International Journal of Engineering & Technology, 7 (3.7) (2018) 62-64

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Fuzzy PCA and Support Vector Machines for Breast Cancer Classification

Mohamad Faiz Dzulkalnine [1*], Roselina Sallehuddin [2], Yusliza Yusoff [3], Nor Haizan Mohamed Radzi [4], Noorfa Haszlinna Mustaffa [5]

[1]Faculty of Computing, Universiti Teknologi Malaysia, Skudai, 81300 Johor, Malaysia
**Corresponding author E-mail: m.faiz.dzul@gmail.com, roselina@utm.my,*

Abstract

Breast cancer is the leading cause of death among women in the world and early detection can increase the chance of survival for the patients. However, expert system and machine learning diagnosis are burdened with the presence of irrelevant data and noise which can reduce the accuracy of prediction and increase computational time. In this paper, Fuzzy Principle Component Analysis (FPCA) and Support Vector Machines (SVM) are proposed for the classification of breast cancer dataset. Experimental results on public breast cancer dataset show that the proposed method FPCA-SVM outperformed the benchmark models in terms of accuracy, specificity, and sensitivity and AUC value. The proposed model can assist doctors and medical practitioners for an early detection of breast cancer.

Keywords: *Feature Selection, classification, accuracy, Fuzzy PCA, SVM.*

1. Introduction

Breast cancer is by far the most common type of cancer among women and the leading cause of tumor-related deaths in the world. However, survival rates can be improved if the cancer is diagnosed earlier. Some reports stated that diagnostic techniques such as mammography and fine-needle aspiration cytology (FNAC) are lacking in the high diagnostic capability. Therefore, developing better diagnostic methods is very important [1]. With the aforementioned needs, machine learning has been introduced to help further improve diagnostic capability in breast cancer. Machine learning can process data faster and more precise, hence reducing the number of errors made by experts during diagnosis.

Classification is an example of a machine learning method widely used in medical diagnosis. However, most studies do not consider the presence of irrelevant features in the data. Irrelevant features can reduce classification accuracy and increase the computational time of the diagnosis [2]

Therefore, in this study, we attempt to improve the classification performance by applying feature selection prior to classification to identify the irrelevant features. SVM is used as classifier since it can avoid local minima and overfitting solutions.
Principal Component Analysis (PCA) is one of the prominent methods frequently used for data analysis and preprocessing classification problems. The advantage of using PCA for preprocessing data in classification is that it can reduce the number of data dimensions, lower computational cost, and increase accuracy performance [3]. In PCA, new attributes called principal components (PC) are defined as mutually orthogonal linear combinations of the original attributes. However, classical PCA often produces outliers that are known to influence the resulting principal component, thus affecting the classification results. Literature has shown that fuzzy membership can deal with the issue of outliers, especially in regression analysis [4]. Therefore, the objective of this study is to improve the capability of classical PCA as feature selection by replacing it with nonlinear Fuzzy PCA (FPCA). FPCA is used to generate a new set of non-correlated features in order to remove noise and to avoid using low variance variables. FPCA will identify the irrelevant features for breast cancer in order to improve SVM's classification performance.

The remainder of this paper is organized as follows. In the next section, literature review and related works are discussed. Then, the implementations of the proposed method are explained. Next, the experimental data and obtained results from the study are presented and finally, the summary and conclusion are discussed.

2. Materials and methods

2.1 Related Studies on Fuzzy Principal Component Analysis (FPCA)

A. Fuzzy Principal Component Analysis

Classical PCA adapts three organizing rules that are designed to work on data that do not contain any outliers. However, real data usually contain some outliers and often difficult to separate from the dataset. In Huber's book, it was demonstrated that the presence of even one outlier can determine an entire principal component. This happened when the outliers have the largest eigenvalues compared to the other attributes. In 1995, Xu and Yuille [5] proposed robust principal component analysis by self-organizing rules based on statistical physics approach to robustify the existing PCA by relating the organizing rules to energy function and proposed an objective function that considered outliers. They proposed an optimization function that replaced the standard eigenvector analyzing algorithm.

The standard method has two major problems. First, the standard method of calculating eigenvector was done in "batch way" but in a real application, the data were analyzed incrementally or "on-line" way. In another word, the standard method was computationally ineffective when a new sample was added. The second problem was the standard method will be heavily affected by the presence of outliers. Xu and Yuille's method was proven effective to resist outliers. However, Xu & Yuille algorithm's has a problem in determining the value of hard threshold during the training process. The small setting value of hard threshold will track the objective function to the minimum as the hard threshold values increase to infinity.

Yang and Wang [6] then extends Xu & Yuille's algorithms by implementing fuzzy membership into the consideration of the data cluster. An element with a high degree of membership that is close to the cluster center will contribute significantly to the weighted average. In contrast, elements with a low degree of membership which is far from the cluster center will not affect weighted average. Therefore, they proposed a fuzzy objective function and gradient descent optimization algorithm that can set threshold automatically. Here, only one parameter namely the fuzziness variable need to preset and it is use to determine the influence of the outliers to the obtained results.

In 2011, Pasi Luukka proposed another nonlinear fuzzy robust principal component analysis extended from Yang & Wang's method by pre-whiten the vector x [7]. The advantage of whitening a vector is the data features will be less correlated with each other. Therefore, when data are tightly clustered, it will be easier to differentiate different features and to identify the distance between the features. In this study, we implemented Pasi Luukka's method in data preprocessing of breast cancer dataset.

B. FPCA Algorithm

FPCA algorithm has nine steps as follow:

Step 1: Initially set the iteration count $t=1$, iteration bound T, learning coefficient α_0 soft threshold η to a small positive value and randomly initializes the weight w.

Step 2: While t is less than T, do steps *3-9*.

Step 3: Compute $\alpha_t = \alpha_0 (1 - t/T)$, set i=1.

Step 4: While i is less than n, do steps *5-8*.

Step 5: Compute the learning rate

Step 6: Update the weight, w.

Step 7: Update the temporary count

Step 8: Add 1 to i.

Step 9: Compute the new soft threshold, η and add 1 to t.

C. SVM steps

Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification. The SVM performance highly depends on the parameters values chosen in the training phase. These parameters include the followings:

1. Regularization parameter c, which determines the trade-off cost between minimizing the training error and the complexity of the model.
2. Parameter gamma g, of the kernel function which defines the non-linear mapping from the input space to some high dimensional feature space.
3. A kernel function used in SVM, which constructs a non-linear hyperplane in an input space. In this study, we employ cross-validation to find out the optimal parameter values of RBF kernel function.

2.2 Proposed Fuzzy Principal Component Analysis and Support Vector Machines

The proposed FPCA-SVM method consists of two phases, namely, feature selection by FPCA and classification by SVM. Firstly, FPCA is used to select the most significant features of the dataset by deleting the least significant features. After that, SVM classified the performance of the reduced dataset. Four statistical errors, namely accuracy, specificity, sensitivity, and AUC value are used as the performance criterion. Next, the differences in accuracy between the original dataset and the reduced dataset are compared. This process is iterated until the accuracy of the classification dropped. FPCA-SVM involves the following four steps.

Step 1: *Select the significant features.* This step obtained the PC scores for each of the features. The significance of the features is determined by the value of the PC scores. The higher the PC score, the higher its significance. In contrast, the lower the PC score, the least significant the feature.

Step 2: *Deletion of the least significant features.* When the PC scores of all the features were obtained and ranked, the feature with the lowest PC scores was deleted. However, before the feature with the lowest PC scores was deleted, the original dataset was first classified to obtain the base performance of the dataset so that we can compare its performance after the least significant feature was deleted.

Step 3: *Training phase by SVM.* The reduced subsets were split into three training-test partitions, namely, 80–20%, 70–30% and 50–50 to ensure the same class distribution in the subset. In order for SVM to generate the optimum training model, cross-validations were used to obtain the best pair of parameters(c, g) for the RBF kernel. In this study, we used 2, 5 and 10-fold cross-validations to obtain the optimal parameters. To choose the best c and g value, we first split the available data into k subsets. The cross-validation errors were then computed using the split error for the SVM classifiers using different values of c and g. The lowest cross-validation error for the value of c and g was then used for training. The performance outcomes from each of the cross-validations were kept and only the best results were presented.

Step 4: *Testing phase by SVM.* After obtaining the training model, we conducted the classification on each testing set accordingly. After the least relevant feature was deleted, the remaining data was classified by SVM to obtain its performance. We compared the classification accuracy from the reduced dataset and the original dataset. If the performance increased, it means that the deleted feature was insignificant and proved irrelevant to the dataset. The process was iterated by deleting the features one-by-one until the classification accuracy remains constant or deteriorates. Only then, the final accuracy of the testing set was obtained. The remaining set of features was then considered the most relevant and important features of the dataset.

3. Results and discussion

3.1. Experimental Data and Performance Measure

To evaluate the performance of the proposed model, Wisconsin Diagnostic Breast Cancer dataset (WDBC) is used. WDBC includes a total of 699 instances represented by 9 features and a predictive class. The class attribute has only two categories namely B (benign) and M (malignant). Class 'B' and 'M' has 458 and 241 instances respectively. In order for the SVM to classify the data, the categorical data are transformed into numerical values which are 0 and 1 for benign and malignant respectively. FPCA and SVM were implemented using MATLAB software.

3.2. Results

As we can see from Table 1, the highest classification accuracy, 100%, was achieved for the 80–20 training-testing partition and was produced by FPCA-SVM with seven features. The second highest accuracy performance, 97.4%, was produced by PCA-SVM and was in the 80-20 dataset with five selected features. Meanwhile, all classification results obtained from SVM using original dataset were lower than those obtained from PCA-SVM and FPCA-SVM. As shown in Table 1, the classification accuracy of the original data was 55.3009%, 77.9904%, and 76.9231% for

the 50-50, 70-30, and 80-20 training testing partitions, respectively.

Table 1: Comparative classification results between FPCA, PCA and original dataset

	Method								
	FPCA			PCA			Original		
Training-Testing partition	50-50	70-30	80-20	50-50	70-30	80-20	50-50	70-30	80-20
Accuracy (%)	92	97.2153	**100**	73.429	95.9638	97.4359	55.3009	77.9904	76.9231
Sensitivity (%)	90	98	**100**	64.42	94.48	96.67	51.98	84.05	83.33
Specificity (%)	98	98	**100**	100	100	100	100	100	100
AUC value	0.9393	0.9615	**1**	0.8221	0.9724	0.9833	0.7959	0.9202	0.9167

The relatively low performance of this classification is due to the existence of irrelevant and useless features, which ultimately decrease the performance of the classifier. FPCA-SVM has the best ability to distinguish relevant and irrelevant features in the related dataset, as it has the highest classification accuracy for every partition.

In addition to accuracy, specificity and sensitivity were also computed as shown in Table 1. The highest sensitivity and specificity rates were 100% for both metrics in the 80-20 partition. 100% sensitivity means that our method would not miss any patients who have the disease. Furthermore, 100% specificity means that our method would not erroneously classify anyone who is disease-free as having the disease. Seven features that were identified by FPCA-SVM, namely 'Clump Thickness', 'Uniformity of Cell Shape', 'Marginal Adhesion', 'Bare Nuclei', 'Normal Nucleoli', 'Single Epithelial Cell Size', and 'Mitoses', produced the highest classification accuracy. These seven features were shown to be the most informative features for classifying breast cancer.

The features identified by FPCA are very different from PCA is terms of quantity and order of importance. For example, FPCA selected seven features while PCA five features. Moreover, 'Clump thickness' rank differently in FPCA compared to PCA. FPCA ranked it as the most significant features while PCA rank it as the least significant selected features. The difference in order is what makes FPCA perform better than PCA. The features that FPCA identified also in line with a previous research (Polat, Şahan et al. 2007) that has been published thus proven the validity of our proposed method.

Lastly, areas under the ROC curves (AUC) are computed and these values can be used for evaluating the classifier performance for different training/testing partitions. ROC curve is a reliable technique based on the values of true and false positives. It is parameterized by the probability threshold values. The true positive rate represents the fraction of positive cases that are correctly classified by the model. Therefore, it provides a trade-off between sensitivity and specificity. The maximum AUC is 1, which indicates to a perfect classifier. Table 1 shows the AUC values of each of the partition. Our proposed method obtained an AUC value of 1 at the 80-20 partition which indicates a perfect classifier.

4. Conclusion

In this paper, a combination of Fuzzy and PCA was adopted as feature selection to find the optimum significant factors that affect breast cancer classification. The purpose of using FPCA instead of PCA is to overcome the limitation of handling outliers. Therefore, FPCA can help to produce better learning and generalization ability in SVM classifier. Therefore, FPCA-SVM model was relatively stable and can converge faster globally. To validate the performance of FCPA-SVM model, comparison with PCA-SVM and SVM were implemented. Experimental results showed that FPCA-SVM model yields the best classification performance in terms of accuracy, specificity, sensitivity and AUC compared to both other models. Therefore, the proposed model can be used to

assist medical practitioners in the healthcare practice for early detection of breast cancer.

Acknowledgement

This study is supported by the Fundamental Research Grant Scheme (FRGS vot : 4F738) that sponsored by Ministry of Higher Education (MOHE). Authors would like to thank Research Management Centre (RMC) Universiti Teknologi Malaysia, and Soft Computing Research Group (SCRG) for the support in research activities.

References

[1] Polat, K., et al. (2007). "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism." Expert Systems with Applications 32(1): 172-183.

[2] Chen, H.-L., et al. (2011). "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis." Expert Systems with Applications 38(7): 9014-9022.

[3] Sun, Y. and D. Wu (2008). "A RELIEF based feature extraction algorithm." Proceedings of the 8th SIAM International Conference on Data Mining: 188–195.

[4] Tanatavikorn, H. and Y. Yamashita (2016). "Fuzzy Treatment Method for Outlier Detection in Process Data." Journal of Chemical Engineering of Japan 49(9): 864-873.

[5] Xu, L. and A. L. Yuille (1995). "Robust principal component analysis by self-organizing rules based on statistical physics approach." IEEE Transactions on Neural Networks 6(1): 131-143.

[6] Yang, T.-N. and S.-D. Wang (1999). "Robust algorithms for principal component analysis." Pattern Recognition Letters 20(9): 927-933.

[7] Luukka, P. (2010). "Nonlinear fuzzy robust PCA algorithms and similarity classifier in bankruptcy analysis." Expert Systems with Applications 37(12): 8296-8302.