# Building Minimal Classification Rules for Breast Cancer Diagnosis

Phonethep Douangnoulack
International College
King Mongkut's Institute of Technology Ladkrabang
Thailand
Email: phonethepdouangnoulack@hotmail.com

Veera Boonjing
International College
King Mongkut's Institute of Technology Ladkrabang
Thailand
Email: kbveera@kmitl.ac.th

***Abstract*— A rule based classifier is widely applied in breast cancer diagnosis. The classifier with a good performance of disease classification have been developed and highly required over the past decades. Since classification rules are derived from previous diagnosis with a large amount of features, it challenges to build a minimal number of rules with high performance while retaining all diagnosis information. The Principal Component Analysis (PCA) is known as a lossless data reduction technique with good classification performance. Therefore, this paper aims at finding the best performance classifier giving minimal classification rules by employing PCA. Based on experiment result on Wisconsin Breast Cancer data set, the J48 decision tree classifier is found to be the best among the three classifiers: J48 decision tree, Reduced Error Pruning Tree, and Random Tree.**

***Keywords—Rule Based Classifier; Decision Tree; PCA; Breast Cancer Diagnosis.***

## I. INTRODUCTION

A rule based classifier plays an important role in modern breast cancer diagnosis. The good classifier equips with high accurate classification rules obtaining from historical diagnosis. Since each diagnosis consists of a large amount of data features, it challenges to build minimal high accurate classification rules from such historical data. Basically, feature reduction techniques could help reduce a number of classification rules. But the trade-off is classification performance. However, if we could find a technique of feature reduction giving high classification accuracy, it would help obtain minimal high accurate classification rules. Fortunately, the Principal Component Analysis (PCA) is a data reduction technique giving new features (less than original features) that strongly differ across the classes. Hence, rules obtained from these new features always give classification performance better than rules of original features. Therefore, this research proposes to use the PCA as a data reduction technique to achieve its goal of obtaining high accurate minimal classification rules. Among decision tree classifiers, these three classifiers namely J48, REP Tree, and Random Tree are known of their ability of providing rules [8] ready to use in a rule based system. Therefore, the research aims at finding the best classifier, in terms of number of rules and classification accuracy, among them on PCA reduced data of Wisconsin Breast Cancer Data set (WBCD).

The rest of the paper is organized as follows. Section II presents related works. Section III describes the methodology. The experimental results are given in section IV. And Section V concludes the paper.

## II. RELATED WORKS

Many researches have been conducted on WBCD to obtain a high performance classifier supporting breast cancer diagnosis. Hind Elouedi et al [1] proposed a hybrid diagnosis approach of breast cancer based on decision trees and clustering. It is evaluated classification results by distinguishing different types of breast cancer. The aim is to improve the quality of classification and clustering of WBCD. The experimental results show that the splitting up of malignant instances into two clusters, and submitting them to the decision tree algorithm, they have gotten better results up to 95.14%. F.Kharbat and H.Ghalayini [2] presented a case study for building ontology from the set of rules which generated by a rule based learning system. The algorithm is used to extract and represent the rules generated from the original data based on WBCD in developing ontology elements. The results show that the rule set with only 25 rules can describe two concepts (Benign and Malignant). P. Hamsagayathri and P. Sampath [3] proposed to find the best performance of the four different decision tree algorithms for breast cancer classification such J48, REP Tree, Random Tree, Random Forest and Priority based decision tree. The experimental results indicated that the Random Forest presents the highest accuracy of 96.70 %, while Priority based decision tree, Random tree, REP Tree, and J48 classifiers gave the accuracies of 94.70 %, 94.13 %, 94.13 %, and 93.56 %, respectively. Among them, the random forest classifier could not produce classification rules. Ronak Sumbaly et al [4] built a detection model of breast cancer in its early stages based on WBCD. J48 decision tree classifier is used to model actual diagnosis. According to the results show that the performance of J48 achieved 94.56 % classification accuracy. J48 has the ability to generate the simple rules, flexible and highly efficient algorithm for breast cancer diagnosis problem and it is also maintain the accuracy in estimation. Chandra Prasetyo Utomo et al [5] applied Artificial Neural Network (ANN) with extreme learning machine technique to compare with BP ANN for diagnosing breast cancer. These techniques are to support in medical decision.

978-1-5386-4015-9/18/$31.00 @2018 IEEE

The results show that ELM ANN classifier can classify better than BP ANN. Nevertheless, these network problems failed to give the rules and take more time in computational complexity. Smita Jhajharia et al [6] implemented hybrid prediction model which combines principal component analysis (PCA) technique with ANN for feature processing and pattern recognition. This hybrid prediction model is compared with other classification algorithms (SVM, NB, DT, IBK, OneR). The experimented results show that PCA+ANN is the most effective. Kathija and S. Nisha [7] presented the performance of selecting the smallest subset of features from WBCD by using SVM and Naive Bayes to build efficiency classifiers.

## III. METHODOLOGY

### A. Principal Component Analysis(PCA)

PCA is a mathematical method used for data analysis. It is one of the most significant features extraction techniques [9]. Normally, PCA transforms a set of dependent variables into a set of independents which handles with uncorrelated variables called Principal Component (PC). Most of the largest possible variances will be retained in the first PC and then the next PCs will decrease the possible variances [10]. The objectives of PCA are to reduce the dimension of the data and select new variables that relevant to the best outcome. There are two approaches using in PCA .i.e Eigenvalues and Eigenvectors. An eigenvector represents the direction of the line (horizontal, vertical, etc) and an eigenvalue is a number of variances in the data's direction of eigenvector. The basic process of reducing data dimension by PCA which can reduce the rules of the model. It can be explained as below:

**Algorithm 1** PCA algorithm

1: Re-center the original dataset to the origin at means zero

2: Compute the sample variance-covariance

3: Compute the eigenvalues and eigenvectors.

4: Decision which principal components should be retained based on the eigenvalues in order to select highest to lowest eigenvalues. It could achieve 95% confidence interval.

5: Find the transformation matrix based on selection of PCs

### B. Decision Tree J48

J48 is an algorithm used to create a decision tree for decision making. It is an implementation of C4.5 algorithm by using Java application in the Weka Data mining tool [11]. Many problems have been solved by decision tree classification approach based on dividing and conquering strategy. It can be used to predict an unseen data set based on the various attribute value of the available data. The decision tree is represented by a rule based (if-then rules) which described by nominal and numeric properties. The construction of decision tree is built from a root node at the top of the tree to any leaf node that defined the feature. A branch feature may stop into a leaf node when searching for subset instance in the same class or may further create the leaf node when the nodes are not the same class. Every branch from the root node to leaf node is represented as a rule. It uses Gain Ratio as a splitting condition to separate the data set for normalizing the data into the form of information gain. The highest value of information gain ratio is selected as a root node and then splitting process is continued until reaching the leaf node.

$$Gain(S,A) = Entropy(S) - \sum_{j \in Values(A)} \frac{|S_j|}{|S|} Entropy(S_j) \quad (1)$$

Where, $j$ is possible values of $A$ , and $A$ is a set of all possible attribute. $S$ is a set of samples $\{X\}$, and $S_j$ is a subset where ($X_A = j$). There are two parts in the equation (1). For the first part is entropy of original collection $S$ and the second part is the expected value entropy which calculates the sum of the entropies of each subset weighted by the fraction of examples

$$SplitEntropy(S,A) = -\sum_{i=1}^{n} \frac{|S_j|}{|S|} \times \log_2 \frac{|S_j|}{|S|} \quad (2)$$

Where, $SplitEntropy(S,A)$ is separated information of $A$ on the value of the categorical attribute $S$ .

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitEntropy(S,A)} \quad (3)$$

### C. Reduced Error Pruning(REP) Tree

REP Tree is constructed by decision tree that is used the information gain as the splitting condition. REP is used for pruning. It is the simplest ideas by using a pruning set to evaluate the performance accuracy of node and leaf depend on the decision tree process. REP Tree could reduce the error rate when applying with unseen data. Missing values are solved by applying C4.5's method with fraction instance. Moreover, it used less time for learning a model [12]. REP Tree is explored by beginning from bottom-up strategy.

### D. Random Tree

Random tree classifier is generated by randomly select features from a set of the tree that is possible with a different instance of the training data [12]. It uses a Bagging idea which provides random data set for creating a decision tree. It is a powerful technique to make a classification which is challenging to over-fit. The combination of a large set of Random Tree generally produces a correct model. Random Tree is selected a test set depends on a given number of the random features of each node. The decision on each node is random the procedure without pruning.

### E. Research Method

The research is an experiment on WBCD. It builds three classifiers (J48, REP Tree, and Random Tree) on the original

dataset and on the PCA-based reduced data set. Evaluation of each classifier is in terms of number of rules obtained and classification accuracy with 10-fold cross-validation. WEKA tool version 3.8.1 framework [13] is used as a tool in this study.

### F. Data Description

The data set of Breast Cancer reported by Dr. Wllliam H. Wolberg was collected as samples of clinical provided by the University of Wisconsin Hospitals. The data set has been stored in the UCI Machine Learning repository [14]. The total number of instances consists of 699 samples with 11 attributes, but some data have missed about 16 samples. There are two classes such as Benign (444 instances) and Malignant (239 instances). The numbers between 1 to 10 are used to record in the domain of each attribute. Sample code number is only an id number of the instance that does not affect the model. In the training and testing phases of the classification do not include the sample code number as in Table I.

### G. Data Preprocessing

Data preprocessing is a preliminary phase to perform on raw data which applies data normalization and separates incomplete data, outliers data and inconsistent data before the data is used to other procedures. PCA is a procedure to transform the data dimension and find a new set of variables by selecting the subset of principal component without losing the important feature. The best significant subset evaluation has been collected and used in the next step for the experiment.

TABLE I. ATTRIBUTES DESCRIPTION

| **Breast Cancer Dataset** | | |
|---|---|---|
| ***Attributes*** | ***Range*** | ***Data Type*** |
| Sample Code Number | No | Numeric |
| Clump thickness | 1-10 | Numeric |
| Uniformity of cell size | 1-10 | Numeric |
| Uniformity of cell shape | 1-10 | Numeric |
| Marginal adhesion | 1-10 | Numeric |
| Sigle epithelial cell size | 1-10 | Numeric |
| Bare nuclei | 1-10 | Numeric |
| Bland chromatin | 1-10 | Numeric |
| Normal nucleoli | 1-10 | Numeric |
| Mitosis | 1-10 | Numeric |
| Class | 2: benign<br>4: malignant | Nominal |

### H. Classification Algorithms

The classification phase learns the data set from previous step by using three different decision tree classification algorithms. There are Decision Tree (J48), REP Tree, and Random Tree.

### I. Performance Evaluation

In this section, three classifiers have been compared the accuracy results between data preprocessing with and without PCA. Among these two techniques, there will be a technique provide the minimal number of rules for breast cancer diagnosis.

The most effective approach to evaluate the performance of the model is based on the confusion matrix as shown in Table II. The confusion matrix is a specific table layout that shows the information about actual and predicted value by a classification model. There are four possible classification methods for each instance: a true positive (TP), a true negative (TN), a false positive (FP) and false negative (FN). Accuracy is computed by the equation (4)

TABLE II. CONFUSION MATRIX [15]

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Posituve | TP | FP |
| Actual Negative | FN | TN |

$$Accuarcy = \frac{TP + TN}{TP + FN + FP + TN} \quad (4)$$

## IV. EXPERIMENTAL RESULTS

The Table III shows the classification's accuracy of rule based system in breast cancer data set without applying PCA between three classification algorithms. For this implementation 9 features from the data set are used. The results show that J48 classifier can classify 96.04 % correctly while REP Tree classifier performed 95.31% and Random Tree classifier computed 94.14%.

TABLE III. THE PERFORMANCE CLASSIFICATION WITHOUT PCA

| ***Classifiers*** | ***J48*** | ***REPTree*** | ***Random Tree*** |
|---|---|---|---|
| Correctly classified | 656 | 651 | 643 |
| Incorrectly classified | 27 | 32 | 40 |
| Accuracy (%) | 96.04 | 95.31 | 94.14 |

From Table IV indicates the results of three classifiers with PCA. From the data set, there are only 7 features used for classification. It shows that J48 classifier proved to be the most accurate classifier for WBCD with the accuracy of 97.36% by 1.32 % increased. In addition, REP tree and Random tree classifiers got the improvement by 1.46 % and 0.58 %, respectively.

TABLE IV. THE PERFORMANCE CLASSIFICATION WITH PCA

| ***Classifiers*** | ***J48*** | ***REPTree*** | ***Random Tree*** |
|---|---|---|---|
| Correctly classified | 665 | 661 | 647 |
| Incorrectly classified | 18 | 22 | 36 |
| Accuracy (%) | 97.36 | 96.77 | 94.72 |

The bar chart gives the information about the comparison of rule based system of three proposed classifiers with and without PCA approaches as shown in Figure 2. According to the chart, the number of rules are decreased in case of the classifiers with PCA. Especially, J48 and REP Tree classifier performed significantly in producing the rules better than Random Tree classifiers. They provide only 2 rules while Random Tree classifier needs 30 rules. However, the number of rules still high when PCA is not used i.e. REP Tree, J48, and Random tree classifiers need 7, 11, 34 rules, respectively.

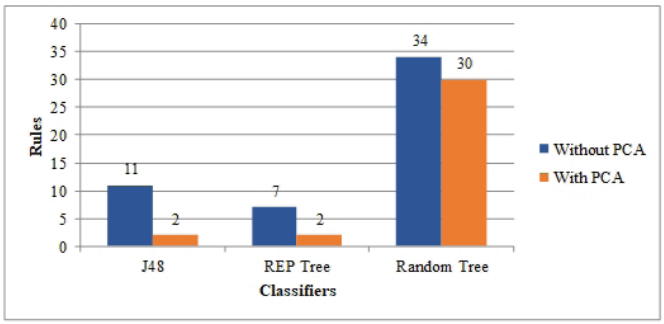


Fig. 2. Comparison performance of three classifiers

## V. CONCLUSION

A rule based system for breast cancer diagnosis has been a powerful tool supporting doctor diagnosis. Such a system requires classification rules derived from historical diagnosis. The desirable rules should be minimal in their number and give a good performance. This paper is to obtain such rules from the Wisconsin Breast Cancer data set. It performed experiments on the data set with PCA reduction to determine the best classifier among J48 decision tree, REP Tree, and Random Tree. It found that J48 classifier giving the best accuracy and smallest number of rules which are 97.36% and 2, respectively.

## ACKNOWLEDGMENT

The author would like to be grateful to International College, King Mongkut's Institute of Technology Lardkrabang for supporting and motivating during my research and thankful to Dr. William H. Wolberg at the University of Wisconsin for providing breast cancer data set which used in this paper.

## REFERENCES

[1] H. Elouedi, W. Meliani, and Z. Elouedi, "A hybrid approach based on decision trees and clustering for breast cancer classification," in 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2014.

[2] F.Kharbat, H.Ghalayini, "New algorithm for Building Ontology from Existing Rules: A Case Study," in International Conference on Information Management and Engineering, 2009, pp. 12-16.

[3] P. Hamsagayathri and P. Sampath, "Performance analysis of breast cancer classification using decision tree classifier," International Journal Of Current Pharmaceutical Research (IRCPR), vol. 9, 2017.

[4] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of breast cancer using decision tree data mining technique," International Journal of Computer Applications, vol. 98, no. 10, p. 0975 8887, 2014.

[5] C. P. Utomo, A. Kardina, and R. Yuliwulandari, "Breast cancer diagnosis using artificial neural networks with extreme learning techniques," International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 3, no. 7, pp. 10–14, 2014.

[6] S.Jhajharia, H.K.Varshney, S.Verma, and R.Kumar, "A neural network based breast cancer prognosis model with PCA processed features," in International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016.

[7] Kathija and S.Nisha,"Breast cancer Data Classification Using SVN and Naive Bayes Techniques," International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE), vol. 4, pp. 21 167–21 175, 2016.

[8] D.Lavanya and D. Rani, "Evaluation of Decision Tree Classifiers on Tumor Datasets," International Journal of Emerging Trends Technology in Computer Science (IJETTCS), vol. 2, pp. 418–423, 2013.

[9] T. M. Mohamed, "Efficient breast cancer detection using sequential feature selection technique," in 7th International Conference on Intelligent Computing and Information Systems (ICICIS), 2015.

[10] J.-W.Liu, Y.H.Chen, and C.H.Cheng, "Owa based information fusion method with PCA preprocessing for data classification," in International Conference on Machine Learning and Cybernetics, 2012, pp. 3322–3327.

[11] T. R. Patil and S. S. Sherekar, "Performance analysis of naive bayes and j48 classification algorithm for data classification," International Journal Of Computer Science And Applications, vol. 6, no. 2, 2013.

[12] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco,CA, USA, 2nd edition, 2005.

[13] A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software:. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[14] W.H.Wolberg, O.Mangasarian, and D.W.Aha.UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[15] K. M. Ting, Confusion Matrix. Springer US, 2010.