PCA based feature reduction to improve the accuracy of decision tree c4.5 classification

M Z F Nasution¹, O S Sitompul²* and M Ramli³

¹Graduate School of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
²Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
³Department of Mathematics, Syiah Kuala University, Aceh, Indonesia

*Corresponding author: opim@usu.ac.id

Abstract. Splitting attribute is a major process in Decision Tree C4.5 classification. However, this process does not give a significant impact on the establishment of the decision tree in terms of removing irrelevant features. It is a major problem in decision tree classification process called over-fitting resulting from noisy data and irrelevant features. In turn, over-fitting creates misclassification and data imbalance. Many algorithms have been proposed to overcome misclassification and overfitting on classifications Decision Tree C4.5. Feature reduction is one of important issues in classification model which is intended to remove irrelevant data in order to improve accuracy. The feature reduction framework is used to simplify high dimensional data to low dimensional data with non-correlated attributes. In this research, we proposed a framework for selecting relevant and non-correlated feature subsets. We consider principal component analysis (PCA) for feature reduction to perform non-correlated feature selection and Decision Tree C4.5 algorithm for the classification. From the experiments conducted using available data sets from UCI Cervical cancer data set repository with 858 instances and 36 attributes, we evaluated the performance of our framework based on accuracy, specificity and precision. Experimental results show that our proposed framework is robust to enhance classification accuracy with 90.70% accuracy rates.

1. Introduction

C4.5 algorithm is a decision tree classification approach based on information entropy. The algorithm uses modified splitting criterion called gain ratio [1]. On decision tree approach, pruning is a step to eliminate several branches which actually does not required [2] and create a large scale of decision tree process called over-fitting. Over-fitting for cost-sensitive learning method produces classification model by improving accuracy in training data. Nevertheless, it does not improved classification model when applied to unseen data in real data set [3]. Over-fitting caused by noisy-data and irrelevant feature [4] and creates misclassification resulting in overfitting and data imbalance. In turn, overfitting reduces accuracy of decision tree in classification model [5]. In order to reduce high data dimensions, a common approach is to use feature reduction to obtain a lower dimensional data, which depends on the feature criteria of the problem setting [6]. Feature reduction is followed by feature selection by correlation and gain ratio methods. Unsupervised method minimizes the information loss whereas the supervised reduction methods maximize the class discrimination [7].
In this research, we propose a framework for feature reduction using principal component analysis (PCA) to eliminate irrelevant features, selecting relevant and non-correlated features without affecting the information contained in the original data and then predictor is developed using decision tree C4.5 algorithm to classify data set used. The aim is to reduce some features of data by PCA and then build a prediction of classification model by decision tree C4.5 to obtain relevant features and to improve the accuracy of decision tree C4.5. To evaluate the proposed approach, we use cervical cancer data set from the UCI Machine Learning repository.

2. Principal Component Analysis (PCA)

PCA gives good result when applied to correlated attributes. In this research, PCA is applied in both training and testing attributes of the cervical cancer data set. The PCA will identify patterns in the data set, and finds their similarities and differences between each attributes. It acts as a powerful model to analyse the data. The original data and the average of the original data are chosen. The covariance matrix is computed whereby the result is used in calculating the eigen vectors and eigen values [8] and the eigen vector with the highest eigen value is chosen as the principle component of the cervical cancer data set as it exhibits the most significant relationship between the data set attributes. The Eigen values are sorted in ascending order to choose the most significant data and discard the least significant one. By this means data with higher dimensions is reduced to lower dimensions [9].

The variance is calculated to find the spread of data in the UCI Cervical cancer data set using equation (1) to determine the deviation of data in the sample data set [10].

\[
\text{Var}(x) = \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \mu_j)^2
\] (1)

Afterwards, covariance is calculated to find the relation between two classes, in which zero value indicates that there is no relation between two dimensions. The covariance is computed using (2) as suggested in [11].

\[
\text{Cov} (x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \mu_{xj}) (y_{ij} - \mu_{yj})
\] (2)

Finally, the Eigenvalues and Eigenvectors for the covariance matrix are calculated. The computed eigenvalues is then transformed (varimax orthogonal rotation) using equation (3) [11].

\[
\text{Det} (A - \lambda I) = 0
\] (3)

3. C4.5 Algorithm

The next step is classifying attributes based on information gain ratio using decision tree C4.5. It is a measure on the number of bits of information obtained for the class prediction. The expected information needed to classify a given sample data set is given by [1].

\[
\text{Entropy} (S) = -(+) \log 2 P(+) - P(-) \log 2 P(-)
\] (4)

Where P is the probability of sample. Information Gain is computed using equation (5) as suggested in [4].

\[
\text{Gain} (S,A) = \text{Entropy} (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy} (S_i)
\] (5)

The entropy of the expected information based on the splitting of subset to data set is given by (6)

\[
\text{SplitInfo}_A(D) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)
\] (6)

and the gain ratio is computed using (7) [4].

\[
\text{Gain Ratio} (A) = \frac{\text{Gain} (A)}{\text{SplitInfo} (A)}
\] (7)
4. Previous Research
An hybrid PCA feature reduction and selection subset by using wrapper filter and information gain-ratio decision tree C4.5 had been proposed in [9]. Feature reduction resulted by the hybrid model did not calculate the performance accuracy of classification decision tree C4.5. A calculation on the classification performance accuracy was proposed by [12] using a prediction system to develop business processes in predictive diagnosis of hypertension by adding functionality prognosis. The prediction is conducted on health information of hypertension data set using decision tree C4.5 algorithms. A 93% output of classification performance accuracy is obtained for the level of health risk patients.

Research work by [13] proposed orthogonal local preserving projection (OLPP) and artificial neural network (ANN) classifier. Here, OLPP was used to reduce feature dimensions. Once the feature reduction is formed, the prediction was subsequently performed based on the classifier. The authors claimed a classification performance accuracy of 90% was achieved. Furthermore, an approach for efficient imputation and classification using imputation based on class based clustering (IMCBC) was developed in [5]. The classification accuracies calculated using the proposed imputation approach was deployed on the existing classifiers such as ID3, J48, Naïve Bayes, KNN and SVM. Naïve Bayes had a significant accuracy of 80%.

In addition, by sharing the same coefficient signs, research work by [14] used a regularization-based transfer learning strategy to enhance both source and target models. Cross-modality individual risk and cross-expert subjective was used to asses colposcopy cervical cancer images quality. A combination of logistic regression and SVM were claimed to obtained positive positive results. Another classification model approach based on cost-selectivity was proposed in [15]. Pre-processing was performed by calculating correlation value for each attributes and target attributes on UCI Cervical cancer data set. In the final step, features reduction method eliminates the highest values attributes correlation. The results achieved a true positive rate (TP rate) by 0.429.

Finally, [16] conducted a principal component analysis and artificial neural network (ANN) classifier to perform non-correlated feature selection. The analysis was performed on traumatic brain injury (TBI) patients, including intracranial hypertension. The predictivity of the components were tested and the results showed a promising accuracy with a mean absolute error of 0.025.

5. The Proposed Method
In this proposed framework, feature reduction method using PCA is conducted as an initial step towards reducing the number of attributes without losing the main purpose and objective information from the original data. The next step is developing a predictor with an improved accuracy in order to classify data set. With the objective to find the decision attributes that have more dependent information from data, the proposed framework consists of the several phases.

- Data acquisition. An experimental data set on Cervical cancer from UCI Machine Learning repository was used to predict cervical cancer indicators and diagnosis.
- Data pre-processing. New attributes should be constructed in order to replace missing values and inconsistencies. Hinselman, schiller, cytology and biopsy are among the targets that are reduced into two values target, namely yes and no when the value of constructed target is 1 and 0, respectively. Missing values of numerical attributes are replaced with average amount of the subset, whereas missing values of nominal attributes are replaced with the most commonly occurring value of the subset.
- Standardizing each attributes. Calculated using variance and standard deviation formula.
- Calculate correlation of each attribute based on covariance formula.
- Calculate eigenvalues of each attributes to produce principal components (PC).
- Find principal component by proportion of variance, these collectively account for 95% of principal component (PC) from eigenvalues.
- Calculate eigenvectors. Performed by transposing and multiplying matrices. Each element of the eigenvectors represents the contribution of a given attributes to principal component (PC).
- Find new attributes based on correlations between the original attributes and the principal component, which is called loading. The loading considered as the cut-off is 0.5.
- Building classification model of decision tree C4.5 classifier. A number of new attributes are classified and evaluated based on performance accuracy, particularity and precision by dividing data into 90% and 10% for training and testing, respectively.

6. Result and Discussion
Cervical cancer data set from UCI Machine Learning repository contains 858 instances and 36 attributes including 4 medical test attributes constituted as the target of this data set. The proposed framework is implemented in Rapid Miner® version 5.3. The detail description of each attribute as described in the UCI repository before attribute reduction using PCA is shown in table 1.

Table 1. Cervical cancer data set description before attribute reduction

| Attributes                  | Type         | Attributes                          | Type         | Attributes                          | Type         |
|-----------------------------|--------------|-------------------------------------|--------------|-------------------------------------|--------------|
| Age                         | Int          | Hormonal Contraceptives (years)     | Int          | STDs: vulvo-perineal condylomatosis | Bool         |
| Number of sexual partners   | Int          | Intra Uterine Device (IUD)          | Bool         | STDs: syphilis                      | Bool         |
| Age of first sexual Intercourse | Int       | IUD (years)                         | Int          | STDs: pelvic inflammatory disease   | Bool         |
| Number of pregnancies       | Int          | Sexually Transmitted Diseases (STDs)| Bool         | STDs: genital herpes                | Bool         |
| Smokes                      | Bool         | STDs (number)                       | Int          | STDs: molluscum contagiosum         | Bool         |
| Smokes (years)              | Bool         | STDs: condylomatosis                | Bool         | STDs: AIDS                          | Bool         |
| Smokes (packs/year)         | Bool         | STDs: cervical condylomatosis       | Bool         | STDs: HIV                           | Bool         |
| Hormonal Contraceptives     | Bool         | STDs: vaginal condylomatosis        | Bool         | STDs: Hepatitis B                   | Bool         |

In order to select relevant attributes, a number of principal components (PC1 to PC12) is performed, which contribute eigenvalues of 99.80% and absolute value of eigenvectors above 0.5. The PCA has reduced 36 attributes to 12 relevant attributes. The detail of each attribute after attribute reduction using PCA is shown in table 2.

Table 2. Result of attributes reduction on cervical cancer data set

| Attributes         | Eigenvector | Attributes                  | eigenvector |
|--------------------|-------------|-----------------------------|-------------|
| Age                | 0.969       | Hormonal Contraceptives     | 0.906       |
| Number of sexual partners | 0.93       | Hormonal Contraceptives (years) | 0.966       |
| First sexual Intercourse | 0.912      | IUD (years)                 | 0.953       |
| Number of pregnancies | 0.973      | STDs (number)               | 0.712       |
| Smokes (years)     | 0.886       | STDs: time since first diagnosis | 0.715       |
| Smokes (packs/year)| 0.889       | STDs: time since last diagnosis | 0.695       |

The new set of attributes is then classified using decision tree C4.5 classifier, which in turns are evaluated based on performance accuracy of the classifier. The data set is divided into 90% of data training and 10% of data testing. The accuracy, particularity and precision obtained can be seen in table 3. The accuracy was enhanced to a negligible margin from 86.05% to 90.70%, whereas for the same data set the number of attributes are decreased from 36 to 12 attributes. The comparison performance of both C4.5-Non-PCA and PCA+C4.5, shows in table 3.
As seen in Table 3, the evaluation metrics are analyzed for two machine learning models, by which we obtained the objective of the proposed cervical cancer data set classification system. The performance of model C4.5-Non-PCA based on accuracy, particularity, and precision values are 86.05%, 94.74% and 33.33%, respectively, whereas the performance of model C4.5+PCA based on accuracy, particularity and precision values are 90.70%, 100% and 100%.

7. Conclusion
The proposed framework has achieved a robust feature reduction approach based on PCA to enhance the accuracy of decision tree C4.5 classifier. The proposed framework was implemented in Rapid Miner® version 5.3. The experiments are conducted using the available data set from UCI Cervical cancer data set repository, which are evaluated based on the accuracy performance by applying feature reduction using PCA to reduce attributes and then classifying attributes using decision tree C4.5. The results obtained based on performance accuracy shows a value as high as 90.70%. Between C4.5-Non-PCA and PCA+C4.5 there are differences in precision as high as 4.65% accuracy, 5.26% particularity and 66.67%, respectively. This result concludes that feature reduction is an important issue in the classification model.

References
[1] John D K, Brian M N, Aoife D 2015 Fundamentals of machine learning for predictive data analytics (Cambridge: The MIT Press) chapter 4 pp 128-190
[2] Abellan J and Castellano J 2016. A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications 73 1-10
[3] Zhang S 2012 Decision tree classifiers sensitive to heterogeneous costs Journal of Systems and Software 85(4) pp 771–779
[4] Tao Wang, Zhenxing Q, Shicao Z and Chengqi Z 2012 Cost-sensitive classification with inadequate labeled data Information Systems 37 (2012) pp 508–516
[5] UshaRani Y and Sammulal P 2016 An Efficient Disease Prediction and Classification Using Feature Reduction Based Imputation Technique IEEE Conference on Engineering & MIS (ICEMIS) Morocco pp 1-5
[6] Ali H S, Guftar M, Qamar U and Muzaffar W A 2015 A feature reduction framework based on rough set for biomedical data sets IEEE International Conference on SAI Intelligent Systems Conference (IntelliSys) London pp 343-9
[7] Sivapriya T R and Nadira B K AR 2013 Hybrid feature reduction and selection for enhanced classification of high dimensional medical data IEEE Int. Conf. on Computational Intelligence and Computing Research (ICCIC) pp 1-4
[8] Kotu V and Deshpande B 2015 Predictive Analytics and Data Mining (Waltham: Morgan Kaufmann)
[9] Kavitha R and Kannan E 2016 An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining IEEE Int. Conf. on Emerging Trends in Engineering Technology and Science (ICETETS) pp 1-5
[10] Jolliffe I T 2002 Principal Component Analysis 2nd Ed. (New York: Springer-Verlag)
[11] Johnson R A and Wichern D W 2007 Applied Multivariate Statistical Analysis 6th Ed. (New Jersey: Pearson Prentice Hall)
[12] Sandi G, Supangkat H S and Slamet C 2016 Health risk prediction for treatment of hypertension IEEE Int. Conf. on Cyber and IT Service Management (CITSM) pp 1-6
[13] Tarle B and Jena S 2016 Improved artificial neural network for dimension reduction in medical
data classification IEEE Int. Conf. on Computing Communication Control and Automation (ICCUBEA) pp 1-6

[14] Fernandes K, Cardoso J S and Fernandes J 2017 Transfer Learning with Partial Observability Applied to Cervical Cancer Screening Iberian Conference on Pattern Recognition & Image Analysis pp. 243-250

[15] Fatlawi H K 2017 Enhanced classification model for cervical cancer data set based on cost sensitive classifier Int. J. of Computer Techniques 4(4) pp. 115-120

[16] Naraei P and Sadeghian A 2017 A PCA based feature reduction in intracranial hypertension analysis IEEE Int. Conf. on 30th Canadian Conference on Electrical and Computer Engineering (CCECE) pp 1-6

**Acknowledgement**

High appreciation should be given to Lembaga Penelitian Universitas Sumatera Utara (LP USU), Graduate School of Computer Science Fasilkom-TI USU and Rector of Universitas Sumatera Utara for their support on the dissemination of this research work as well as the facilities provided.