Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

Entropy Based Feature Selection for Chronic Kidney Disease Classification

Muhammad Nasir Dankolo
Kebbi State University of Science and Technology,
Aliero, Kebbi State

Salisu Ibrahim
Department of Computer science
Shehu Shagari College of Education Sokoto, Sokoto State

Amina Imam Abubakar
Department of Computer science
Shehu Shagari College of Education Sokoto, Sokoto State

***Abstract* – Chronic kidney disease is a general term for heterogeneous disorders affecting kidney structure and function. It is recognized now that even mild abnormalities in measures of kidney structure and function are associated with increased risk for developing complications in other organ systems which lead to mortality, all of which occur more frequently than kidney failure. Data mining has been a current trend for attaining diagnostic results. Huge amount of un mined data is collected by the healthcare industry in order to discover hidden information for effective diagnosis and decision making. Data mining is the process of extracting hidden information from massive dataset, categorizing valid and unique patterns in data. In this research we are going to use entropy feature selection method to improve the classification of chronic kidney disease. The experimental result shows that there is a significance improvement in performance of classifiers when feature selection is applied.**

I. INTRODUCTION

Data mining is one of the most interesting fields of research with the purpose of finding useful information from huge amount of data sets collected for different purpose. It has been used in several areas such as recommendation system, sentiment analysis, face recognition, learning analytics, medical diagnosis, etc. Its applications include outlier detection, fraud detection, weather forecast, market basket analysis, medical data analysis, social network analysis, etc.

Chronic kidney disease is a general term for heterogeneous disorders affecting kidney structure and function. It is recognized now that even mild abnormalities in measures of kidney structure and function are associated with increased risk for developing complications in other organ systems which lead to mortality, all of which occur more frequently than kidney failure [6].

Nowadays Chronic Kidney Disease (CKD) is regarded as a global health issue and is an area of concern both medical and computing where data mining approach become helpful tool to enable extraction of knowledge from huge amount of data generated by healthcare industry about patients, diseases, hospitals, medical equipment, claims, treatment cost etc; since such datasets requires careful processing and analysis Yadollahpour [7].

This paper focused on building a classification model with feature selection using information gained to classify Chronic Kidney Disease (CKD) using Classification algorithms such as Decision tree, Naive Bayes, Support Vector Machine and Random Forest. The remaining part of this paper is organized as follows: Section II review of related works; Section III the methodology used while experimental procedures and results are discussed and analyzed in section IV. Finally, section V concludes the paper.

II. RELATED WORK

[5] compared three (3) techniques; for predicting Kidney Dialysis Survivability where they used three data mining techniques (Artificial Neural Networks, Decision tree and Logical Regression). However, their results were not reflected in the paper.

[2] Analyzed a dataset using six classification algorithms for classification task to study their classification accuracy and performance over the Chronic-Kidney-Disease data set. The classifiers in Weka have been categorized into different groups such as Bayes, Functions, Lazy, Rules, Tree based classifiers etc. A good mix of algorithms has been chosen from these groups which are used in distributed data mining. They include Naive Bayes (from Bayes), Multilayer Perceptron, SVM, J48, Conjunctive rule and Decision Table. These algorithms have been compared with classification accuracy to each other on the basis of correctly classified instances, time taken to build model, time taken to test the model, mean absolute error, Kappa statistics and ROC Area. In the experiments Multilayer perceptron algorithm gives better classification accuracy and prediction performance to predict chronic kidney disease (CKD) using relevant dataset available at UCI machine learning repository. Hence it is concluded that Multilayer perceptron classifier performs well if we take all parameters in to consideration.

[3] Compare and evaluated 12 classification techniques by applying them to the Chronic Kidney Disease data. In order to calculate efficiency, results of the prediction by candidate methods were compared with the actual medical results of the subject. The various metrics

used for performance evaluation are predictive accuracy, precision, sensitivity and specificity. The results indicate that decision-tree performed best with nearly the accuracy of 98.6%, sensitivity of 0.9720, precision of 1 and specificity of 1.

[4] Used J48 and SMO algorithms in the estimation process, which is performed by trained testing and training data in two different sizes. At Test-1, algorithms are trained by creating data in approximately 66% of all data. The estimation procedure is done for algorithms by creating test data with the remaining 34% of data. At Test-2, algorithms are trained by creating data in approximately 10% of all data. The estimation procedure is done for algorithms by creating test data with the remaining 90% of data. In the classification stage, decision tree has been more successful than the support vector machine recognition of 97% to 100% recognition. Forecasting and phase accuracy, it has been observed that made an accurate estimate of 100% rate with the decision tree in the Test-1 data set and an accurate estimate of 91.67% rate in the Test-2 data set. Also it has been observed that made an accurate estimate of 97.06% rate with the SVM in the Test-1 data set and an accurate estimate of 96.12% rate in the Test-2 data set.

All the above authors, however, did not use feature selection in their work. This shows that there is need to investigate the performance of the classifiers when feature selection is applied.

III. METHODOLOGY

This section provides description of the research methodology used in this study. It gives a brief overview of the dataset used, proposed system design and the selected algorithms to be used.

A. CHRONIC KIDNEY DISEASE DATASET DESCRIPTION

The dataset used in this study was obtained from the UCI machine learning repository Kunwar et al. (2016). The dataset contains data of 400 instances and a total of 25 attributes including the class. Table 1 the dataset. In the preprocessing of the data the missing values were dealt with by replacing numeric and discrete integer values by attribute mean of the all the instances with the same class-label as that of the instance under consideration and nominal values were replaced using attribute mode.

Table 1: Dataset Description.

Attribute	**Domain**
Age	Discrete Integer Values
Blood pressure	Discrete Integer Values
Specific gravity	Nominal Values (1.005,1.010,1.015,1.020,1.025)
Albumin	Nominal Values (0,1,2,3,4,5)
Sugar	Nominal Values (0,1,2,3,4,5)
Red blood cells	Nominal Values (Normal, Abnormal)
Pus cell	Nominal Values (Normal, Abnormal)
Pus cell clumps	Nominal Values (Present, Not-Present)
Bacteria	Nominal Values (Present, Not-Present)
Blood glucose random	Discrete Integer Values
Blood urea	Discrete Integer Values
Serum creatinine	Numeric Values
Sodium	Discrete Integer Values
Potassium	Numeric Values
Hemoglobin	Numeric Values
Packed cell volume	Discrete Integer Values
WBC count	Discrete Integer Values
RBC count	Numeric Values
Hypertension	Nominal Values (Yes, No)
Diabetes mellitus	Nominal Values (Yes, No)
Coronary artery disease	Nominal Values (Yes, No)
Appetite	Nominal Values(Good, Poor)
Pedal edema	Nominal Values(Yes, No)
Anemia	Nominal Values(Yes, No)

B. PROPOSED DESIGN

Figure 1, shows the proposed system design adopted for this study

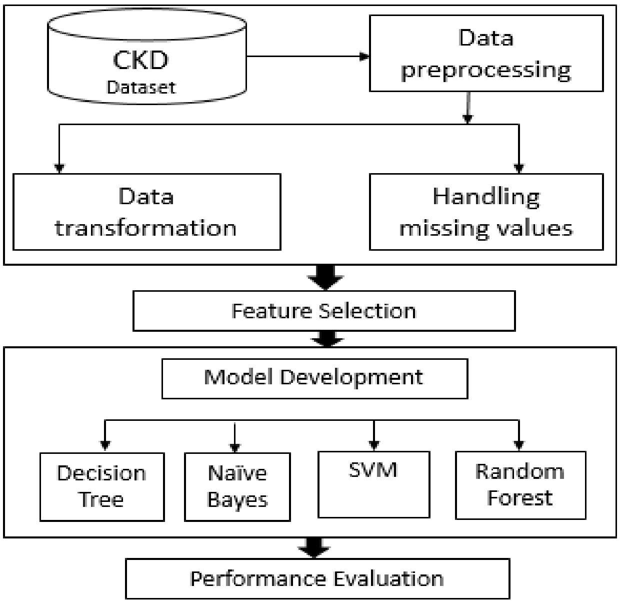

Figure 1

C. CLASSIFIERS

This section describes the classification algorithms used construct the models. A total of 6 classification algorithms have been used. The classifiers used include Decision Tree, Naïve Bayes, Support Vector Machine and Random Forest

Decision Tree Algorithm

Decision tree classifiers classify data by making use of tree structure algorithms [6]. The underlying algorithm begins with the training samples and corresponding class labels. The training set is partitioned recursively based on a feature value into subsets. Each internal node represents a test on attribute; each edge (branch) represents an outcome of the test. A decision tree classifier identifies the class label of an unknown sample by following path root to the leaves, which represent the class label for that sample. The feature (attribute) i.e. selected as the root node is the one that best divides the training data. There are number of ways for finding the feature that best divides the training data, some of them are namely Information gain, Gini Index, Gain Ratio, G-statistics, chi-square, MDL etc. One cannot generalize any measure to be better than others. Any measure that results in a multiway tree (hence reduced complexity) and more balanced splits may be used depending on the dataset. Some of the commonly used decision tree algorithms are ID3, CART, C4.5, Random and J48.

Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. It assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. Naive Bayes classifier is referred to as naive since it makes the assumption that each of its inputs are independent of each other and an assumption which rarely holds true, and hence the word naive. Naive Bayesian algorithm is seen as a simple powerful tool in the world of classification and machine learning. Naive Bayesian requires small amount of training data to estimate the parameters. It can solve problems involving categorical and continues values and it can be easily implanted and used.

Support Vector Machine (SVM)

These classifiers are based on structural risk minimization principal and statistical learning theory with an aim of determining the hyperplanes (decision boundaries) that produce the efficient separation of classes. The underlying algorithm is Support Vector Classification (SVC) and it revolves around the perception of a "margin"- on either side of a hyperplane that divides two data classes. Maximizing the margin creates the largest possible distance among the hyperplane and the instances on either side of the hyperplane reduce an upper bound on the anticipated generalization error. It works on two types of data i.e. linearly separable data and linearly Non-separable data. In case of linearly separable data only one hyperplane is needed for separating the data but in the case of latter more than one hyperplanes are needed. Figure 2 depicts an example of a two-class problem with one separating hyperplane.

Random Forest

Random Forest (Breiman, 2001) is supervised ensemble machine learning approach for classification, regression and other tasks that operates by constructing a number of decision trees during training and producing as its output the class that is mode of the classes of the individual trees. Unlike in decision tree where each node is split using the best among the attributes, in Random Forest each node is split using the best among a subset of predictors randomly chosen at the node. This strategy makes Random Forest perform very well when compared to many other classification algorithms including Neural Network, Support Vector Machine and Discriminant Analysis among others and it is robust against overfitting.

IV. EXPERIMENT AND RESULT

A. Data set

The clinical data of 400 records considered for analysis has been taken from UCI Machine Learning Repository. The data obtained and used after transformation, cleaning and handling missing values.

The data has been implemented using Rapid Miner tool. There are 25 attributes in the dataset. The numerical attributes include age, blood pressure, blood glucose random, blood urea, serum creatinine, sodium, potassium, emoglobin, packaged cell volume, WBC count, RBC count. The nominal attributes include specific gravity, albumin, sugar, RBC, pus cell, pus cell clumps, bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia and class. Number of Instances: 400

Number of Attributes: 25 Class: {CKD, NOTCKD}, Missing Attribute Values: yes

Class Distribution: [63% for CKD] [37% for NOTCKD]

B. MODEL CONSTRUCTION

The whole experiment was conducted using rapid miner analytical tool.

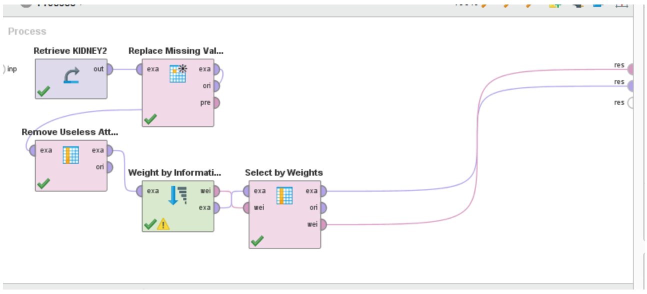

Figure 2, the construction of the model in rapid miner

C. RESULTS

After applying entropy feature selection in the above model, we used information gained to weight the attributes in the data set and return then in ascending order of weight. Initially, we select top 10 attributes and observe the performance of the classifier, we continuously change the setting until we arrive at the best choice which is top 11. The experiment shows that the selected top 11 are the features with most information gained. The figure below show the top 11 feature how they are correlated.

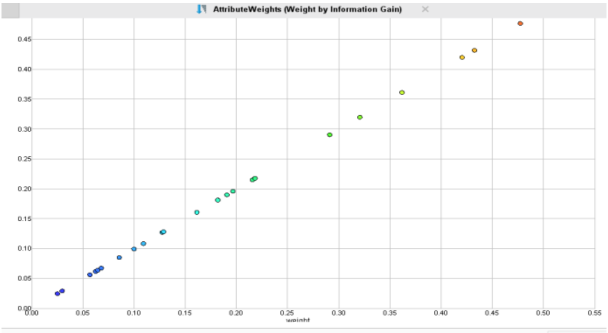

Figure2, most informative features

After we apply the models, we have record a high increase in accuracy compared to the results that was done without feature selection. Below is the summary table of results for each classifier used.

CLASSIFIER	ACCURACY	PRECISION	RECALL	F-MEASURE
Decision Tree	97.5%	97.75	97.22%	97.48%
Naïve byes	98.00%	97.83%	98.12	97.9%
SVM	94.75%	94.34%	95.37%	94.85
Random Forest	95.75%	97.01%	96.36%	96.68%

Table2, classifiers accuracy

To evaluate our results, we compare our results with the results from other authors that used the same data set as our own but did not apply feature selection

V. CONCLUSION

The study revealed that, the accuracy of the prediction and diagnosis of Chronic Kidney Disease can be improved by enhancing performance of the classification algorithms through feature selection. We also shows that entropy feature selection can be used as feature selection technique to improve the performance of the classifiers.

Our results show a significant difference in performance of the classifiers compared to the results of the same classifiers with the same data set but without feature selection.

The result also shows that Naïve Bays classifiers out performed SVM, decision tree and Random forest with an accuracy of 98.00%.

REFERENCE:

[1] Kunwar, Veenita, et al. "Chronic Kidney Disease analysis using data mining classification techniques." *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*. IEEE, 2016.

[2] Jena, Lambodar, and Narendra Ku Kamila. "Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease." (2015).

[3] Sharma, Sahil, Vinod Sharma, and Atul Sharma. "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis." *arXiv preprint arXiv:1606.09581* (2016).

[4] Celik, Enes, Muhammet Atalay, and Adil Kondiloglu. "The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods." (2016): 27-31.

[5] Lakshmi, K. R., Y. Nagesh, and M. Veera Krishna. "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability." *International Journal of Advances in Engineering & Technology* 7.1 (2014): 242.

[6] Pavithra, N., and R. Shanmugavadivu. "Survey on Data mining Techniques used in Kidney related Diseases."

[7] Yadollahpour, A. "Applications of expert systems in management of chronic kidney disease: a review of predicting techniques." *Orient J Comp Sci Technol* 7.2 (2014): 306-15.