DECISION SUPPORT SYSTEM FOR CLASSIFICATION OF CHRONIC KIDNEY DISEASE WITH PRINCIPLE COMPONENT ANALYSIS

A. K. Shrivas, Dr. C. V. Raman University, Bilaspur (C.G.), India (akhilesh.mca29@gmail.com)
Sanat Kumar Sahu, Govt. Kaktiya P.G. College, Jagdalpur (C.G.), India (sanat.kosa1@gmail.com)
S. K. Singhai, Govt. Engineering College, Bilaspur (C.G.), India (singhai_sanjay@yahoo.com)

ABSTRACT

Now a day, Decision Support System (DSS) is widely used in medical science for diagnosis of disease both for doctors and medical students. Data is increasing in every day in every organization due to increasing number of patients. Data mining based decision tree techniques play very important role to identification and classification of diseases. In this research work we have used data mining based decision tree classifiers to develop the robust classifier for classification of chronic kidney disease. We have used four decision tree techniques and its ensemble model to compare the classification accuracy. The proposed ensemble of Random Forest (RF), Classification and Regression Technique (CART), C5.0 tree and J48 as robust classifier and gives 99% of accuracy. An ensemble model is combination of two or more than two classifiers which give better classification accuracy compare to individuals classifiers. We have also used Principle Component Analysis (PCA) for dimension reduction to computationally increase the performance of model. PCA is applied on proposed ensemble model with different feature components and achieved satisfactory results as 99.75% with less computational time.

Keywords: Classification, Ensemble Model, Principal Component Analysis (PCA), Decision Support System (DSS), Chronic Kidney Disease (CKD).

INTRODUCTION

Now a day in field medical science, doctors and medical students are facing the problem of various diseases like blood pressure, sugar, heart disease, kidney disease etc. In which kidney disease is very serious problem that is facing by most of the people. Chronic diseases usually cannot be prevented by vaccines or cured by medication, nor do they just disappear. Classification is very important technique to identify and classification of various chronic diseases. Classification of chronic kidney disease data will be beneficial to doctors, pharmacists, and medical. Decision support system (DSS) is very important role in field of computer science for decision making process. In this study, we have used the four decision tree based classifiers like C4.5, C5.0, Random forest and classification and regression tree (CART) and their ensemble models as decision support system for construct rule and make decision process. These decision tree models are used to classify the CKD data and compare the performance of each individual's model and ensemble model. Feature selection play major role to remove the irreverent feature from original feature space and improve the performance of model.

A lot of researchers have worked in the field of medical science which related to various chronic diseases like cancer, heart, and other diseases diagnosis. There are some authors have studied about CKD. S. Ramya and D.N.R. (2016) have used three different classification techniques such as Back Propagation Neural Network, Radial Basis Function and Random Forest for classification of CKD. Radial basis function network gives the highest accuracy as 85.3%. Arora & Sharma (2016) have worked on three classification algorithms i.e. Naïve Bayes, J48 and SMO with WEKA environment and compared the performance of accuracy. J48 classifier performs the best classification accuracy compare to others. Kumar (2016) used Random Forest (RF) classifiers, Sequential Minimal Optimization (SMO), Naive Bayes, Radial Basis Function (RBF) and Multilayer Perceptron (MLP), Simple Logistic (SLG) techniques for the predictions task of CKD. The Random forest performs better than other classifiers. Sinha (2015) suggested two classification techniques as support vector machine (SVM) and K-Nearest Neighbour (KNN).The performance of KNN classifier is better than SVM. S & S (2015) have used SVM and Bayes network as classifier for classification of chronic kidney disease and compare the performance of classifier where SVM gives high performance compare to Naive Bayes Classifiers.

PROPOSED DECISION SUPPORT SYSTEM

Healthcare diagnosis system is capable to identify and solve problems (i.e. health issues) and make decisions. Decision Support System (DSS) is simple means of an interactive computer-based system for uses by medical students, physician, and medical researchers focus on understanding and improving the decision process. In this study, we have proposed a DSS for classification of CKD. The main goal of DSS is to improve the performance and identify the diagnosis. Our proposed DSS is Health Care Diagnosis Systems (HCDS), through a variety of data mining techniques being applied to support physicians in diagnosing, have received great attention due to the success of the novel system.

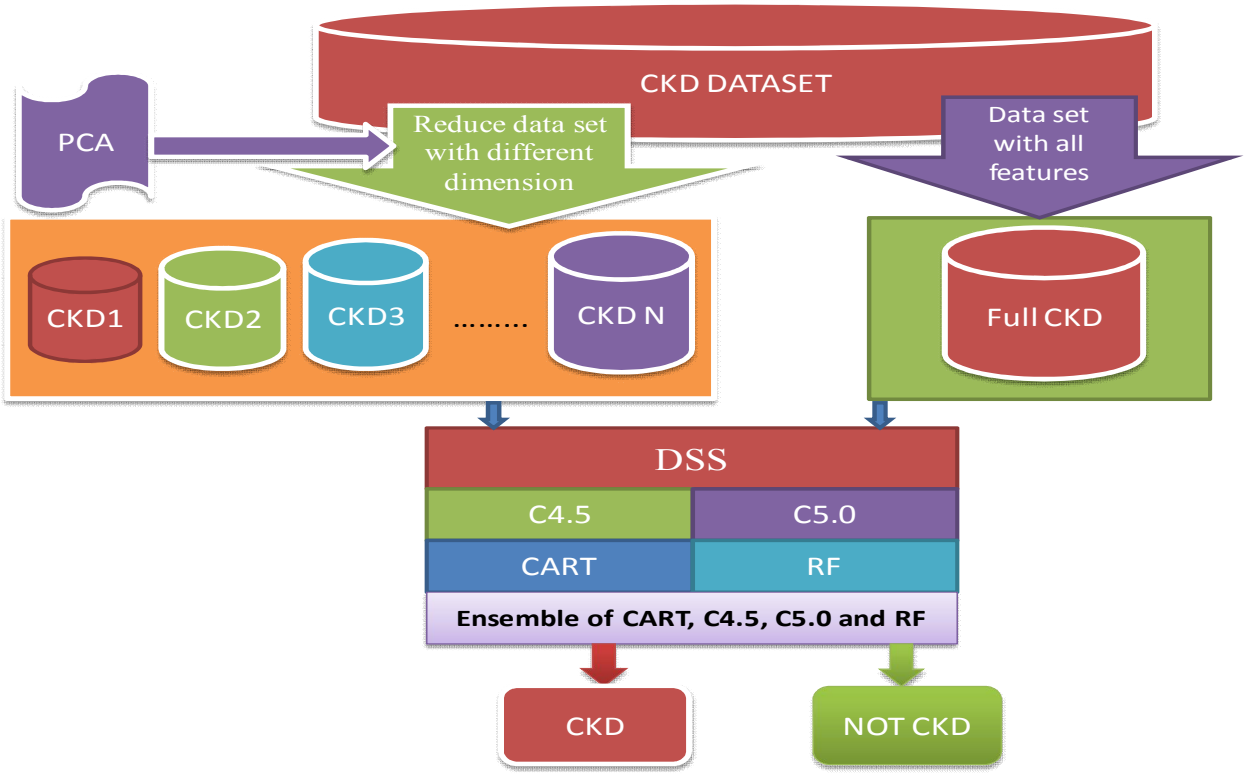

Figure1: Proposed Decision Support System for CKD

Our proposed DSS based on classification of binary class CKD datasets collected from UCI repository. DSS is used to classify the CKD data set into CKD and NOTCKD with all features and reduced dimension of data set as shown in Figure1 where CKD1, CKD2 CKD 3 etc. are reduced dimension data set. In this research work we have used different decision tree as decision support system with feature dimension reduction technique. We have used Random Forest (RF), C4.5, C5.0 and CART as classifiers and their ensemble model to develop robust DSS. We have used PCA in dimensionality reduction is accomplished by selected sufficient eigenvectors to the explanation for a quantity of percentage of the variance in the original data. PCA is a feature optimization technique applied on decision tree to develop computation robust model.

DATASET AND K-FOLD VALIDATION

This research work focus on the classification of CKD. The CKD data set is collected from UCI repository having 24 features, 400 instances and 1 class with binary nature (ckd or notckd).The features of data set contain numerical and nominal value. K-fold cross validation is technique for data partition as training and testing. In this research work we have used 10-fold cross validation for partition of data into training and testing.
The process of 10-fold cross validation as given below:

i. Split sample dataset into 10 equal sizes.
ii. Used train for 9 sample of data set and test for 1 sample size.
iii. Repeat 10 times of steps 2 until each partition used as testing and obtain the mean accuracy.

DECISION TREE

Decision tree (Han, Kamber, & Pei, 2012) is the knowledge of decision trees starting class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each interior node (nonleaf node) indicates a test on an attribute, each one limb corresponds to an outcome of the test, and each leaf node (or terminal node) holds a class label. The uppermost node in a tree is the root node.

(i) C4.5

C4.5 (Han et al., 2012) is an algorithm used to produce a decision tree developed by Ross Quinlan. C4.5 is implementing a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. The decision trees (Pujari, 2013) produced by C4.5 be capable of use for classification, and for this rationale, C4.5 is often referred to as a statistical classifier.

(ii) C5.0

This is a decision tree supported classifier developed by Ross Quinlan and is an extension of C4.5. It without human intervention extracts classification rules in the form of decision tree from specified training data. C5.0 has several benefits over C4.5 in provisions of time and memory space required; the tree generated by C5.0 is also very small as compared to C4.5 algorithm which ultimately improves the classification accuracy(Han et al., 2012).

(iii) Classification and Regression Technique (CART)

CART is a classification technique is foundation on decision tree induction (Han et al., 2012) which is a learning of decision trees from class label training tuples. The Classification and Regression (CART) tree scheme make use of recursive partitioning to split the training records into fragment with related output field values. The CART tree node beginning by investigative the input fields to find the finest split, measured by the reduction in an impurity index that results from the split. CART bring into play Gini index splitting records measures in selecting the splitting attribute. Pruning is done in CART by using a training data set. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary (only two subgroups).

(iv) Random Forest

Random forests (Han et al., 2012; Pujari, 2013) are an ensemble learning method widely used for classification and regression tasks. For each tree in the forest, a bootstrap sample is selected from the original data. The bootstrapped sample is obtained by randomly selecting instances from the original data with replacement and is of the same size as the original data set. A decision tree is then grown to the maximum extent possible without pruning on the bootstrapped sample using a modified decision-tree learning algorithm. The tree-learning algorithm is modified as follows: At each node, best split is selected by examining a random subset of features rather than the complete feature set. Since deciding the best-split is the most computationally expensive aspect of the learning process, choosing a subset of features will drastically speed up the learning of the tree. Once all the trees are constructed this way, final predictions are obtained by averaging individual predictions of the trees.

ENSEMBLE MODEL

An ensemble (Han et al., 2012)of model is combination of two or more trained model Each combines a series of k learned models (classifiers or predictors), M_1, M_2, ::: , Mk and creating a composite model, M*. The main aim of ensemble the model is achieve the more classification accuracy compare to individuals model. In this research work we have used four classification techniques like RF, J48, C5.0 and CART are combined together to form ensemble model.

PRINCIPAL COMPONENT ANALYSIS (PCA)

It is a mathematical tool from applied linear algebra. It is a simple, non-parametric method of extracting relevant information from confusing datasets. Basics of statistical measures e.g. variance and covariance(Han et al., 2012).This method that applies an orthogonal transformation to convert a set of explanation of probably related variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is a less than or equal to the smaller of the number of original variables or the number of explanation. This change is defined in such a way that the first principal component has the biggest possible variance (that is, accounts for as much of the variability in the data as possible), and each subsequent component in rotate has the highest variance possible under the constraint that it is orthogonal to the earlier components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables. The plan of PCA is to decrease the dimensionality of dataset that contains a large number of correlated attributes by transforming the original attributes space to a new space in which attributes are uncorrelated. The algorithm then ranks the variation between the original dataset and the new one. Transformed attributes with most variations are saved; meanwhile discard the rest of attributes (Yildirim, 2015)**.** The steps of PCA as given below:

i. Eigen decomposition - Computing Eigenvectors and Eigen values
ii. Selecting Principal Components
iii. Projection onto the New Feature Space

RESULT AND DISCUSSION

This research work is carried out using R3.2.5 and WEKA 3.6 open source data mining tools for classification and dimension reduction respectively. The experimental work is done into two sections: first section consist analysis and development of robust classifier for classification of CKD while second section consist dimension reduction of CKD data set to improve the performance of model. We have used 10-fold cross validation for data partition as training and testing.

In first section, we have used various decision tree techniques like J48, C5.0, CART and Random forest to analysis and classification of CKD. We have proposed new ensemble model which is combination of decision trees due to individual models are not giving satisfactory results. The proposed ensemble model gives better results compare to individual decision tree techniques.Table1 shows that accuracy of individuals and ensemble model with 10-fold cross validation. The proposed ensemble model gives 99.0% of accuracy. Figure 2 shows that accuracy of individual and ensemble classifier.

In second section, we have used PCA as dimension reduction technique to develop computationally efficient model. We have applied the PCA techniques on CKD data set to reduce the dimension where reduce the dimension or component using reduce the variance covered value. We have applied the reduce dimension of CKD data set into proposed ensemble model. Table 2 shows that accuracy of proposed ensemble model with different reduce dimension or component. The proposed ensemble classifier gives satisfactory result with all the reduce component but achieved better result as 99.75% of accuracy with 4 component. Finally the proposed ensemble model with PCA is satisfactory DSS for classification of CKD.

Table 1: Accuracy of models with 10-fold cross validation

Models	**Accuracy**
CART	97.25%
J48	96.75%
C5.0	96.75%
Random Forest	98.25%
RF+J48+ C5.0+CART	**99.00%**

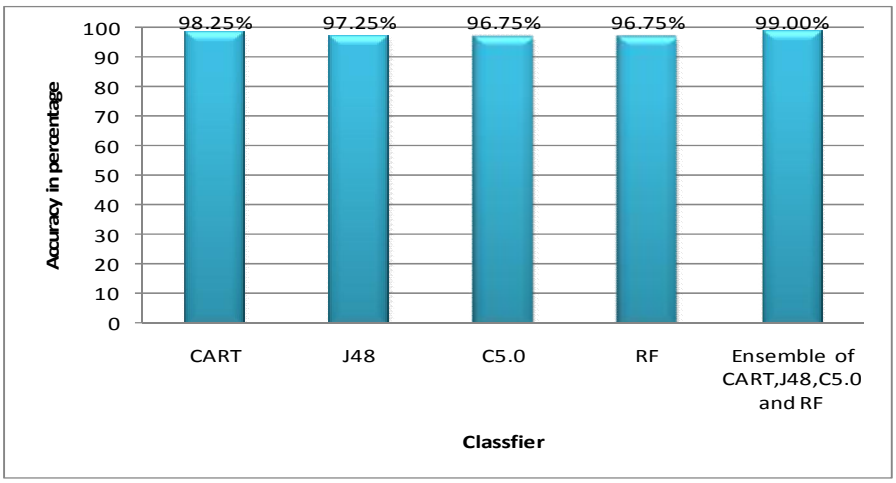

Figure 2: Accuracy of individual and ensemble classifier

Table 3: PCA applied on best model (RF+J48+ C5.0+CART)		
Variance covered	**Components**	**Accuracy**
95%	20	99.75%
90%	16	99.75%
85%	14	99.75%
80%	12	99.00%
75%	11	99.50%
70%	09	99.50%
65%	08	99.75%
60%	07	99.75%
55%	06	99.50%
50%	05	99.75%
45%	04	**99.75**%

CONCLUSION

Diagnosis of heath condition is very challenging and critical issue in field of medical science. DSS play very important role to diagnosis of health condition. We have developed robust DSS using decision tree techniques for classification of CKD. We have proposed ensemble model which is combination of J48, C5.0, CART and Random forest. The proposed ensemble model is gives better classification accuracy. We have used PCA as dimension reduction technique to reduce the dimension of data and computationally increase the performance of proposed model. Finally we recommended our proposed ensemble model with PCA is robust and computationally efficient model for classification of CKD.

REFERENCES

Arora, M., & Sharma, E. A. (2016). Chronic Kidney Disease Detection by Analyzing Medical Datasets in Weka. *International Journal of Computer Application*, *6*(4), 20–26.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques* (Third ed.), Elsevier.

Kumar, M. (2016). Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. *International Journal of Computer Science and Mobile Computing*, *5*(2), 24–33.

Pujari, A. (2013). *Data mining technique (Second ed.)*,University press.

Ramya, S., & D. N. R. (2016). Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(1), 812–820.

S, V. , & S., D. (2015). Data Mining Classification Algorithms for Kidney Disease Prediction. *International Journal on Cybernetics & Informatics*, *4*(4), 13–25.

Sinha, P., & Sinha, P. (2015). Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM, *International Journal of Engineering Research & Technology (IJERT)*, *4*(12), 608–612.

UCI Machine Learning Repository of machine learning databases. Retrieved from *http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease* (Last access date: July 2016).

Yildirim, P. (2015). Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. *International Journal of Machine Learning and Computing*, *5*(4), 258–263.