International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: http://dx.doi.org/10.7324/IJASRE.2018.32767

Volume 4, Issue 6
June - 2018

Comparative Study of Principal Component Analysis (PCA) based on Decision Tree Algorithms

Aung Nway Oo

University of Information Technology

Yangon Division

Myanmar

ABSTRACT

Data mining (DM) can be viewed as a result of the natural evolution of information technology. The role of data mining approach is very important in computer science and knowledge engineering. A number of data mining approaches are used for classification. Classification is the process of finding a model that describes and distinguishes data classes or concepts. The decision tree (DT) approach is most useful in the classification problem. The research work analyses the efficiency of the Principal Component Analysis (PCA) based decision tree algorithms, namely J48, Classification and Regression Tree (CART) and Random Forest.

Keywords: Data mining (DM), Classification, Decision Tree (DT), Principal component analysis (PCA).

1. INTRODUCTION

Data mining is a process to discover interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from large amounts of data stored in databases or other information repositories. In the procedure of data mining the former data is explained and future rules are calculated by data analysis. Data mining is a major advancement in the type of analytical tools. Data mining is a multi-disciplinary field which is a combination of machine learning, statistics, database technology and artificial intelligence.

In data mining, a decision tree (it may be also called Classification Tree) is a predictive model that can be used to represent the classification model. Classification trees are useful as an exploratory technique and are commonly used in many fields such as finance, marketing, medicine and engineering [7-10]. The use of decision trees is very popular in data mining due to its simplicity and transparency.

Feature selection is one of the key topics in data mining; it improves classification performance by searching for the subset of features. In problem of high dimensional feature space, some of the features may be redundant or irrelevant. Removing these redundant or irrelevant features is very important; hence they may deteriorate the performance of classifiers. Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features [11].

In this paper, we evaluate the ionosphere dataset and compare different decision tree algorithms, where PCA is applied for feature selection. The rest of this paper is organized as follows. We outline overview of decision tree algorithms in section 2 and PCA in Section 3. The experimental results and conclusions are presented in Section 4 and 5 respectively.

2 DECISION TREES (DT)

A decision tree is a flow chart like tree structure. The internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution. A decision tree consists of nodes. Each node represents some information. Decision tree learning is started from the root node and discrete values are produced at each node by testing the values of attribute [1]. These discrete values acts as target function. Then by using target function, value of attribute for next node is evaluated. This process is repeated for each new node. The learned tree is represented by if then rules. Decision tree algorithms can be applied on large amounts of data and valuable predictions can be produced.

2.1 J48 Algorithm

J48 is an extension of ID3 algorithm. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. J48 decision tree can handle specific characteristics, lost or missing attribute values of the data and differing attribute costs. Here precision can be increased by pruning [2]. The steps of the algorithm are as follows:

Step 1: The leaf is labeled with the same class if the instances belong to the same class.

Step 2: For every attribute, the potential information will be calculated and the gain in information will be taken from the test on the attribute.

Step 3: Finally the best attribute will be selected based on the current selection parameter.

2.2 Classification and Regression Tree (CART) Algorithm

The CART (Classification and regression trees) was jointly developed by Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone in 1984. It builds both classification and regression trees. The classification tree construction by CART is based on the binary splitting of the attributes. The algorithm will consider the set of samples-question about the data features will lead to the data minimization and continues till some stop criteria is reached [3]. It is also based on Hunt's model of decision tree construction, and can be implemented serially. It uses the gini index splitting measure in selecting the splitting attribute. Pruning is done in CART by using a portion of the training data set. The CART uses both numeric and categorical attributes for building the decision tree, and has in-built features that deal with missing attributes. The steps of the algorithm are as follows:

Step1: The first is how the splitting attribute is selected.

Step2: The second is deciding upon what stopping rules need to be in place.

Step3: The last is how nodes are assigned to classes.

2.3 Random Forest Algorithm

The first algorithm for random decision forests was created by Tin Kam Ho[4] using the random subspace method. A random forest algorithm is a supervised classification algorithm. As the name suggests, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives higher accuracy results. The steps of the algorithm are as follows:

Step1: The first is the creation of random forest.

Step2: The second is the prediction with a trained random forest classifier.

DOI: dx.doi.org/10.7324/IJASRE.2018.32767

3. PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is an essential technique in data compression and feature extraction [12]. It is well known that PCA has been widely used in data compression and feature selection. Feature selection refers to a process whereby a data space is transformed into a feature space, which has a reduced dimension. Overview of PCA is briefly described as follows.

Assume that $\{x_t\}$ where t =1, 2..., N are stochastic n dimensional input data records with mean (μ). It is defined by the following Equation:

$$\mu = \frac{1}{N}\sum_{t=1}^{N} x_t \quad (1)$$

The covariance matrix of x_t is defined by

$$C = \frac{1}{N}\sum_{t=1}^{N}(x_t - \mu)(x_t - \mu)^T \quad (2)$$

PCA solves the following eigenvalue problem of covariance matrix C:

$$C_{vi} = \lambda_i v_i \quad (3)$$

where λ_i (i =1 ,2,...,n) are the eigenvalues and v_i(i = 1,2,...,n) are the corresponding eigenvectors. To represent data records with low dimensional vectors, we only need to compute the m eigenvectors (called principal directions) corresponding to those m largest eigenvalues (m<n). It is well known that the variance of the projections of the input data onto the principal direction is greater than that of any other directions.

Let

$$\varphi = [v_1, v_2, \ldots, v_m],\ \Lambda = diag[\lambda_1, \lambda_2, \ldots, \lambda_m] \quad (4)$$

Then

$$C\Phi = \Phi\Lambda \quad (5)$$

The parameter *v* denotes to the approximation precision of the m largest eigenvectors so that the following relation holds.

$$\frac{\sum_{i=1}^{m}\lambda i}{\sum_{i=1}^{n}\lambda i} \geq v \quad (6)$$

Based on (5) and (6) the number of eigenvectors can be selected and given a precision parameter *v*, the low dimensional feature vector of a new input data x is determined by

$$x_f = \Phi^T x \quad (7)$$

4. EXPERIMENTAL RESULTS

The experimental results of classifiers are discussed in this section. The main aim of this research is to analyze PCA based decision tree classification algorithms. The ionosphere dataset from UCI is used for comparative analysis. The data set contains 351 instances and 35 attributes. The WEKA application is used for the evaluation. For each classifier 2/3 of the dataset is used for training and 1/3 of datasets is used for. A comparative study of classification accuracy in J48, CART and Random Forest algorithm is carried out in this work. The following formula is used to calculate the proportion of the predicted positive cases, Precision P using TP = True Positive Rate and FP = False Positive Rate as,

$$\text{Precision P} = \frac{TP}{TP+FP} \quad (8)$$

It has been defined that Recall or Sensitivity or True Positive Rate (TPR) means the proportion of positive cases that were correctly identified. It will be computed as

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

Where FN =False Negative Rate

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

The above formula will calculate the accuracy, which is the proportion of the total number of predictions that were correct. The following tables compare the accuracy of three different classifiers for without PCA.

Table 1. Accuracy of algorithms without using PCA

J48	CART	Random Forest
84.8739 %	84.0336 %	91.5966 %

Table 2. Results of J48 without using PCA

Class	TP Rate	FP Rate	Precision	Recall
Bad	0.774	0.091	0.872	0.774
Good	0.909	0.226	0.833	0.909

Table 3. Results of CART without using PCA

Class	TP Rate	FP Rate	Precision	Recall
Bad	0.792	0.121	0.840	0.792
Good	0.879	0.208	0.841	0.879

Table 4. Results of Random Forest without using PCA

Class	TP Rate	FP Rate	Precision	Recall
Bad	0.868	0.045	0.939	0.868
Good	0.955	0.132	0.900	0.955

The experimental results of PCA based decision tree classifiers are shown in the following tables.

Table 5. Accuracy of algorithms using PCA

PCA-J48	PCA-CART	PCA-Random Forest
88.2353 %	89.0756 %	94.1176 %

Table 6. Results of J48 using PCA

Class	TP Rate	FP Rate	Precision	Recall
Bad	0.849	0.091	0.882	0.849
Good	0.909	0.151	0.882	0.909

Table 7. Results of CART using PCA

Class	TP Rate	FP Rate	Precision	Recall
Bad	0.849	0.076	0.900	0.849
Good	0.924	0.151	0.884	0.924

Table 8. Results of Random Forest using PCA

Class	TP Rate	FP Rate	Precision	Recall
Bad	0.925	0.045	0.942	0.925
Good	0.955	0.075	0.940	0.955

The following figure visualizes the accuracy results of different classifiers.

DOI: dx.doi.org/10.7324/IJASRE.2018.32767

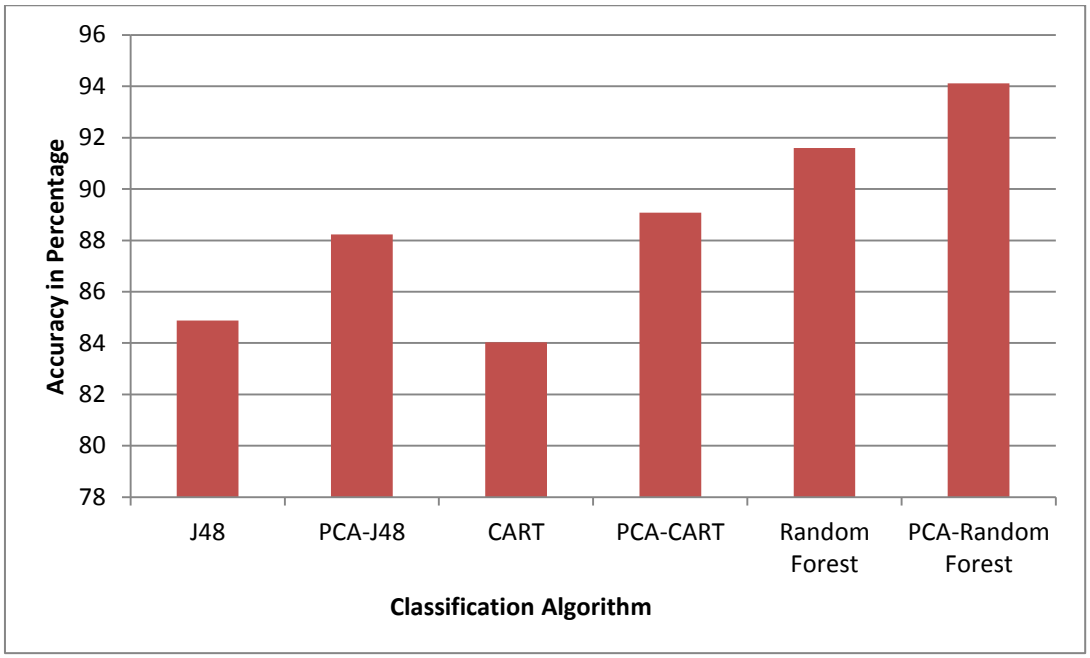

Figure 1. Accuracy results of different classifiers.

5. CONCLUSION

The paper presented comparative results of J48, CART and Random Forest decision tree classifier with PCA. This analysis provides as a better understanding of these algorithms. The experimental results shows that the highest accuracy 94.1176 % is found in PCA-Random Forest classifier and the results of every PCA based algorithms is higher than other normal decision tree algorithms for choosing data set. We can apply them on different types of data sets and can attain a best result by knowing that which algorithm will give the best result on a given type of data set.

REFERENCES

[1] Aloraini A. Different machine learning algorithms for breast cancer diagnosis. International Journal of Artificial Intelligence and Applications, 3(6):21–30, 2012.

[2] Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22):13–7, 2014.

[3] Wang KJ, Adrian AM. Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm. International Journal of Computer Science and Electronics Engineering, pp. 408–12, 2013.

[4] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, pp. 278–282, 1995.

[6] Jiawei han and micheline kamber, Data mining concepts and techniques, second edition,285-291.

[7] H.S. OH and W.S. SEO, , "Development of a Decision Tree Analysis model that predicts recovery from acute brain injury", Japan Journal of Nursing Science, pp. 1742-7924, 2012.

[8] G. Zhou and L. Wang, , "Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation", Transportation Research: Part C,21(1), pp. 287-305., 2011.

[9] S. Sohn and J. Kim, . "Decision tree-based technology credit scoring for start-up firms: Korean case ",Expert ystems With Applications,vol. 39(4), pp. 4007-4012,2011.

[10] J. Choand P.U. Kurup, "Decision tree approach for classification and dimensionality reduction of electronic nose data",Sensors & Actuators B: Chemical,vol. , pp. 542-548, 2011.

[11] D. Koller, and M. Sahami, "Toward optimal feature selection" In proceedings of international conference on machine learning, Bari, (Italy) pp. 284-92, 1996.

[12] E. Oja, "Principal components, minor components, and linear neural networks" Neural Networks, vol. 5, pp. 927-935, 1992.