is found that the PCA dimensionality reduction method plays a positive role for boosting the prediction accuracy, but the PCA method does not significantly improve the prediction accuracy for the data that has been processed and the dimension is not high.

Keywords: breast cancer, PCA, machine learning.

1. Introduction

Health is the foundation of comprehensive human development. With the develop of the society, lung cancer replaces the lung cancer as the most common types of cancer. In recent years, the incidence of breast cancer has gradually increased, which has seriously affected women's health [1,2]. At present, one of the treatment principles for breast cancer is still to detect and inhibit the cancer cells spreading as soon as possible. Moreover the pathological characteristics of breast cells are complex and diverse, which undoubtedly adds difficulty to the diagnosis of doctors [3,4]. Existing research can already build models based on machine learning methods to classify breast cancer tumors through detection data and labels [5,6].

Machine learning is one of the most important components of computer science. It aims at automatically learning descriptive patterns from samples for boosting performances [7]. This ability allows computers to solve problems without human guidance and improve over time. Machine learning is now widely used, from autonomous driving to speech recognition, from natural language processing to image classification, all of which can be achieved using machine learning techniques [8].

DOI: 10.54254/2755-2721/17/20230936

Common machine learning methods could be categorized into supervised, unsupervised, semi-supervised, reinforcement, and other learning strategies. The supervised setting requires humans to provide large amounts of training data to help computers learn. Unsupervised learning does not require manual provision of training data, but allows the computer to discover patterns from the data itself. The semi-supervised setting acts as a compromise between unsupervised and supervised learning. Reinforcement learning is the problem of training an intelligent agent for optimizing a goal under a specific environment [9].

This article will be systematically summarized and investigated on the dimensionality reduction of data using PCA methods and the accuracy of different nuclear functions for predicting breast cancer [10]. The discussion of such problems is very meaningful, and the analysis and induction of the patient's breast pathological data will help doctors diagnose breast cancer diseases and predict the recurrence of cancer cells in diagnosed patients.

2. Method

2.1. Dataset and visualization analysis

The data for this study comes from a usable machine learning repository collected from the University of California, Irvine. The dataset is made up of 569 instances, including both malignant and benign cases.

Exploratory Data Analysis (EDA) is an essential way for analyzing the data by visualization and statistics to explore the protentional data distributions and provide better data understanding. This method is important which could offer intuitive understanding of the data to assist researchers.

In the data analysis of this experiment demonstrated in Figure 1, it can be concluded from the observation of the images that concavity_point, area, compactness, concavity, perimeter, radius are supposed to follow an exponential distribution.

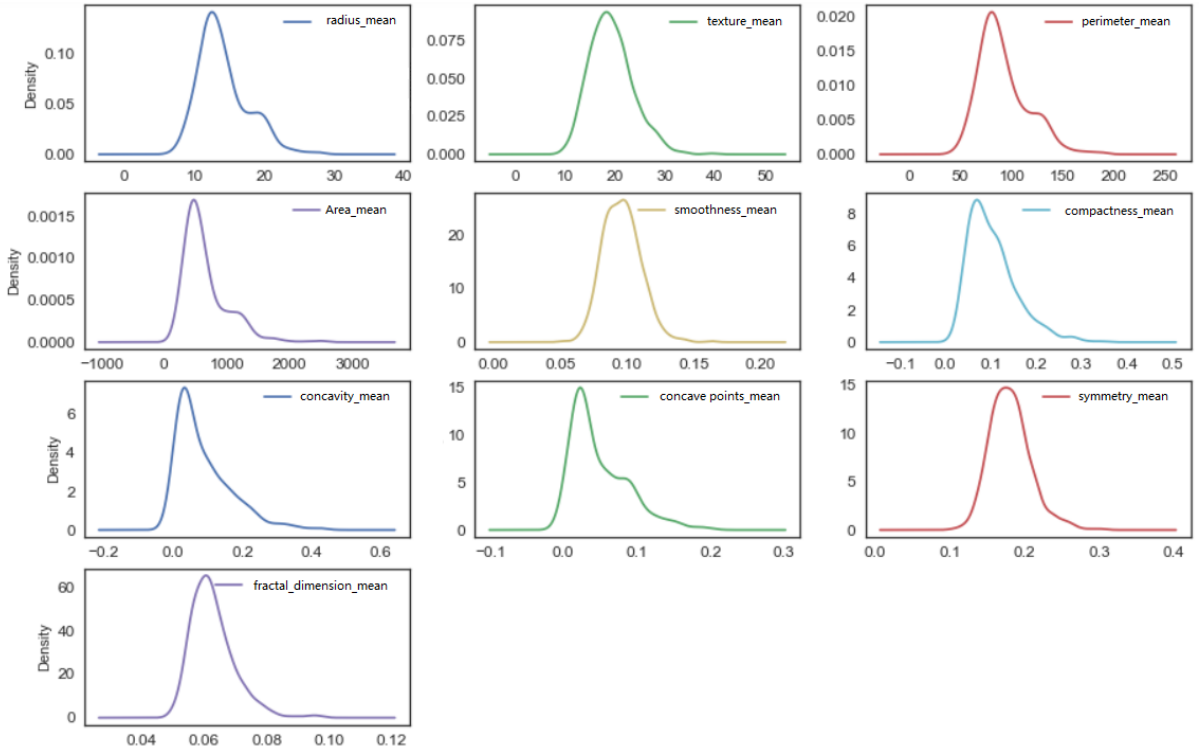

Figure 1. Distributions of features in the dataset.

Through the analysis of the scatterplot and the associated matrix as respectively demonstrated in Figure 2 and 3, it could be concluded that values from 1 to 0.75 indicates strong correlation. Moreover,

DOI: 10.54254/2755-2721/17/20230936

there is a strong positive correlation between the area of the tissue nucleus and the mean values of radius and parameter.

The fractal dimension is a measure of the complexity of an object's shape, and it has been observed that cancer cells with a more irregular shape tend to have a lower fractal dimension. In addition, certain parameters related to cancer cells, such as radius and texture, have been found to have a strong negative correlation with fractal dimension. Specifically, as the fractal dimension of cancer cells decreases, parameters such as radius and texture tend to increase. This suggests that larger tumors may have a more irregular shape, and cancer cells with a more irregular shape may also have a more varied texture.

These findings inspire the classification of cancer. Mean values of various parameters, such as radius and texture, can be useful in identifying malignant tumors, with larger values often indicating a higher likelihood of malignancy. Therefore, the negative correlation between fractal dimension and these parameters can be utilized as a tool to aid in the accurate classification of cancer.

In conclusion, the observed negative correlation between fractal dimension and certain parameters related to cancer cells provides valuable insights into the characteristics of malignant tumors. The use of mean values of these parameters in combination with fractal dimension can enhance the accuracy of cancer classification and potentially improve patient outcomes.

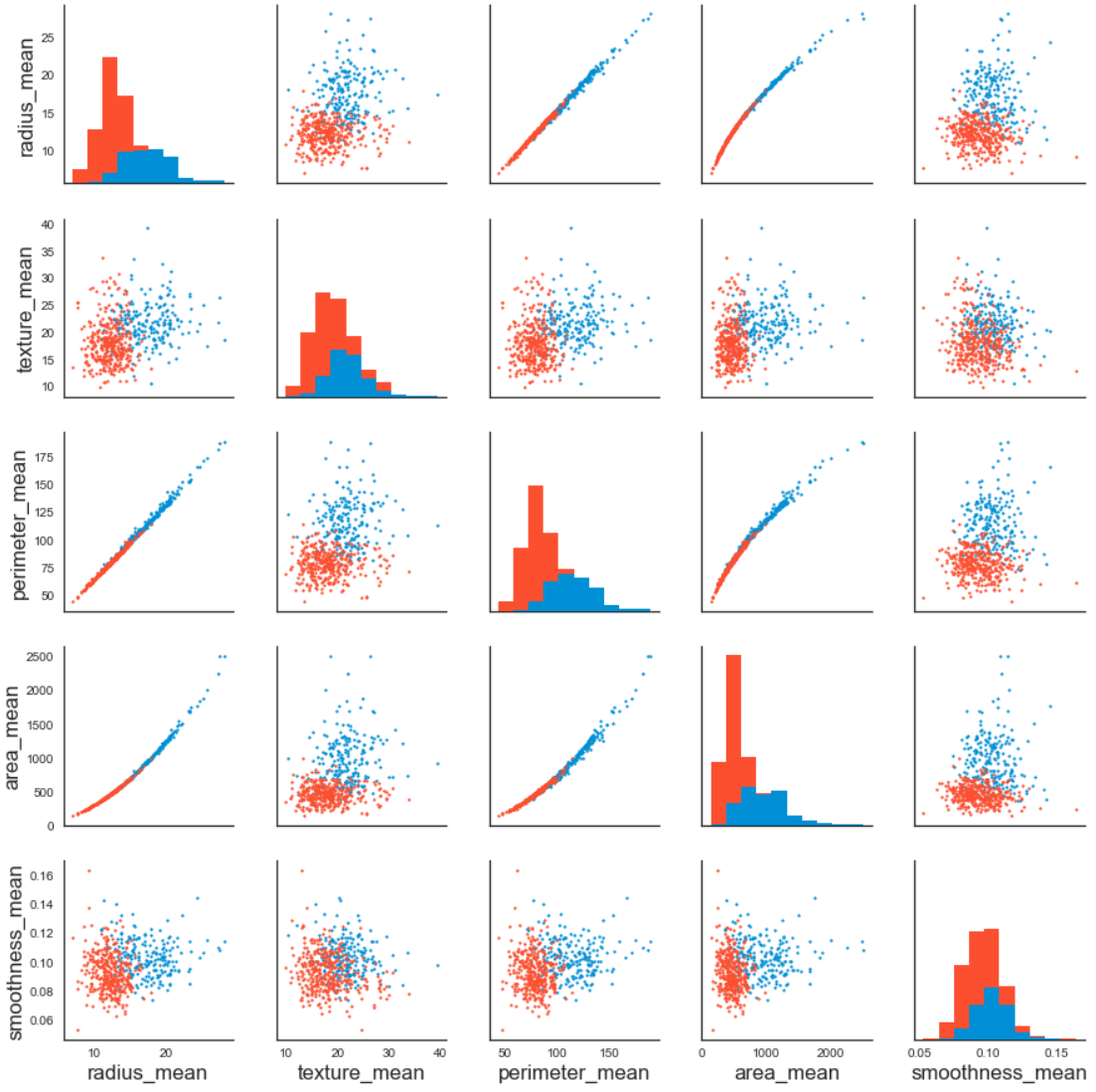

Figure 2. Corelations of representative features.

DOI: 10.54254/2755-2721/17/20230936

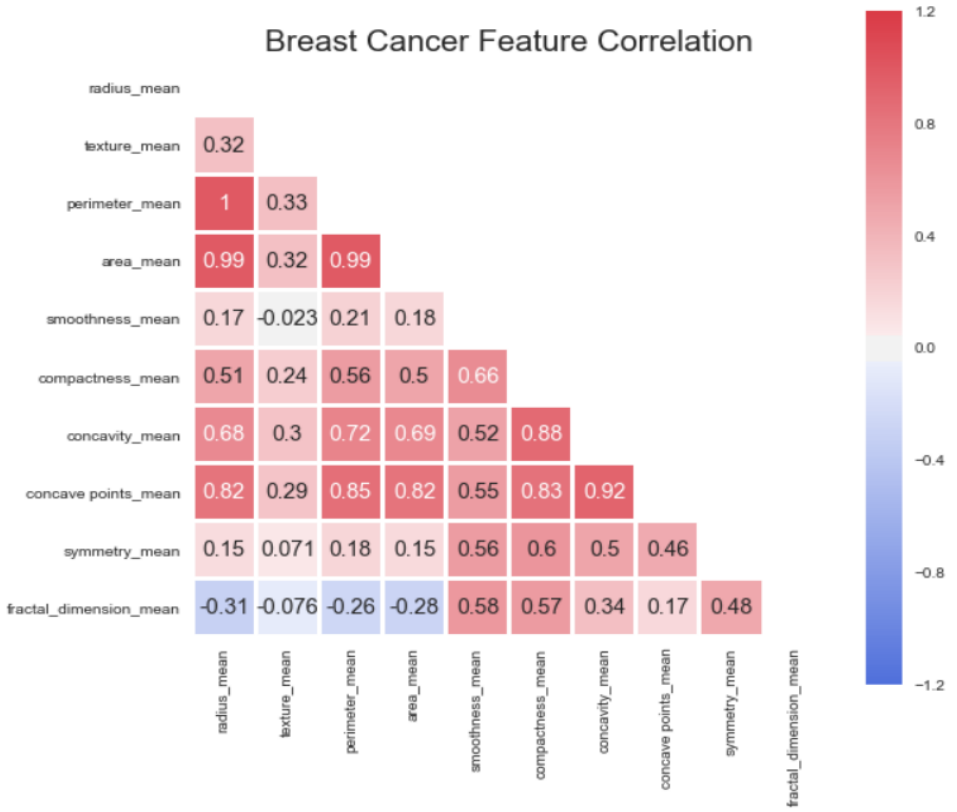

Figure 3. Correlation scores among various features.

2.2. PCA

PCA is a technique leveraged for dimensionality reduction of data. The method involves transforming high-dimensional embeddings into a series of orthogonal embeddings, known as principal components. They are selected based on the variance of the original data, with the first component being the direction that has the largest variance, followed by the next direction that is orthogonal to the previous one and maximizes the remaining variance. The primary aim of PCA is to decrease the data complexity while keeping its essential characteristics. This is achieved by compressing the data along its most prominent dimensions, while minimizing the information loss. By identifying the principal components of a dataset, it is possible to visualize and analyze the relationships between variables, as well as to identify trends and patterns in the data. PCA is a powerful tool in data analysis, and its applications are numerous, ranging from image processing and bioinformatics to finance and marketing. It is particularly useful in situations where there are many features and a small sample size, or when the features are highly correlated. By decreasing the data dimensionality, PCA simplifies the analyzing process and the interpretation of the results, making it easier to extract meaningful insights from the data.

Specifically, by keeping the first k axes with the most variance, where k is usually much smaller than the original dimensionality, the data can be projected onto a lower-dimensional space, while preserving most of the relevant information. PCA is a powerful tool for data analysis and feature selection in machine learning and other fields. It can help identify patterns and relationships within data sets, and can be used to extract the most important features benefiting to the overall variability of the data. This makes it a valuable tool in areas such as image and signal processing, where large datasets with high-dimensional feature spaces are common. Moreover, PCA can help in reducing noise and improving data visualization, which can aid in the interpretation of results.

DOI: 10.54254/2755-2721/17/20230936

2.3. SVM

Support Vector Machines (SVM) is a classic supervised model that distinguishes data points into separate categories. The basic SVM is a linear classifier that seeks the largest margin in the embedding space, which separates the data points into different classes. Unlike the perceptron, it has several kernel tricks, allowing it to operate as a nonlinear classifier. The learning strategy involves maximizing the interval, which can be formulated as a convex quadratic programming question. This approach is the same with minimizing the normalized hinge loss. The optimization algorithm for solving this problem is used in SVM's learning algorithm, which aims to identify hyperplanes for separating the training data with largest feature space interval. The hyperplane is defined as $w \cdot x + b = 0$, where x denotes the features and w is the weights, and b is the bias.

The kernel functions of SVM are very powerful, making the model fits well for various datasets. It embodies the classifier with non-linear segmentation capacity, allowing complicated decision boundaries. When applied on dataset with few samples but high dimensions, it could also achieve promising performances.

3. Result

Comparing the results using the PCA in Table 1, together with the results without the PCA in Table 2, it can be concluded that the performances could be slightly improved using PCA. However, there is little difference from the accuracy without the PCA method. The results reconfirmed that PCA method based on the data of this experiment had no significant effect on the improvement of accuracy, and the hypothesis that the use of PCA had a significant effect on accuracy was not proven.

However, the use of PCA method will have a great impact on the decision boundary of different classifiers, and the image obtained by the research results refers to Figure 4 and Figure 5, and the research hypothesis that the PCA dimensionality reduction method will have an impact on the decision boundary is confirmed.

Table 1. Results with PCA.

	precision	recall	f1-score
0	0.95	0.99	0.97
1	0.98	0.91	0.94
macro avg	0.96	0.95	0.96
weighted avg	0.96	0.96	0.96

Table 2. Results without PCA.

	precision	recall	f1-score
0	0.95	0.99	0.97
1	0.98	0.91	0.94
macro avg	0.96	0.95	0.96
weighted avg	0.96	0.96	0.96

DOI: 10.54254/2755-2721/17/20230936

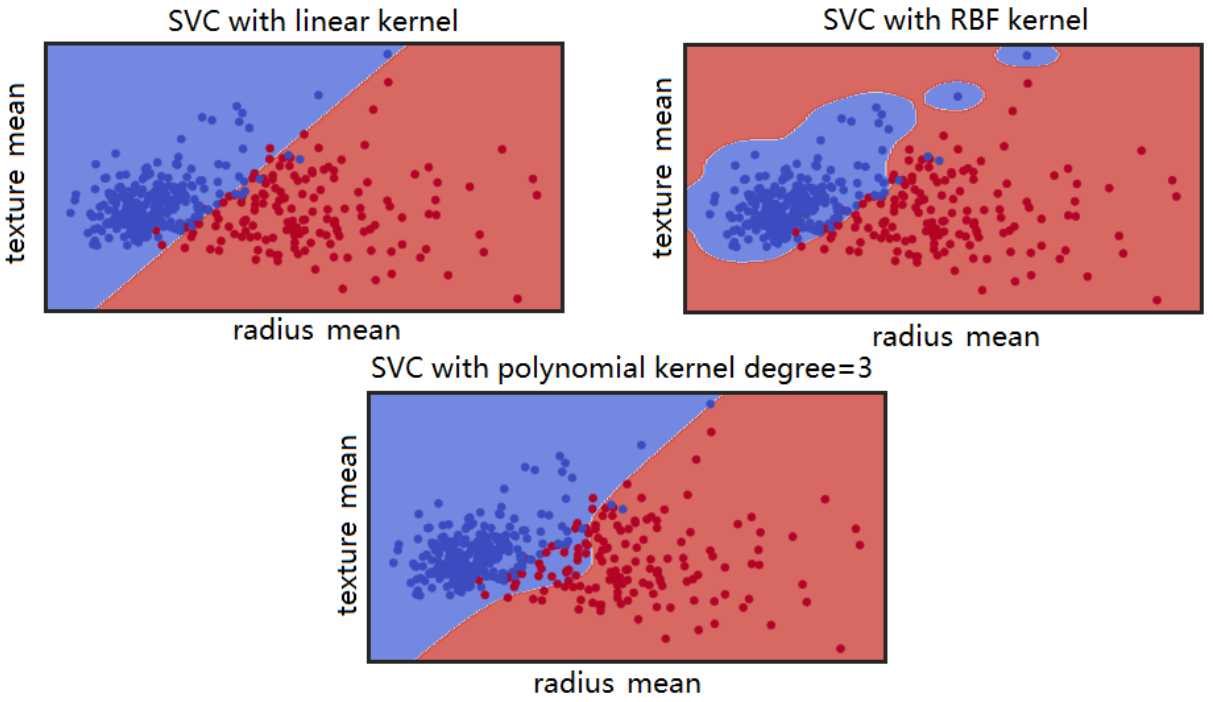

Figure 4. Sample distributions in feature space with various kernels with PCA.

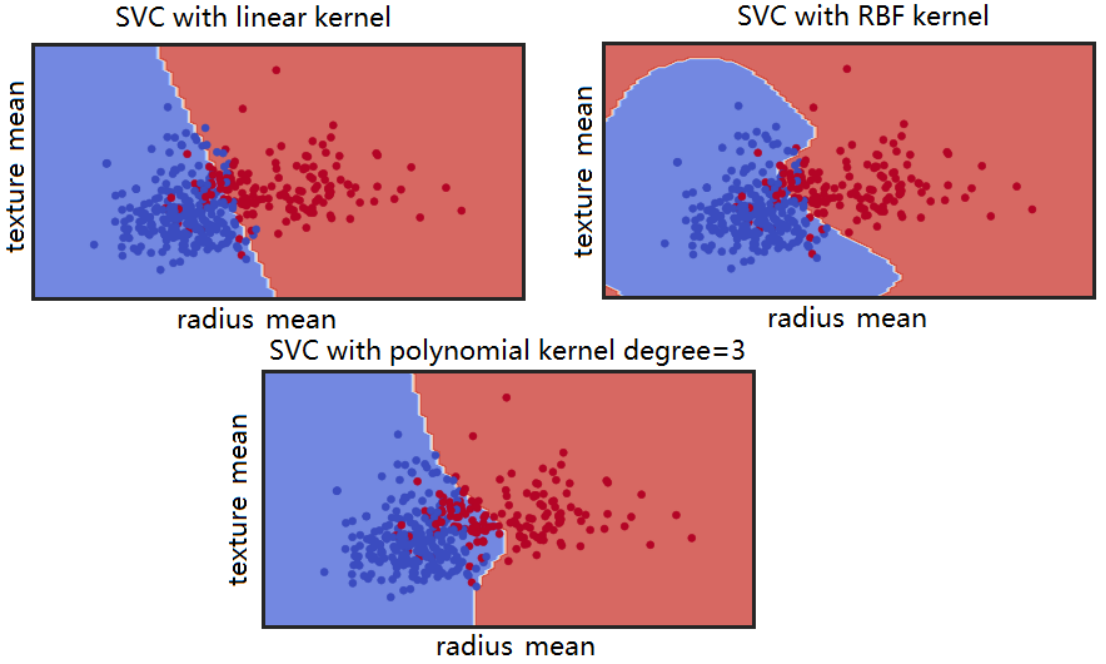

Figure 5. Sample distributions in feature space with various kernels without PCA.

4. Discussion

The hypothesis that the use of PCA had a significant effect on accuracy was not proven. The research hypothesis that the PCA dimensionality reduction method will have an impact on the decision boundary is confirmed. This research shows the breast cancer diagnosis problem leveraging SVM. It can be seen from the comparison that data normalization could boost the SVM performs.

DOI: 10.54254/2755-2721/17/20230936

The use of PCA dimensionality reduction could greatly impact the decision boundary of the RBF kernel function, while the image of the decision boundary of the linear and polynomial kernel function after using the PCA function, the decision boundary is symmetrical with the left diagonal, and the edge is smoother and the classification is more accurate. The clearer and more granular the boundary conditions, the more rational the decisions made on this basis and the more it will be resolved that must be solved. It can be concluded that the PCA dimensionality reduction method will still be helpful for forecasting. It could be calculated from the training dataset, which could be further generalized to the validation dataset.

PCA is helpful for prediction, but the accuracy improvement in this experiment is not great. After research, it may be for the following reasons. First, the data after data analysis and selection has the characteristics of Gaussian function. In fact, non-Gaussian functions can also use the PCA dimensionality reduction method, but it is not optimal. Under the Gaussian hypothesis, PCA is the optimal solution in the minimum sense of reconstructing the mean squared error. When the data is non-Gaussian, the optimal solution of the maximum likelihood is ICA, that is, ICA is a natural extension of PCA under non-Gaussian, and PCA is the reduction of ICA under Gaussian data, which is essentially due to independent and uncorrelated differences. PCA can of course be used for non-Gaussian, but it will miss non-Gaussian information, because PCA is essentially just manipulation of the first second moment, which only describes the Gaussian distribution and is whitening in ICA.

The PCA dimensionality reduction method has several distinct advantages. First, the principal components are orthogonal, which can eliminate the factors that affect each other between the original data. Second, the calculation process of PCA dimensionality reduction is not complicated, so it is simpler and easier to implement. Third, under the premise of retaining most of the main information, it has a dimensionality reduction effect. However, in some special cases, the PCA may not be the best method. At this time, if the PCA is applied for dimensionality reduction, it will have the following disadvantages. The meaning of the feature dimension of the principal component is vague and poorly explained. The criterion for PCA dimensionality reduction is to select the principal component that makes the original data have the largest difference above the new axis. But small variance features are not necessarily unimportant, and such a single criterion may lose some important information.

It should be noted that there are limitations in the operationalization of some variables in the text. Through the analysis of several existing dimensionality reduction methods, the following suggestions can be made. The KPCA algorithm will show obvious algorithm characteristics when selecting the appropriate kernel function. Ordinary PCA technology is difficult to make corresponding optimization choices in dealing with nonlinearity, so in the actual industry, the KPCA algorithm is a reliable choice for the processing of nonlinear data. In addition, the NMF dimensionality reduction algorithm in machine learning is also a very good dimensionality reduction method.

5. Conclusion

This study found that PCA dimensionality reduction methods have an impact on cancer prediction accuracy, but this effect will vary depending on data relationships and data dimensions. Compared with the predictions made before the experiment, the PCA dimensionality reduction method will indeed improve the accuracy, but the impact on the accuracy improvement of data analysis and low-dimension data is not as great as expected. In contrast, the PCA dimensionality reduction method has a greater impact on the decision boundary, which can be explained by the following facts. Before using the PCA dimensionality reduction method, the accuracy was 0.95, and is after that it was 0.96, and the improvement was not obvious. However, comparing the decision boundaries of SVC images shows that the image of the decision boundary of the linear and polynomial kernel function after using the PCA function, the decision boundary is symmetrical with the left diagonal, and the edge is smoother and the classification is more accurate. The clearer and more granular the boundary conditions, the more rational the decisions made on this basis and the more it will be resolved that must be solved. It can be concluded that the PCA dimensionality reduction method will still be helpful for forecasting.

DOI: 10.54254/2755-2721/17/20230936

In this study, the objective effect of PCA dimensionality reduction method on the accuracy of breast cancer prediction was first scientifically evaluated. This is conducive to the public understanding of the true role and impact of PCA dimensionality reduction method in the process of cancer prediction. Secondly, the analysis of the accuracy table and SVC image finds the specific performance of the PCA on the accuracy and decision boundary. Finally, the data in this study have already used the methods of data analysis before using the PCA dimensionality reduction method, so each dimension is not fully considered. As described in this research literature, for data analysis and low-dimensional data, PCA dimensionality reduction method is helpful for improving the accuracy of breast cancer prediction, but the improvement is not very large. In the future, the operation of the above related variables can be further refined to facilitate in-depth research on this topic.

References

[1] Waks, A. G., & Winer, E. P. (2019). Breast cancer treatment: a review. Jama, 321(3), 288-300.
[2] Fahad Ullah, M. (2019). Breast cancer: current perspectives on the disease status. Breast Cancer Metastasis and Drug Resistance: Challenges and Progress, 51-64.
[3] García-Aranda, M., & Redondo, M. (2019). Immunotherapy: a challenge of breast cancer treatment. Cancers, 11(12), 1822.
[4] Barzaman, K., Karami, J., Zarei, Z., Hosseinzadeh, A., Kazemi, M. H., et, al. (2020). Breast cancer: Biology, biomarkers, and treatments. International immunopharmacology, 84, 106535.
[5] Houssein, E. H., Emam, M. M., Ali, A. A., & Suganthan, P. N. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. Expert Systems with Applications, 167, 114161.
[6] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8, 150360-150376.
[7] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
[8] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 9, 381-386.
[9] Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2016). Machine learning basics. Deep learning, 98-164.
[10] Kurita, T. (2019). Principal component analysis (PCA). Computer Vision: A Reference Guide, 1-4.