Comparing the Performance of SVM and Decision Tree with PCA Feature Selection on Breast Cancer Wisconsin (Diagnostic) Dataset

Feng Chuang
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxfc4@nottingham.ac.uk

Shao-Sen Chueh
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxsc15@nottingham.ac.uk

Abstract— There have been many researchers using the Breast Cancer Wisconsin Dataset to train models to predict the classes, whether benign or malignant, of cancer. We aimed to provide an experimental result that demonstrates the best classification model to improve breast cancer diagnosis.

We first reviewed the research papers to compare the performances of algorithms that have been conducted previously. Then, we introduced our research methods, using principal component analysis (PCA) as our feature selection method, support vector machines (SVM), and decision trees as our algorithms to train our model. Finally, we proved the SVM model significantly outperformed other algorithms with the Breast Cancer Wisconsin Dataset and discussed the limitations of our findings.

Keywords— machine learning, R language, breast cancer, prediction, feature selection, classification

I. Introduction

Breast cancer is the most prevalent cancer diagnosed among women. There were more than 2.26 million women diagnosed with breast cancer in 2020 [1]. Early detection followed by medical treatment can increase the survival rate of a patient up to 90 percent or higher. Breast cancer presents in a variety of symptoms even without pain, therefore a detailed medical examination of the characteristics of cell samples is significant to identify if the sample is benign or malignant. A mature and sophisticated classification model consisting of machine learning algorithms can increase the accuracy of a classification and improve the efficiency of practitioners in disease diagnosing. In this paper we compared the performance of SVM and decision tree with PCA feature selection and suggested a more accurate model to apply in the future.

II. Literature Review

Breast cancer is the second leading cause of death among women. Reliable and efficient detection and classification system can improve the accuracy of finding whether the cell is malignant or benign. Using different characteristics of cells, a combination of multiple classifiers has shown higher accuracy levels than those of a single classifier. [1]

There are many research papers about the Breast Cancer Wisconsin Dataset. In order to identify the main features that could result in breast cancer, researchers have tried many classification and clustering methods. For example, Dubey et al [2016] experimented k-means algorithms with different methods of calculation, such as centroid (foggy/ random), distance (Euclidean/ Manhattan/ Pearson), split (simple/ variance), etc [2]. Also, Obaid et al [2018] experimented with three machine learning algorithms: Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbor (k-NN). The result showed that SVM outperforms the other 2 algorithms and achieved 98.1% accuracy [3]. Also, Anuradha and Telangana [2021] indicated that SVM is a highly accurate, sensitive and specific algorithm, and they showed that SVM-based approaches have achieved great accuracy on the Breast Cancer Wisconsin Dataset. [4]

Except for the studies about classification and clustering methods, other research have also been conducted in order to improve the accuracy of classification and clustering algorithms. Lavanya and Rani [2011] applied feature selection methods with classification algorithms on the Breast Cancer Wisconsin dataset [5]. The result showed that the accuracy of classification can be significantly improved with feature selection.

Ibrahim et al [2021] used correlation analysis and principal component analysis (PCA) as feature selection to perform classification algorithms on Breast Cancer Wisconsin dataset. Their result shows that with the 2 feature selection techniques, the accuracy has outperformed other state-of-the-art research results [6]. Therefore, it is fair to say that feature selection can considerably improve the performance of classification algorithms.

III. Methodology

From the literature review, we discovered that there are 2 important components in the analysis of the dataset. Firstly, a suitable classification or clustering algorithm can decide the accuracy of the result. Secondly, the result of classification and clustering can be further improved with feature selection.

The first method we propose is a method using PCA (feature selection) and the SVM model to analyze the breast cancer dataset. The reason we choose PCA for feature selection is that there are 9 features in the Breast Cancer Wisconsin Dataset and we were unsure about the importance of each element. Thus, PCA is applied to reduce redundant features. As for the classifier, the first method we choose to use is the SVM model because there is much previous research that shows SVM has performed well on this Breast Cancer Wisconsin Dataset and especially in the research by Obaid et al [2018]. SVM has outperformed other algorithms on the same dataset. The second classification method is the decision tree.

Another effective method we found during our research was the decision tree model, which achieves a high classification accuracy of 94.3% in the previous study by Yi [2017]. We decided to use this well performed decision tree model as a comparison to the SVM model. Therefore, we also applied the result of feature selection, classified into test sets and train sets, to our decision tree model.

A. Data Preprocessing

In the original Breast Cancer Wisconsin dataset, there were several empty cells in the *bare.nuclei* column. To ensure and enhance the performance, we had to drop all the cells that do not provide any information.

First, we coerced the character data type to numeric data type and assigned the value of NA to those empty cells. Then, we dropped the rows that contain the value of NA.

B. Principal Component Analysis (PCA)

PCA is a very useful feature extraction method that can reduce the dimensionality of a dataset. The purpose of PCA calculation lies in finding the components that account for the largest proportion of variance. The reason is that in theory, normally the data points mapped onto a dimension with larger variance are more dispersed and thus easier to be classified. Therefore, we chose PCA as the feature selection method in this research.

Before calculating PCA, we firstly normalized our data, adjusting the mean of each feature to 0 and the standard deviation of each feature to 1. The reason is that each feature has a different unit and thus the variance of each feature varies a lot. These features can affect the calculation of PCA.

C. Data Split

To improve data accuracy and avoid overfitting of the given dataset, the breast cancer dataset was divided into training and testing datasets. We specified the percentage of the rows that should be stored in datasets. Subsequently, we decided to assign an 80:20 partition of the training set which stores 547 objects, and a testing set which stores 136 objects. The same datasets were used and compared for the later analyses of the Support Vector Machine (SVM) and decision tree.

D. Support Vector Machine (SVM)

SVM is a very frequently used supervised machine learning algorithm. It is meant for classification and regression tasks in data science. The advantage of SVM is that it can perform not only linear classification, but also non-linear classification with some different kernel settings. The classification threshold used in SVM is choosing the hyperplane that maximizes the margin (maximal margin classifier) between the edges of different groups. However, not all the groups are cleanly separated and there might be a region that contains mixed data points from different classes. Therefore, allowing for misclassification is important when implementing SVM.

In this research, we used 10-fold cross-validation to allow some misclassification and chose the suitable classifier. Also, we experimented with some different settings. First of all, we compared the linear kernel model and the non-linear kernel model. Also, we experimented using a different number of principal components (PCs) to try to determine the best model.

E. Decision Tree

The decision tree model is another machine learning algorithm that performs both classification and regression tasks. The advantage of the decision tree is that it is easily interpretable and not sensitive to non-linear relationships and outliers. However, the decision tree model tends to overfit in many cases. To accommodate overfitting in the analysis, we minimized the data complexity through preprocessing and principal component analysis; and downsized our training set to 20% of the complete dataset.

IV. Results

To assess the accuracy and effectiveness with our proposed methods we performed several techniques to achieve it. We started with PCA analysis to understand the correlation between each column in the Breast Cancer Wisconsin Dataset and ultimately reduced the dimensionality of it. Hence, we split the dataset into a testing set and a training set according to the predefined partition of 80:20. With thorough preprocessing works, we were able to successfully complete the comparison of the support vector machine model and decision tree model.

A. Principal Component Analysis (PCA)

After the calculation of PCA, we used some visualization techniques to show the results of the PCA calculation and understand these components. First of all, we examined how much variance each component accounts for (See Figure 1).

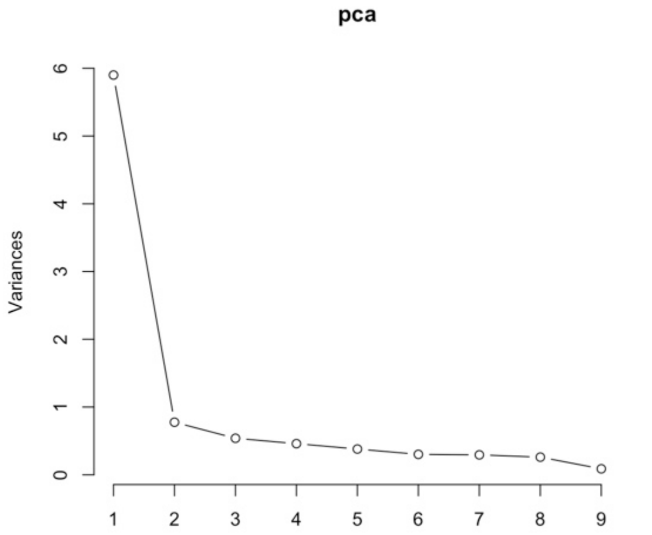

Fig. 1. The X-axis shows the principal components calculated, and the Y-axis shows the variance each component account for.

From Figure 1 we observed that the first component accounts for about 0.6 proportion of variance, while others account for less than 0.1 proportion of variance. It was difficult for us to choose how many components we are going to use in this research. Thus, we first picked the top 5 principal components that account for almost 90 % of cumulative proportion variance and attached them to our dataset. These components were applied to both of the SVM model and decision tree model independently to optimize the accuracy.

Also, we drew scatterplots of each PC to observe how well the two classes are separated in each PC. From Figure 2 we observed that the two classes are well separated in PC1, while significantly sparse in PC5.

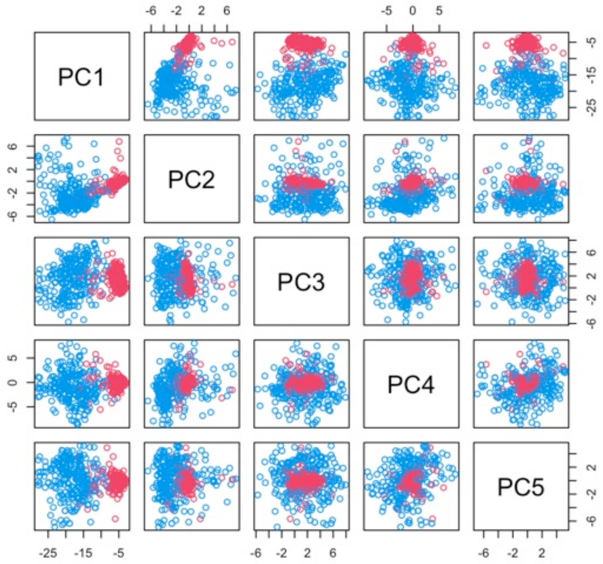

Fig. 2. Scatterplots of each PC, couloured with cancer classes (benign or malignant).

Finally, We drew the relationship between PC1, PC2 and the original features to understand the relationship between our components and the original features. (See Figure 3.)

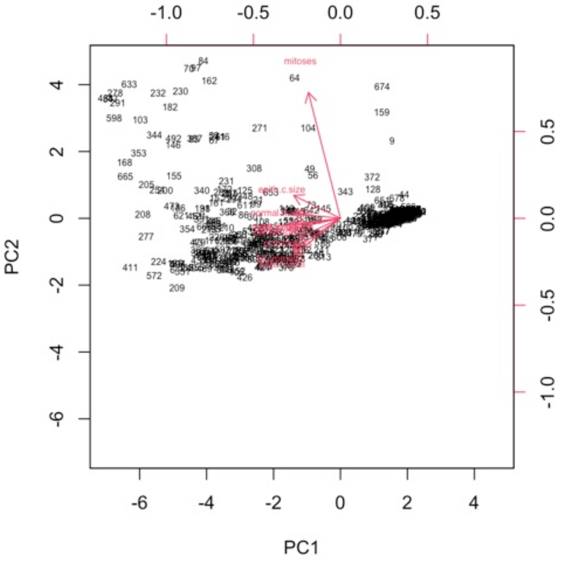

Fig. 3. The relationship bvetween PC1, PC2 and the original features.

From Figure 3 we can see that all of the original features are negatively correlated with PC1, while the correlations differ with PC2. Some correlations clearly tell that they are either positively or negatively correlated, while some correlations are not that obvious with PC2.

B. *Support Vector Machine (SVM)*

As we were unsure whether the classification of different classes is linear or nonlinear, we used PC1 and PC2 to draw a scatter plot to observe the distribution of the two classes- benign and malignant classes.

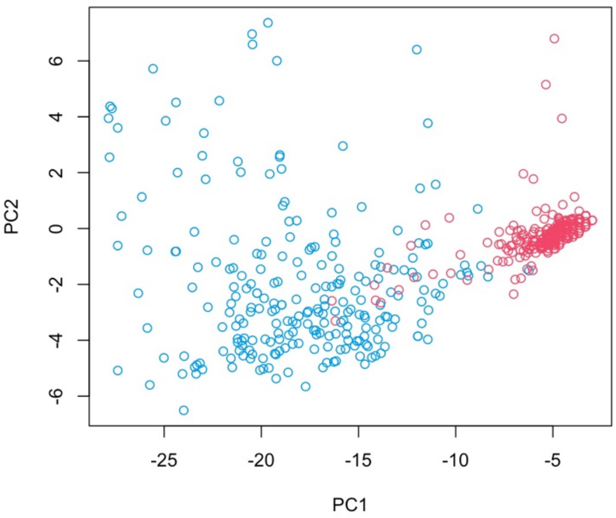

Fig. 4. Scatter plot of PC1 and PC2, coloured with the 2 classes (benign and malignant).

From the plot, we observed that the classification is more of a linear one. Therefore, we chose to use the linear kernel in the SVM. However, we also experimented with nonlinear kernels to compare the performance of each setting. Also, when calculating SVM, we tried to use a different number of PCs to train our model because we had not been clear about how many PCs are suitable for training. Excepting PC1, the proportions of variance the other PCs account for are quite close.

In order to evaluate the performance of the SVM model and the following decision tree model, we defined and computed the classification accuracy with the formula in the following figure. (See Fig. 5.)

$$Accuracy=(TP+TN)/(TP+FP+FN+TN)*100\%$$

Fig. 5. Formulation of evaluating model accuracy

PREDICT\TRUE	2 (benign)	4 (malignant)
2 (benign)	87	2
4 (malignant)	2	45

Fig. 6. SVM with Linear Kernel

Figure 5. is a confusion matrix of the result of SVM with the linear kernel. The result shows that with this model, we can achieve 97% accuracy in classification. We have tried it with different numbers of PCs (from PC1 to PC5), but there is no big difference in the results (with accuracy between 96%~97%). However, with nonlinear kernels, the results are a lot worse. The accuracy of nonlinear kernels achieves accuracy between 65% to 80% (with a different number of PCs). Therefore, we can be sure that the two classes in this dataset can be linearly classified. Figure 6. shows the scatterplot of PC1 and PC2 coloured with the predicted classification. In comparison with Figure 4, we can observe that the distribution is very similar.

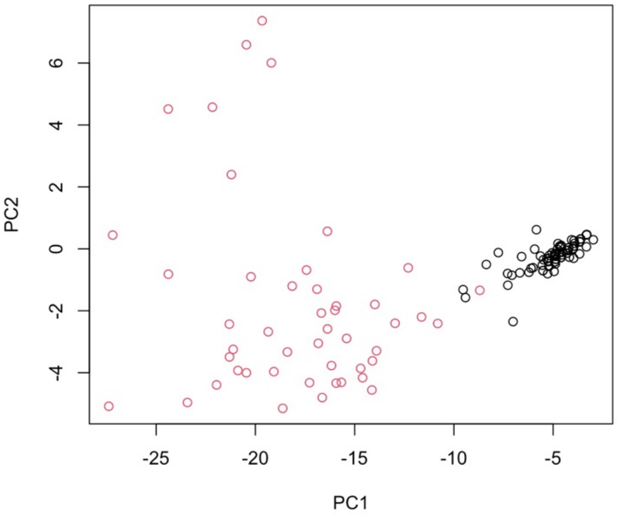

Fig. 7. The scatterplot of PC1 and PC2, coloured with the 2 predicted classes of cancer (benign and malignant).

C. *Decision Tree*

We constructed the decision tree model based on the result of PCA. Using the selected features, we were able to make a benign-or-malignant judgment. The result of the decision tree model is shown in Figure 7. We had tried it with different numbers of PCs with the decision tree model. The accuracy was slightly lower as we tried with 4PCs and 5 PCs, which only reaches around 92% accuracy. Trying 3PCs, the decision tree model achieves the best result, achieving 94% accuracy. Although the discrepancy between different numbers of PCs used in the current size of the dataset and model is minute, it can mislead the consequence if we input a larger size of data. Thus, we kept the result of implementing 3 PCs in the decision tree model as our optimal result.

PREDICT\TRUE	2 (benign)	4 (malignant)
2 (benign)	86	3
4 (malignant)	5	42

Fig. 8. Decision Tree

As we wanted to examine the condition of patients with benign or malignant cancers, the decision tree in Figure 8 illustrates the percentage in each test of different PCs. We obtained the preliminary result of the experiment. The result showed that the probability of diagnosing a benign tumor is 71% and the probability of diagnosing a malignant tumor is 29%.

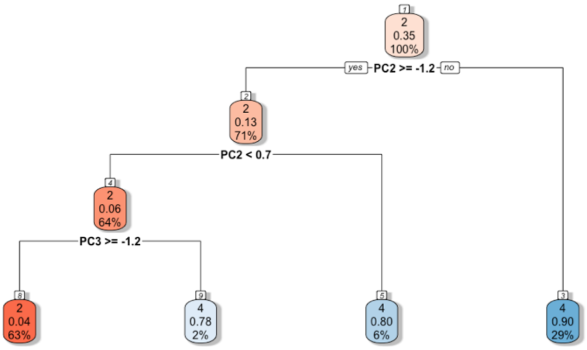

Fig. 9. The plot of the decision tree. The number 2 presents benign cases; the number 4 presents malignant casers.

V. Discussion

A. Limitation

Although we have tried to make sure all aspects of the methodology are attended to, there are still a few limitations of this research.

First of all, we have not been able to make sure the principal analysis works. Although our results with PCA have achieved accuracy as good as previous research with different feature selection techniques, we did not implement our algorithms without PCA, and thus we are not sure how much difference it would make. The second limitation is that the results of splitting data into a training set and testing set are different each time of calculation, which would also affect the results of accuracy of classification. Although accuracy of the result does not change a lot in most cases, the concern still exists and we do not have an effective way of solving the problem in this research.

B. Future Work

Due to the limitation of time and knowledge, the questions above are to be further researched in the future. Our result has shown a promising judgment on the performance of the support vector machine and the decision tree model. Both models perform competently, while the support vector machine performs slightly more accurately. In the future work, we suggest that the researcher can design and test a model combined with more than one classification method. Moreover, all kinds of models should be tested with larger sizes of data to evaluate the needs for improvement and verify the algorithm design.

VI. Conclusion

Classification Model	Support Vector Machine	Decision Tree
Accuracy	97%	94%

Fig. 10. The plot of the decision tree. The number 2 presents benign cases; the number 4 presents malignant casers.

Through the testing of two types of the classification model, we validated that the performance of the support vector machine (SVM) model meets our expectations and outperforms the comparative decision tree model. With a linear approach, both graphical and numerical results support that the two classes can be linearly classified with high accuracy of 97%. In contrast, a decision tree model can only attain an accuracy of 94% after adjustments. (See Fig. 9) Although the difference was not significant in percentage points, the number of erroneous classifications is likely to skyrocket when inputting a large dataset. Hence, we proved that the SVM or SVM-based models do outperform among different methods with Breast Cancer Wisconsin Dataset.

VII. References

[1] 2022. Breast cancer statistics. *WCRF International*. Retrieved April 21, 2022 from https://www.wcrf.org/cancer-trends/breast-cancer-statistics/

[2] G. Salama, M. B. Abdelhalim, and M. A. Zeid. 2012. Breast Cancer diagnosis on three different datasets using multi-classifiers. (2012). Retrieved April 21, 2022 from https://www.semanticscholar.org/paper/ab6c4f08484db95f8950d2637 6dbd22c03b19b21

[3] Ashutosh Kumar Dubey, Umesh Gupta, and Sonal Jain. 2016. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. *Int. J. Comput. Assist. Radiol. Surg.* 11, 11 (2016), 2033–2047. DOI:https://doi.org/10.1007/s11548-016-1437-9

[4] Omar Ibrahim Obaid, Mazin Abed Mohammed, Mohd Khanapi Abd Ghani, Salama A. Mostafa, and Fahad Taha AL-Dhief. 2018. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. Int. j. eng. technol. 7, 4.36 (2018), 160–166. DOI: https://doi.org/10.14419/ijet.v7i4.36.23737

[5] Anuradha Reddy. 2021. Support vector machine classifier for prediction of breast malignancy using Wisconsin breast cancer dataset. *ASIAN JOURNAL OF CONVERGENCE IN TECHNOLOGY* 7, 3 (2021), 57–60. DOI: https://doi.org/10.33130/ajct.2021v07i03.010

[6] Sara Ibrahim, Saima Nazir, and Sergio A. Velastin. 2021. Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. J. Imaging 7, 11 (2021), 225. DOI:https://doi.org/10.3390/jimaging7110225

[7] D. Lavanya, S. Padmavathi, and K. U. Rani. 2011. Analysis of feature selection with classification: Breast Cancer Datasets. (2011). Retrieved April 13, 2022 from https://www.semanticscholar.org/paper/c9aababc88cb7c7882535eddd3ac370027235024

[8] Yi, L., & Yi, W. 2017. Decision tree model in the diagnosis of breast cancer. *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC)*, 176–179. Retrieved April 24, 2022 from https://ieeexplore.ieee.org/document/8789297