PERFORMANCE EVALUATION OF CLASSIFIERS FOR CLASSIFICATION OF CHRONIC KIDNEY DISEASE

A.K. Shrivas, Dr. C. V. Raman University, Bilaspur (C.G.), (akhilesh.mca29@gmail.com)
Pallavi Ambastha, Holy Cross Women's College, Ambikapur (C.G.), (pallavi.ambastha26@gmail.com)
Laxmi Gupti, Dr. C. V. Raman University, Bilaspur (C.G.), (pinkyroy8407@gmail.com)

ABSTRACT

In our day to day life we see in our surroundings many persons suffering from many sever health diseases in which chronic kidney disease is one of them. The bean-shaped pair organs i.e. Kidney helps to purify our blood by removing the waste product from our body in form of urines. This paper aims to classify the data as person suffering from chronic kidney disease or not with better accuracy. Various classifiers such as Decision Tree, Random Forest, and Random Tree, Naïve Bayes, K-NN, ID3 and CHAID have used individually with different data partition and calculate the performance in terms of accuracy. We have achieved best accuracy 97% in case of ID3 classifier. We have also applied the feature optimization technique with best ID3 classifier to achieve better classification accuracy. The suggested ID3 classifier with Weighted Information Gain feature selection technique gives i.e. 97% of accuracy with reduced 20 feature subset.

INTRODUCTION

In living body, disease is any circumstance that affects the way of living and working. Today we see our surrounding with full of patients suffering from several major or minor types of diseases like heart disease, lungs disease, blood disease etc . In this research work we have focused on chronic kidney disease. Kidneys, pair of organs located at lower back of our body. They help in filtering our blood and remove waste products or toxins from our body. These toxins are then send to our bladder and later removed through urination. When these normal functioning of body gets affected disease occurs. Chronic kidney disease may not become straightforward until our kidney function is knowingly undersized [www.healthline.com]. This paper contains several data mining techniques like decision tree, random forest, random tree, naïve Bayes, K-NN, ID3 and CHAID as classifiers for classification of chronic kidney disease. There are various authors worked in the field of various diseases and suggested various techniques for classification. Larose D. (2005) used various data mining techniques for classification like decision tree, random forest, random tree, naïve Bayes etc. for classification of kidney disease. Rajam K. (2016) also used other data mining techniques like C4.5, CART, and SVM etc. for classification of same disease. Breiman L., et al. used Classification and Regression Treee (CART) and Shrivas A.K. (2017) used CART, C5.0 and J48 classifier along with PCA applied on ensemble model for classification of kidney disease. Shrivas A.K. (2018) used CART and SVM to classify Chronic Kidney Disease also applied Ranking based feature selection Techniques to increase performance of individual as well as ensemble model.

METHOD AND MATERIALS

Methods and material play very important role in field of research work. In this research work we have used different classifiers and materials

- **Classifiers**
  Classification is one of the important applications of data mining. In this research work we have used different data mining based classification techniques used in this research work.

a) Decision Tree

Decision Trees are the most popular architectures widely used in data mining (Larose D. et al., 2005). These architectures use a divide-and-conquer strategy in order to partition the instance space into decision regions. The decision tree architectures consist of a root node, branches, internal n

odes, and leaf nodes. There are three main steps for classification by using decision trees: The first step is the learning process. The model is constructed on the training data. Hence, this model is presented by classification rules. In the second step, a test is selected in order to calculate the model accuracy. The model is accepted according to the value of this test. If this value is considerably accepted, the model could be used for the classification of a new datum. At last, the third step includes the usage of the model for a classification or prediction of a new data.

b) Random Forest

Random Forest is a group of tree predictors (Breiman L. et al., 2001) A random vector is used. It is sampled independently by using the same distribution θkis handled from the old vectors θ1, θ2, θk−1. X is defined as an input vector. The construction of the tree is handled on the training set by using the random vector θk. The resulting is defined with h (X, θk). If a large numbers of trees are generated, they are voted in order to find the most popular class. The procedure is called as random forests. It is a classifier. Each tree has a cost as a vote for the class selected the most popular at input X.

c) Random Tree

RANDOM Tree is a multiple random tree algorithm (Witten I. et al., 2000 and Fan W. et al. 2003). A non-tested attribute is selected randomly from the whole data set without using a training set. A limit is predefined. The tree has been constructed until the depth of the tree exceeds this predefined limit. If the depth of the tree exceeds this limit, it stops. The training set is used in order to update the statistics of each node. The class attribute contains different classes. Each node sets down the number of records classified as different classes. The same process is applied for the classification by using a decision tree. Each data record is read in order to update multiple random trees. It is necessary to complete one scan of the data. The classification of a datum x is performed by averaging the probability outputs from multiple random trees.

d) Naïve Bayes

Bayesian classification [Rajam K. et al. (2016)] is derived from the Bayes theorem. It is also known as "Simple Bayesian Classifier". In this classifier each data sample is represented by an n dimensional vector and measurements samples are formed by n-attributes .Suppose there are m classes, C1, C2 …..C3 having a unknown data sample, X, Naive Bayesian classifier will predict that belongs to class having highest probability conditioned on X an unknown sample X to the class Ci.

e) K-NN

In pattern recognition, the KNN algorithm is a method for classifying objects based on closest training examples in the feature space. KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification [Cover et al. ,1967). The KNN is the fundamental and simplest classification technique when there is little or no prior knowledge about the distribution of the data.

f) ID3

ID3 is a supervised learning algorithm (Shrivastava A. et.al., 2012) builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. ID3 algorithm builds tree based on the information (information gain) obtained from the training instances and then uses the same to classify the test data. ID3 algorithm generally uses nominal attributes for classification with no missing values.

g) CHAID

CHAID (Chi-Squared Automated Interaction Detector) (Kass G. ,1980) is one of the classification tree algorithms, is the name given to one version of the Automatic Interaction Detector that has been developed for categorical variables. In fact, CHAID is a technique that recursively partitions (or splits) a population into separate and distinct segments. These segments, called nodes, are split in such a way that the variation of the response variable (categorical) is minimized within the segments and maximized among the segments.

➢ **Feature optimization technique**

Feature optimization is an optimization technique in which results are obtained with some selected features only. Optimization selection uses both forward and backward eliminations. Here it is applied on Naïve Bayes model with different partitions for better accuracy performance.

DATA SET

Dataset for chronic kidney disease is acquired from UCI machine repository with 24 features, 400 example sets and 1 class with binary nature. We have used 10 fold cross validation method for data partitioning with testing and training.

RESULT AND DISCUSSION

The Figure 1 shows that architecture of proposed work. In this research work we have used Chronic Kidney disease dataset form UCI repository and partition data set into training and testing. The training data set is used trained the classifiers and testing data set is used to test the trained classifier. Finally we have selected the best classifier and applied the feature optimization technique on best ID3 classifier to achieve high accuracy.

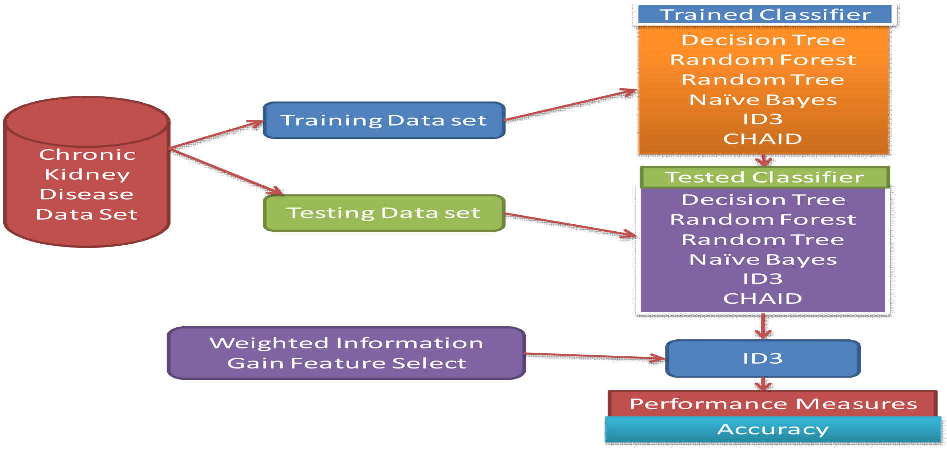

Figure 1: Architecture of proposed model

In this research work we have performed analysis in RapidMiner tools in window environment along with various classifier models in different partitions. We have applied like 70%-30%, 75%-25%, and 10 fold cross validation method. The experimental work divided into two sections: (i) Analysis of classifier (ii) Feature selection.

The classifiers like decision tree, random tree, random forest, naïve Bayes, ID3 and CHAID are trained and tested individually with given partitions. From which we got the best classifier named ID3 which gives best accuracy compared to others. Table 1 and Figure2 shows the accuracy of classifiers with different data partitions. The suggested classifier ID3 gives 97.0% of accuracy as best classifier with 10-fold cross validation data partition.

Table 1: Accuracy of Classifiers with different partitions

Model Name	70-30%	75-25%	10-fold Cross validation
Decision Tree	96.67%	96.00%	89.25%
Random Forest	93.33%	90.00%	88.75%
Random Tree	81.67%	87.00%	73.00%
Naïve Bayes	93.33%	95.00%	96.75%
ID3	**96.67%**	**96.00%**	**97.00%**
CHAID	83.33%	84.00%	89.50%

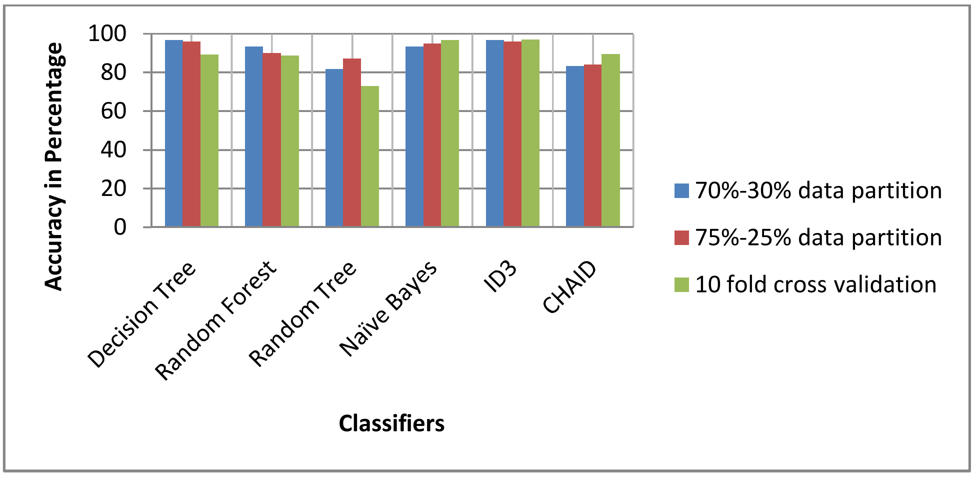

Figure 2: Accuracy of classifiers with different data partitions

We have applied Weighted Information Gain feature selection technique in best classifier ID3 with 10 fold cross validation (partition) which optimizes some features for better performance accuracy. The Weighted Information Gain feature selection technique ranks the features of chronic kidney disease like 9, 21, 24, 8,5,23,7,2,22,1,20,6,17,19,4,13,14,3,11,10,18,16,12,15 as feature_Id. Table 2 and Figure 3 shows that accuracy of ID3 classifier with optimizes feature subset. The performance accuracy of ID3 classifiers gives 97% of accuracy with 20 features, 96.50% of accuracy with 15 features, 96.25% of accuracy with 10 features and 84.25% of accuracy with 05 features.

Table 2: Accuracy of ID3 using Weighted Information Gain FST

No. of Features	Accuracy
20	97.00%
15	96.50%
10	96.25%
05	84.25%

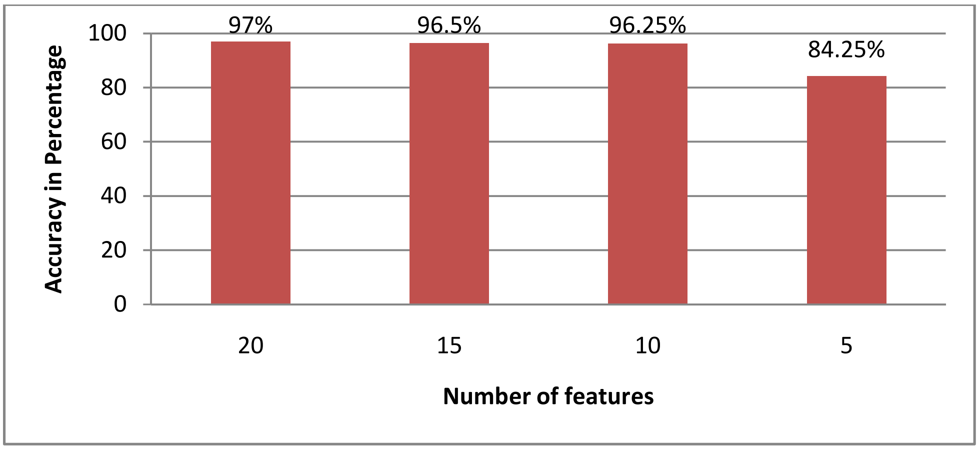

Figure 3: Accuracy of best ID3 classifier with different feature subsets

CONCLUSION AND FUTURE WORK

Medical science is the fields were any research work cannot be stated as it's completed. Various researchers have given their whole life in this field. We cannot conclude the result here only. Classification is just a technique to classify the data and mark its performance accuracy. In this paper also we have used various data mining techniques to build powerful classifiers. We have tested various individual models in different partitions and select the best classifier with high accuracy. In our proposed model we have used feature selection only on ID3 with 10 fold cross validation which gives better result as 97.00% of accuracy with 20 features. In future we will try for develop new model with different classifiers to obtain much better result.

REFERENCE

Shrivas A.K. , Sahu S. K. and Singhai S.K., "Decision Support System For Classification Of Chronic Kidney Disease With Principle Component Analysis" Review of Business and Technology Research, Vol. 14, No. 2, 2017.

Shrivas A.K., Sahu S. K. and Hota H.S., "Classification Of Chronic Kidney Disease With Proposed Union Based Feature SelectionTechniques" 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), pp. 503-507, 2018.

Ankur Shrivastava and Vijay Choudhary, "Comparison between ID3 and C4.5 in Contrast to IDS Surbhi Hardikar", VSRD-IJCSIT, Vol. 2 (7), pp. 659-667, 2012.

Breiman L., Friedman J. and Olshen R., "Classification and Regression trees", Chapman & Hall, London, 1984.

Cover, T.M. & Hart, P.E. (1967) "Nearest neighbor pattern classification", IEEE Trans. Inf. Theory, Vol. 13. pp. 21–27, 1967.

Kass G. V., "An Exploratory Technique for Investigating Large Quantities of Categorical Data", Applied Statistic, Vol. 29, pp. 119-127, 1980.

QUINLAN J.R., Induction of Decision Trees, Vol. 1, pp. 81-106, 1986.

Kohavi R., Scaling up the accuracy of naïve ayes classifiers: a decision tree hybrid, In: Proc. of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, pp. 202-207, 1996.

Larose D., "Discovering knowledge in data: An introduction to data mining", John Wiley & Sons, Vol. 385, 2005.

Rajam K.., "A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol. 5(5), pp. 354-358, 2016.

Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Mateo, CA, (2000), W. Fan, H. Wangi, P. S. Yu, et al; Is a random model better? On its accuracy and efficiency, In: Proceedings of Third IEEE International Conference on Data Mining (ICDM), Vol. 51, 2003. *http://dx.doi.org/10.1109/ICDM.2003.1250902*

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).
www.healthline.com