International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-11)

Research Article | **November 2015**

Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease

Lambodar Jena
Department of Computer Science & Engineering,
Gandhi Engineering College, Bhubaneswar, India

Narendra Ku. Kamila
Department of Computer Science & Engineering,
C.V.Raman College of Engineering, Bhubaneswar, India

Abstract:

Since, classification is the most commonly applied data mining technique, and employs a set of pre-classified examples to develop a model that can classify the population of records at large. The major goal of the classification technique is to predict the target class accurately for each case in the data. The present study is focused on the usage of classification techniques in the field of medical science and bioinformatics. The main focus of this paper is Chronic- Kidney-Disease prediction using weka data mining tool and its usage for classification in the field of medical bioinformatics. It firstly classifies dataset and then determines which algorithm performs better for diagnosis and prediction of Chronic- Kidney-Disease. Prediction begins with identification of symptoms in patients and then identifying sick patients from a lot of sick and healthy ones. Thus, the prime objective of this paper is to analyze the data from a Chronic- Kidney-Disease dataset using classification technique to predict class accurately in each case. Many authors have used WEKA tool in their work to compare the performance of different classifiers applied on various datasets. But, none of the author worked on prediction of accuracy for chronic-kidney-disease dataset. Here, we have considered six number of classifiers to study their performance based on various parameters obtained by applying them in the dataset.

Keywords: Classification, Chronic-kidey-desease, Accuracy, prediction, distributed data mining.

I. INTRODUCTION

In present days, computers have brought significant improvements to technology that lead to the creation of huge volumes of data. Moreover, the advancement of the healthcare database management systems creates a huge number of medical databases. Creating knowledge and management of large amounts of heterogeneous data has become a major field of research, namely data mining. Data Mining is a process of identifying novel, potentially useful, valid and ultimately understandable patterns in data [1]. Data mining techniques can be classified into both unsupervised and supervised learning techniques. Unsupervised learning technique is not guided by variable and does not create a hypothesis before analysis. Based on the results, a model will be built. A common unsupervised technique is clustering [2]. Supervised learning technique requires the building of a model that is used in prior performing analysis. Supervised learning techniques that are used in both medical and clinical research are Classification, Statistical regression and Association rules [3].

In the present study, we have focused on the usage of classification techniques in the field of medical science and bioinformatics. Classification is the most commonly applied data mining technique, and employs a set of pre-classified examples to develop a model that can classify the population of records at large. The major goal of the classification technique is to predict the target class accurately for each case in the data.

The main focus of this chapter is Chronic- Kidney-Disease prediction using weka data mining tool and its usage for classification in the field of medical bioinformatics. It firstly classifies dataset and then determines which algorithm performs better for diagnosis and prediction of Chronic- Kidney-Disease. Prediction begins with identification of symptoms in patients and then identifying sick patients from a lot of sick and healthy ones. Thus, the prime objective of this chapter is to analyze the data from a Chronic- Kidney-Disease dataset using classification technique to predict class accurately in each case. The major contributions of this chapter are:

- To extract useful classified accuracy for prediction of Chronic-Kidney-Disease.
- Comparison of different data mining algorithms on Chronic- Kidney-Disease dataset.
- Identify the best algorithm based on performance for prediction of diseases.

There are several classification mechanisms that are used in analyzing medical data. These include Decision trees, K-Nearest Neighbor (KNN), Bayesian network (Naive Bayes), Neural networks, Fuzzy logic, J48 and Support Vector Machine (SVM). All these classifiers are basically learning methods and adopt sets of rules. In this chapter we have used WEKA Data Mining tool for classification techniques. Weka, a machine learning workbench implements algorithms for data pre-processing, classification, regression, clustering and association rules [4]. Implementation in weka is classified as:

1. Implementation scheme for classification;
2. Implementation schemes for numeric prediction;
3. Implemented meta-schemes.

Learning methods in weka are called classifiers which contain tunable parameters that can be accessed through a property sheet or object editor. The exploration modes in weka allow data pre-processing, learning, data processing, and attribute selection and data visualization modules in an environment that encourages initial exploration of data.

II. RELATED WORK

Many researchers have used different data mining techniques for future prediction. Dhamodharan [1] has done prediction of liver disease using Bayesian Classification through Naïve Bayes and Functional Tree (FT) algorithms. With the help of data mining techniques he has predicted and analyzed liver diseases using weka tool. Finally the author has compared the outputs obtained from Naïve Bayes and FT algorithms and concluded that Naive Bayes algorithm plays a major role in predicting liver diseases. Solanki [2] has used weka as a data mining technique for classification of sickle cell disease prevalent. The author has compared J48 and Random tree algorithms and given a predictive model for classification with respect to a person's age of different blood group types. From the experiment it is inferred that Random tree is better algorithm as it produces more depth decisions with respect to J48 for sickle cell diseases. Similarly, Joshi et al.[3] have done diagnosis and prognosis of breast cancer using classification rules. By comparing classification rules such as Bayes Net, Logistic Model Tree (LMT), Multilayer Perceptron, Stochastic Gradient Descent (SGD), Simple Logistic, Sequential Minimal Optimization (SMO), AdaBoostM1, Attribute Selected, Classification via Regression, Filtered Classifier, Multiclass Classifier and J48, they suggested that LMT Classifier gives more accurate diagnosis i.e. 76 % healthy and 24 % sick patients. However, David et al.[4] have used classification techniques for leukaemia disease prediction. They have compared the output obtained by using K- Nearest Neighbor, Bayesian Network, Random tree and J48 tree on the basis of accuracy, learning time and error rate. According to them Bayesian algorithm performs well on classification. But in 2013, Vijayarani and Sudha [5] have compared the analysis of classification function techniques for heart disease prediction. The classification was done using algorithms such as LMT, Multilayer Perceptron and Sequential Minimal Optimization algorithms for predicting heart disease. In this classification comparison LMT algorithm is found best classifier for heart disease with more accuracy and least error rate. In the same time Kumar [6] used alternating decision (AD) trees for early diagnosis of dengue fever. The AD Tree correctly classifies 84 % of cases as compared to J48 (which classifies only 78% of cases correctly). In the same year Durairaj and Ranjani [7] have compared different data mining applications in healthcare sector. They have used algorithms such as Naïve, J48, KNN and C4.5 for classification in order to diagnose diseases like Heart Disease, Cancer, AIDS, Brain Cancer, Diabetes, Kidney Dialysis, Dengue, IVF and Hepatitis C. Comparison study reveals that data mining techniques in all health care applications obtain high accuracy i.e. 97.77% for cancer prediction and around 70% for IVF treatment through data mining techniques.

In 2011, Sugandhi et al. [8] analyzed a population of cataract patient's database by weka tool. In their study they concluded that Random Tree gives 84% classification accuracy, means better performance as compared to other algorithms used for classification. Thus according to them Random Tree is a best classification algorithm for cataract patient disease. In the same year Yasodha and Kannan [9] performed analysis of a population of diabetic patient database using weka tool. They have considered different algorithms such as Bayes Network, REP Tree, J48 and Random Tree algorithms for their works and compared the outputs. The main objective of their study was to develop Diabetic expert system. By entering patient's daily glucose rate and insulin dosages the system would produce a graph from insulin history, thereby predicting and consulting the patients for their next insulin dosage.

Bin and Yau [10] have compared different classification techniques using weka for breast cancer. In their study they have used different algorithms for simulating results of each algorithm and its training. They have simulated the errors by using Bayes Network, Radial Basis function, Decision Tree and pruning and Single Conjugation Rule Learner algorithms. From their work it is concluded that Bayes Network performs well for breast cancer data. Time taken to build model is 0.19 second and accuracy 89.7 % and least error at 0.2140 as compared to other algorithms being used. Similarly Mihaila and Ananiadou [21] have compared two data mining tools i.e. weka and CRF Suite on the basis of features like Lexical, Syntactic and semantic with various parameters to compare their impacts on each algorithm. The experiments have been employed in CRF Suite implementation by using Conditional Random Field algorithm and in weka by algorithms like Support Vector machine and Random Forests to identify discourse causality trigger in the biomedical domain. Classification tasks have been performed on the basis of statistics such as F score, precision and recall. As per them, performance of CRF classifier is remarkably good, achieved F score = 79.35 % by combining three features as compared to other classifiers.

Thitiprayoonwonsge et al. [22] have analyzed dengue infection using data mining decision tree. In their work two datasets have been used from two different hospitals Srinagarindra Hospital and Songklanagarind Hospital, each having more than 400 attributes. Four classification algorithms have been used in their work for experiment purpose. The first and second experimental tests obtained an accuracy of 97.6% and 96.6%. The third experiment extracts useful knowledge when they integrated two data sets. Another objective of this paper was to detect day abatement of fever also referred as day0. In fourth experiment of day0 accuracy is very low as compared to other three experiments. Therefore physician needs day0 amongst patient in order to treat them.

All the above authors have used WEKA tool in their work to compare the performance of different classifiers applied on various datasets. But, none of the author worked on prediction of accuracy for chronic-kidney-disease dataset. Here, we have considered six number of classifiers to study their performance based on various parameters obtained by applying them in the dataset.

III. METHODOLOGY

In order to carry out experiments and implementations WEKA is used as the data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches in the field of bioinformatics. In this work we have used the data mining techniques to predict the survivability of Chronic-Kidney disease through classification of different algorithms accuracy [12, 13]. The experiments have been carried out on the Chronic-Kidney-Disease dataset downloaded from UCI machine learning repository, USA. The different classification algorithms are implemented on the dataset by using the WEKA interface.

The interface of WEKA data mining tool has four applications:

- Explorer: The explorer interface has several panels like pre-process, classify, cluster, associate, select attribute and visualize. But in this interface our main focus is on the Classification Panel.
- Experimenter: This interface provides facility for systematic comparison of different algorithms on basis of given datasets. Each algorithm runs 10 times and then the accuracy gets reported.
- Knowledge Flow: It is an alternative to the explorer interface. The only difference between this and others is that here user selects WEKA component from toolbar and connects them to make a layout for running the algorithms.
- Simple CLI: Simple CLI means command line interface. User performs operations through a command line interface by giving instructions to the operating system. This interface is less popular as compared to other three.

IV. PRELIMINARY

Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level. Classification algorithm requires classes to be defined based on the data attribute value. It describes these classes according to the characteristics of the data that is already known to belong to the classes. The classifier training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination.

In Classification, training examples are used to learn a model that can classify the data samples into known classes. The Classification process involves following steps:

- Create training data set.
- Identify class attribute and classes.
- Identify useful attributes for classification (Relevance analysis).
- Learn a model using training examples in Training set.
- Use the model to classify the unknown data samples.

A. Classifiers Used

In this work six classification algorithms have been used 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, Conjuctive rule and Decision Table. The following sections explain a brief about each of these algorithms.

- *Naïve Bayes Classifier*

It is one of the fastest statistical classifier algorithm works on probability of all attribute contained in data sample individually and then classifies them accurately. It is used to predict class membership probabilities i.e. probability about the tuple that belongs to the particular class or not. Bayesian classification is based on Bayes theorem. Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $X= (x_1, x_2, ..., x_n)$ representing some n features (independent variables), it assigns to this instance probabilities $p(C_k \mid x_1, ..., x_n)$ for each of k possible outcomes or classes.

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k|X) = \frac{p(C_k)p(X|C_k)}{p(X)}$$

In other words, using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{prior \times likeli\,hood}{evidence}$$

- *Multilayer Perceptron*

It is the most popular network architecture in today's world. Each unit performs a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output. The units are arranged in a layered feed forward topology. The network has a simple input-output model, with the weights and thresholds. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. The important issues in Multilayer Perceptron are the design specification of the number of hidden layers and the number of units in these layers [20].

Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a back propagation neural network with one or more layers between input and output layer. The following diagram illustrates a perceptron network with three layers [13].

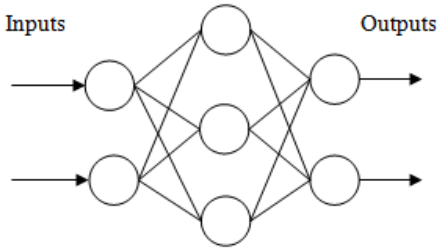

Fig. 1: Multilayer perceptron

- *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier.

- *J48*

J48 classifier is a simple C4.5 decision tree for classification. It is supervised method of classification. It creates a small binary tree. It is univarient decision tree. It is an extension of ID3 algorithm. In this classifier Divide and Conquer approach is used to classify the data. It divides the data into range based on the attribute value for that value that are found in training sample.

As this approach is range based and univarient [11], it does not perform better than multivarient approach. As this is decision tree approach it is very much useful in predicting the values. J48 accuracy of correctly classified instance is much more than that of the other algorithms which are univarient in nature [10].

- *Conjuctive Rule*

It is a decision-making rule in which the intending buyer assigns least values for a number of factors and discards any result which does not meet the bare minimum value on all of the factors i.e. a superior performance on one factor cannot recompenset for deficit on another. Conjuctive rule uses the AND logical relation to correlate stimulus attributes. Conjuctive rule is a simple well interpretable 2-class classifier.

- *Decision Table*

A decision table is a predictive modeling tool that performs classification. It incorporates an inducer (an algorithm for generating decision table models), and a visualizer. Unlike the evidence model, the Decision Table model does not assume that the attributes are independent. A decision table is a hierarchical breakdown of the data, with two attributes at each level of the hierarchy. The Decision Table inducer identifies the most important attributes (columns) for classifying the data, and the accompanying visualizer displays the resulting model graphically. It summarizes the dataset with a decision table which contains the same number of attributes as the original dataset. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table.

B. Characteristics required for Classification Algorithm

In this work, we have focused on the following three measures namely correctly classified instances, incorrectly classified instances, and accuracy.

(i) Correctly classified instance:

These are the instances which are correctly classified by any classification algorithm. Percentage of correctly classified instances is called as accuracy.

(ii) Incorrectly classified instances:

These instances are not correctly classified by the algorithm. Sometimes it is observed that the data which is incorrectly classified may contain inconsistent data, noisy data or data out of scope.

(iii) Accuracy:

Accuracy is how a measured value is closed to the true value.

The general formula is given below:

$$\text{Accuracy} = \frac{Tp+Tn}{P+N} \quad (1)$$

where, Tp indicates True positive, Tn indicates True negative, P indicates total positive, N indicates total negative. And P = Tp + Fn , N = Fp + Tn.

In classification system, the algorithm with highest accuracy will be selected for the prediction. Accuracy of the algorithm varies according to the dataset used. So before using the algorithms for prediction system, we must check the accuracy of the algorithm. So it will reduce the cost of doing trial and error of using algorithms in the prediction system.

C. Performance Evaluation

10-fold cross validation technique is used to evaluate the performance of classification methods, Data set is randomly sub divided into ten equal sized partitions. Among the partitions nine of them are used as training set and the remaining one is used as a test set. Evaluation of performance is compared using Mean absolute error, Root

mean squared error, Receiver Operating Characteristic (ROC) Area and Kappa statistics. Large test sets gives a good assessment of the classifier's performance and small training sets which result in a poor classifier.

- *Kappa Statistics*

Kappa Statistics measure degree of agreement between two sets of categorized data. Kappa result varies between 0 to 1 intervals. Higher the value of Kappa means stronger the agreement. Kappa is a normalized value of agreement for chance of agreement.

$$K = \frac{P(A)-P(E)}{1-P(E)}$$

Where P(A) = percentage of agreement
P(E) = chance of agreement.
If K =1 agreement is perfect between the classifier and ground truth.
If K=0 indicates there is a chance of agreement.

- *Mean Absolute Error (MAE)*

The mean absolute error (MAE) is a quantity used to measure predictions of the eventual outcomes. The mean absolute error is given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

The mean absolute error is an average of the absolute errors ei = | fi – yi |, where fi = prediction , yi = true value.

- *Root Mean Squared Error (RMSE)*

Root mean squared error is the square root of the mean of the squares of the values. It squares the errors before they are averaged [18] and RMSE gives a relatively high weight to large errors.

The RMSE Ei of an individual program i is evaluated by the equation:

$$E_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{P_{(i,j)}-T_j}{T_j}\right)^2}$$

where, P(i,j) = the value predicted by the individual program
i = fitness case
Tj =the target value for fitness case j.

- *Receiver Operating Characteristic (ROC) Area*

ROC Area is defined as area under the ROC curve which is the probability of randomly chosen positive instance that is ranked above randomly chosen negative one.

Receiver Operating Characteristic represents test performance guide for classifications accuracy of diagnostic test based on: excellent (0.90-1), good (0.80-0.90), fair (0.70-0.80), poor (0.60-0.70), fail (0.50–0.60).

V. DATASET USED

Dataset is a collection of data or a single statistical data where every attribute of data represents variable and each instance has its own description. For prediction of chronic kidney disease we have used Chronic-Kidney-Disease dataset for prediction and classification.

The dataset used for our experiment contains 25 (24 + class = 25 (11 numeric, 14 nominal)) attributes and 400 instances for chronic kidney disease classification and accuracy presented in Appendix. We have applied different algorithms using WEKA data mining tool for our analysis purpose. Table 1 shows the description of the attributes in Chronic-Kidney-Disease dataset.

TABLE 1 DESCRIPTION OF ATTRIBUTES IN THE CHRONIC-KIDNEY-DISEASE DATASET

Sl.No	Attribute	Description	Type	Permissible values
1	age	Age	numerical	age in years
2	bp	blood pressure	numerical	in mm/Hg
3	sg	specific gravity	nominal	(1.005,1.010,1.015,1.020,1.025)
4	al	albumin	nominal	(0,1,2,3,4,5)
5	su	sugar	nominal	(0,1,2,3,4,5)
6	rbc	red blood cells	nominal	normal,abnormal
7	pc	pus cell	nominal	normal,abnormal
8	pcc	pus cell clumps	nominal	present,notpresent
9	ba	bacteria	nominal	present,notpresent
10	bgr	blood glucose	numerical	in mgs/dl
11	bu	blood urea	numerical	in mgs/dl

12	sc	serum creatinine	numerical	in mgs/dl
13	sod	sodium	numerical	in mEq/L
14	pot	potassium	numerical	in mEq/L
15	hemo	haemoglobin	numerical	in gms
16	pcv	packed cell volume	numerical	in cells/cumm
17	wc	white blood cell count	numerical	in cells/cumm
18	rc	red blood cell count	numerical	millions/cmm
19	htn	hypertension	nominal	yes,no
20	dm	diabetes mellitus	nominal	yes,no
21	cad	coronary artery disease	nominal	yes,no
22	appet	appetite	nominal	good,poor
23	pe	pedal edema	nominal	yes,no
24	ane	anaemia	nominal	yes,no
25	class	class	nominal	ckd, notckd

VI. RESULT AND DISCUSSION

We have conducted two experiments based on the dataset with all above discussed classification algorithms; first without using feature selection and second with using Genetic search feature selection. First the results of the classification algorithms based on parameters such as accuracy of classification, kappa statistics, MAE, RMSE, model building time, model testing time, and ROC are shown in the following Table 2, where model building time and model testing time are generated by WEKA Tool itself during classification.

TABLE 2 CLASSIFICATION RESULTS FROM WEKA WITHOUT FEATURE SELECTION

Algorithm	Accuracy	Kappa statistics (K)	MAE	RMSE	ROC	Time to build the model (sec)	Time to test model on training data(sec)
Naïve Bayes	95%	0.8961	0.0479	0.2046	1	0.02	0.02
Multilayer Perceptron	99.75%	0.9947	0.0085	0.0622	1	8.63	0.02
SVM	62%	0	0.375	0.6124	0.5	0.25	0.16
J48	99%	0.9786	0.0225	0.0807	0.999	0.08	0
Conjuctive Rule	94.75%	0.8869	0.081	0.2237	0.942	0.05	0
Decision Table	99%	0.9786	0.1815	0.2507	0.992	0.22	0.02

A. Analysis of the parameters

(a) Kappa statistics

From table 2 it is seen that, each classifier produces K value greater than 0 except SVM, that means each classifier is doing better than chance of agreement. Since K=0, SVM classifier has a chance of agreement. As per the table Multilayer perceptron classifier has greater value K= 0.9947 in comparison with other classifiers for the Chronic-Kidney-Disease dataset. Therefore, it leads to the agreement which is perfect between the classifier and ground truth.

(b) Mean Absolute Error (MAE)

The mean absolute error shown in Table 2 for all classifiers, represents that Multilayer perceptron classifier achieves minimum MAE values (0.0085) and SVM classifier gives maximum MAE values i.e. 0.375. So the prediction result in case of Multilayer perceptron classifier is very close to the true value of the chronic kidney disease parameters. Whereas the prediction result of SVM is farther than the true value. However the other classifiers give average prediction result.

(c) Root Mean Squared Error (RMSE)

Train a classifier with sufficient data sets generally minimizes the error rate for testing. Error rate for training set is comparatively higher than that of the test set. From the Table 2 it is observed that Multilayer perceptron classifier has the lowest error rate i.e. =0.0622 compared to other algorithms. If two algorithms have the same mean absolute error (MAE) rate then root mean squared error (RMSE) rate is taken into consideration for choosing the best classification algorithm.

Moreover it is observed from the above table that Multilayer perceptron gives ideal results taking above three parameters in to consideration. However, SVM gives poor results for above three parameters shown in table 3 and depicted in figure 2.

TABLE 3 KAPPA STATISTICS AND ERROR RATE FOR CLASSIFIED INSTANCES

Algorithm	Kappa statistics	MAE	RMSE
Naïve Bayes	0.8961	0.0479	0.2046
Multilayer Perceptron	0.9947	0.0085	0.0622
SVM	0	0.375	0.6124
J48	0.9786	0.0225	0.0807
Conjuctive Rule	0.8869	0.081	0.2237
Decision Table	0.9786	0.1815	0.2507

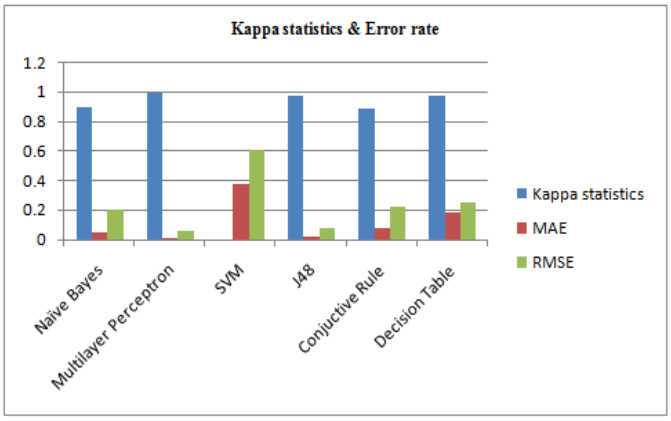

Fig. 2 Kappa statistics and Error rates

(d) Classification Accuracy

The accuracy in Table 2 has been evaluated as per equation (1) and shown in Table 4 and depicted in Figure 3. From both table and figure it is observed that Multilayer perception algorithm gives more classification accuracy i.e. 99.75% comparing to all other classifiers. However, it is interesting to note that all algorithms have classification accuracy more than 90% except SVM which performs very poor. Hence it is concluded that Multilayer Perceptron performs well in case of chronic-kidney-disease dataset.

TABLE 4 CLASSIFICATION ACCURACY

Algorithm	Accuracy
Naïve Bayes	95%
Multilayer Perceptron	99.75%
SVM	62%
J48	99%
Conjuctive Rule	94.75%
Decision Table	99%

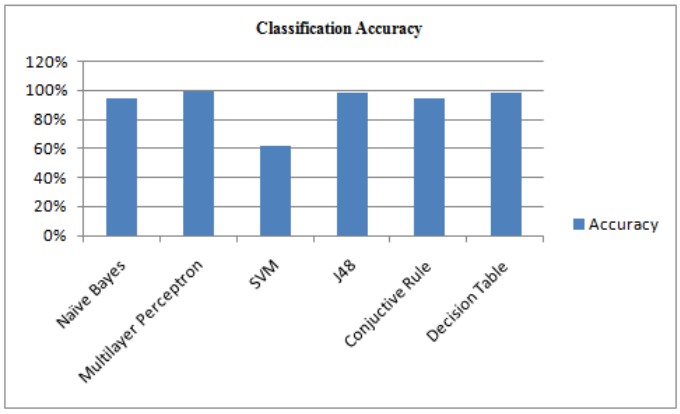

Fig. 3 Classification Accuracy of different classifiers

(e) ROC, Time to build the model, Time to test model on training data

The ROC, Time to build the model and Time to test model on training data of all the classification algorithms are shown in the Table 5 and depicted in figure 4.

TABLE 5 COMPUTATION OF ROC, TIME TO BUILD THE MODEL & TIME TO TEST MODEL ON TRAINING DATA

Algorithm	ROC	Time to build the model (sec)	Time to test the model (sec)
Naïve Bayes	1	0.02	0.02
Multilayer Perceptron	1	8.63	0.02
SVM	0.5	0.25	0.16
J48	0.999	0.08	0
Conjuctive Rule	0.942	0.05	0
Decision Table	0.992	0.22	0.02

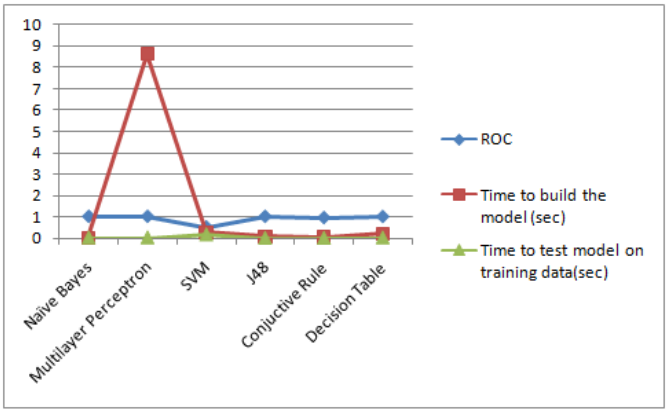

Fig. 4 Value of ROC, Time to build the model & Time to test model on training data

From the above Table 5 and the Figure 4 it is observed that ROC value in case of SVM is very low, whereas ROC of other algorithms is approximately same. The model building time for Multilayer perceptron classifier is relatively high. Similarly model testing time for Multilayer perceptron classifier is less. Indirectly it presumes that high model building time is considered only when the classifiers get trained with sufficient data sets for which model testing time gets reduced. Moreover, ROC value is high in Multilayer perceptron classifier which indicates that classification accuracy is high. Hence it is concluded that Multilayer perceptron classifier performs well if we take all parameters in to consideration.

VII. CONCLUSION

The main objective of this chapter is to predict chronic kidney disease. We have used six algorithms i.e. Naive Bayes, Multilayer Perceptron, SVM, J48, Conjuctive Rule and Decision Table for our experiments. These algorithms are implemented using WEKA data mining tool to analyze accuracy which is obtained after running these algorithms in the output window. 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.

REFERENCES

[1] Dhamodharan S., Liver Disease Prediction Using Bayesian Classification, Special Issues, 4th National Conference on Advance Computing , Application Technologies, May 2014

[2] SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014.

[3] Joshi J, Rinal D, Patel J, Diagnosis And Prognosis of Breast Cancer Using Classification Rules, International Journal of Engineering Research and General Science,2(6):315-323, October 2014.

[4] David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.

[5] Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, 1(3): 735-741, 2013.

[6] Kumar M. N., Alternating Decision trees for early diagnosis of dengue fever .arXiv preprint arXiv: 1305.7331,2013.

[7] Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.
[8] Sugandhi C , Ysodha P , Kannan M , Analysis of a Population of Cataract Patient Database in WEKA Tool , International Journal of Scientific and Engineering Research ,2(10) ,October ,2011.
[9] Yasodha P, Kannan M, Analysis of Population of Diabetic Patient Database in WEKA Tool, International Journal of Science and Engineering Research, 2 (5), May 2011.
[10] Bin Othman M. F , Yau, T. M. S., Comparison of different classification techniques using WEKA for breast cancer, In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, Springer Berlin Heidelberg, 520-523,January 2007.
[11] Wikipedia, http://en.m.wikipedia.org/wiki/Dengue_fever, accessed in January 2015.
[12] Wikipedia, http://en.m.wikipedia.org/wiki/weka (machine learning), accessed in January 2015.
[13] KirkbyR, Frank E, WEKA Explorer User Guide for version 3-4-3, November2004.
[14] Wikipedia, http://en.m.wikipedia.org/wiki/Naive_Bayes_classifier, accessed in January 2015.
[15] Hall M, Reutemann P, WEKA Knowledge Flow Tutorial for version 3-5-8, July 2008.
[16] Scuse D, Reutemann P, WEKA Experimenter Tutorial for version 3-5-5, January 2007.
[17] Shomona Gracia Jacob, R.Geetha Ramani, Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications (0975– 8887)Volume 32– No.7, October 2011.
[18] Mihaila C, Ananiadou S, Recognising Discourse Causality Triggers in the Biomedical Domain, Journal of Bioinformatics and Computational Biology, 11(06), October 2013.
[19] Thitiprayoonwongse D., Suriyaphol P., Soonthornphisaj N., Data mining of dengue infection using decision tree, Entropy, 2: 2, 2012.
[20] Alam M., Shakil K.A., Cloud Database Management System Architecture, UACEE International Journal of Computer Science and its Applications, 3(1):27-31, 2013.
[21] Alam M., Shakil K.A., A decision matrix and monitoring based framework for infrastructure performance enhancement in a cloud based environment, International Conference on Recent Trends in Communication and Computer Networks, Elsevier, pp. 174-180, November 2013.
[22] Alam M., Shakil K.A., An NBDMMM Algorithm Based Framework for Allocation of Resources in Cloud, arXiv preprint arXiv: 1412.8028, 2014.
[23] Shakil K.A. and Alam M., Data Management in Cloud Based Environment using k-Median Clustering Technique, IJCA Proceedings on 4th International IT Summit Confluence 2013 - The Next Generation Information Technology Summit Confluence 2013, pp. 8-13, January 2014.
[24] Alam M., Shakil K.A., and Sethi S. ,Analysis and Clustering of Workload in Google Cluster Trace based on Resource Usage, arXiv preprint arXiv: arXiv: 1501.01426, 2014.
[25] Alam M., Shakil K.A., Recent Developments in Cloud Based Systems: State of Art, arXiv preprint arXiv: arXiv: 1501.01323, 2015.
[26] Shakil K.A., Sethi S., Alam M., An Effective Framework for Managing University Data using a Cloud based Environment, arXiv preprint arXiv:1501.07056, 2015
[27] Zareen F.J. and Jabin S., A Comparative Study of recent trends in biometric signature verification,IC3, IEEE, 354-358, 2013.

