OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Generating comparative analysis of early stage prediction of Chronic Kidney Disease

L.Jerlin Rubini, Dr.P.Eswaran

[a]Research Scholar, Department of Computer Science and Engineering, Alagappa University, Karaikudi
[b]Assistant Professor, Department of Computer Science and Engineering, Alagappa University, Karaikudi

ABSTRACT: Chronic Kidney Disease prediction is one of the most important issues in medical decision making. The discovery of ckd prediction is an important task because it depends on experts of doctor knowledge. Construct effective ckd prediction in time is essential to prevent healthy patients. Chronic kidney disease is one of the leading cause of death and early prediction of chronic kidney disease is important. Prediction is most interesting and challenging tasks in day to life. Data mining play a essential role for prediction of medical dataset. It extract unknown information from hidden knowledge. This paper can proposed a new chronic kidney disease dataset with three classifiers such as radial basis function network, multilayer perceptron, and logistic regression. The obtained result of this experiment shows in terms of prediction accuracy, type I error, type II error, type I error rate, type II error rate, sensitivity, specificity, F-score. Kappa value represents that measure of agreement between the classification made by the experts and classifiers. Accuracy of the three classifiers are evaluated for the new CDK dataset from UCI repository. Thus, the paper discussed the result of comparative study of classifiers in medical ckd dataset.

Keywords: CKD, classification, RBF network, MLP, Logistic Regression.

I. INTRODUCTION

Chronic kidney disease is one of the Kidney disease in medical field. Kidneys are a pair of organs located toward lower back of the body .It can be placed on either side of the spine. Main function of the Kidney act as a filtration system for blood and to remove toxins from body. The kidney shifts the toxins to the bladder then it later removed from the body through urination. Kidney failure occurs when the kidneys unable to filter waste from the blood.If kidneys cannot perform their regular job then body becomes overloaded with toxins. This can lead to kidney failure and can result in death. Kidney failure suffers from one or more of the following causes: Loss of Blood Flow to the Kidneys, Damage to the Kidneys and Urine Elimination Problems. Kidney problems can be either acute or chronic (fig:1). Acute kidney disease is the sudden loss of kidney function that occurs when high levels of waste products of the body's metabolism accumulate in the blood.

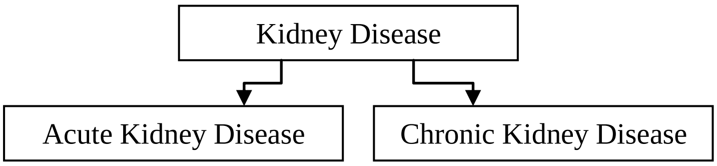


Fig:1 Types of kidney disease

Chronic kidney disease is a gradual development of permanent kidney disease.It is the most common type of kidney disease and occurs when the kidneys are damaged or are not functioning for some months or longer. Some of the leading causes of chronic kidney disease are diabetes, hypertension, lupus and complications from some medications. Medications can control hypertension and diabetes and changes in diet and lifestyle. Complications of chronic kidney disease such as anemia and weak bones leading to fractures.Chronic kidney disease includes a number of conditions affecting function of kidney.Many people may be in the early stages of kidney disease and not have any indication .There are certain symptoms,as follows for chronic kidney disease.

- Diabetes
- High blood pressure
- Coronary artery Disease
- Anemia
- Bacteria and albumin in urine
- Deficiency of Sodium and Potassium in blood and Family history of kidney disease.

Blood and urine tests, ultrasound and other tests can check the status of kidney function. The treatment options for end stage chronic kidney disease are dialysis and kidney transplantation. The main objective of the this paper is to test new chronic kidney disease dataset with classifiers.Section2 shows the related work of the data mining and three classifiers.Section3 shows CKD dataset. Section4and 5 shows result and discussion.Section6 shows conclusions.

II. RELATED WORK

Data mining (DM), also known as "knowledge discovery in databases" (KDD), is the process of discovering meaningful patterns in huge databases (Han & Kamber, 2001). In addition, it is also an application that can provide significant competitive advantages for making the right decision. (Huang, Chen, & Lee, 2007). DM is an explorative and complicated process involving multiple iterative steps. Fig.3 shows an overview of the data mining process (Han & Kamber, 2001). It is interactive and iterative, involving the following steps:

- *Step 1*. Application domain identification: Investigate and understand the application domain and the relevant prior knowledge. In addition, identify the goal of the KDD from the administrators' or users' point of view.
- *Step 2*. Target dataset selection: Select a suitable dataset, or focus on a subset of variables or data samples where data relevant to the analysis task are retrieved from the database.
- *Step 3*. Data Preprocessing: the DM basic operations include 'data clean' and 'data reduction': In the 'data clean' process, we remove the noise data, or respond to the missing data field. In the 'data reduction' process, we reduce the unnecessary dimensionality or adopt useful transformation methods. The primary objective is to improve the effective number of variables under consideration.
- *Step 4*. Data mining: This is an essential process, where AI methods are applied in order to search for meaningful or desired patterns in a particular representational form, such as association rule mining, classification trees, and clustering techniques.
- *Step 5*. Knowledge Extraction: Based on the above steps it is possible to visualize the extracted patterns or visualize the data depending on the extraction models. Besides, this process also checks for or resolves any potential conflicts with previously believed knowledge.
- *Step 6*. Knowledge Application: Here, we apply the found knowledge directly into the current application domain or in other fields for further action.
- *Step 7*. Knowledge Evaluation: Here, we identify the most interesting patterns representing knowledge based data on some measure of interest. Moreover, it allows us to improve the accuracy and efficiency of the mined knowledge.

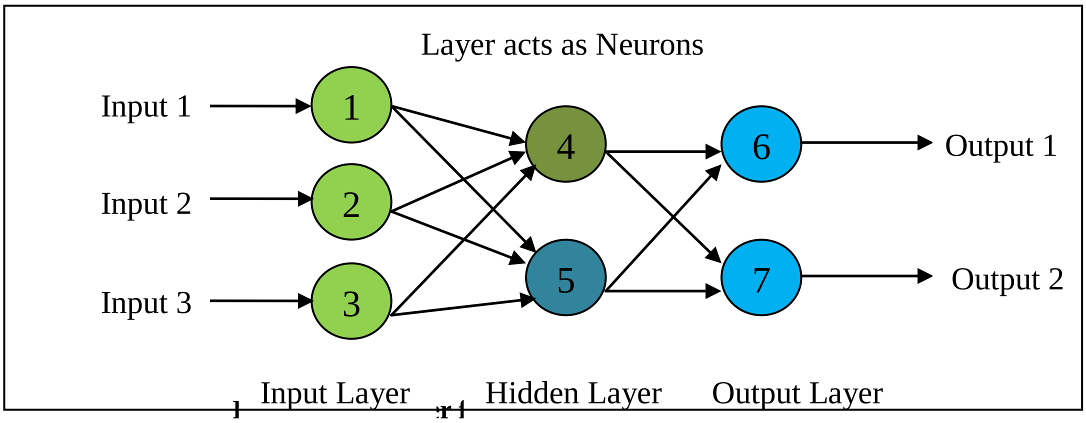
In data mining field, neural networks are used mostly for prediction tasks.Neural networks more accurately called Artificial Neural Networks (ANNs), are computational models. They were originally developed from the inspiration of human brains. Like human brains, Artificial neural networks have been developed eagerly as a subfield of machine learning for decades. There are many kinds of artificial neural networks that have been reported very successful [3,4,5,6,7].

There are two kinds of neural networks based on how the networks are interconnected – feed-forward neural networks and recurrent neural networks [8]. RBFNs are one of the most popular feed-forward networks. RBFNs have three layers including the input layer like multilayer perceptrons, Many researchers have reported successful application of RBFNs [1,2].However, the two most widely used ANNs are a)Feed Forward Networks information flows from the input layer by use of the hidden layers to the final output layer in one direction along connecting pathway fig:2. There is no feedback loops that is the output of any layer does not affect that same or preceding layer. And b)Recurrent networks differ from feed forward network because there is at least one feedback loop. It could exist one layer with feedback connections and also be neurons with self-feedback links that is the output of a neuron is fed back into itself as input.

There are also many other types of networks like Hopefield networks, Pulse networks, and Radial-Basis Function networks. The most important class of neural networks for real world problems solving includes a) Multilayer Perceptron b)Radial Basis Function Networks.

a. Multilayer Perceptron

Multilayer Perceptron is a feed forward artificial neural network model.It has any number of inputs and has one or more hidden layers with any number of units. It uses linear combination functions in the input layers and generally sigmoid activation functions in the hidden layers. It has any number of outputs with any activation function. MLPs are said to be distributed-processing networks because the effect of a hidden unit can be distributed over the entire input space. These techniques are successfully applied to many domains such as finance, medicine, engineering, biology and agriculture



b.Radial basis functions networks

Radial basis functions (RBF) networks are also feed forward, but have only one hidden layer. A RBF network has any number of inputs and has only one hidden layer with any number of units.It uses radial combination functions in the hidden layer, based on the squared Euclidean distance between the input vector and the weight vector and uses exponential activation functions in the hidden layer. It has any number of outputs with any activation function and has connections between the input layer and the hidden layer, and between the hidden layer and the output layer. Gaussian RBF networks are said to be local-processing networks because the effect of a hidden unit is usually concentrated in a local area centered at the weight vector.

c.Logistic Regression

Logistic Regression is a classification method.It returns the probability that the binary dependent variable may be predicted from the independent variables.Maximum Likelihood Estimation is a statistical method for estimating the coefficients of the model.The Likelihood Ratio test is used to test the statistical significance between the full model and the simpler model.

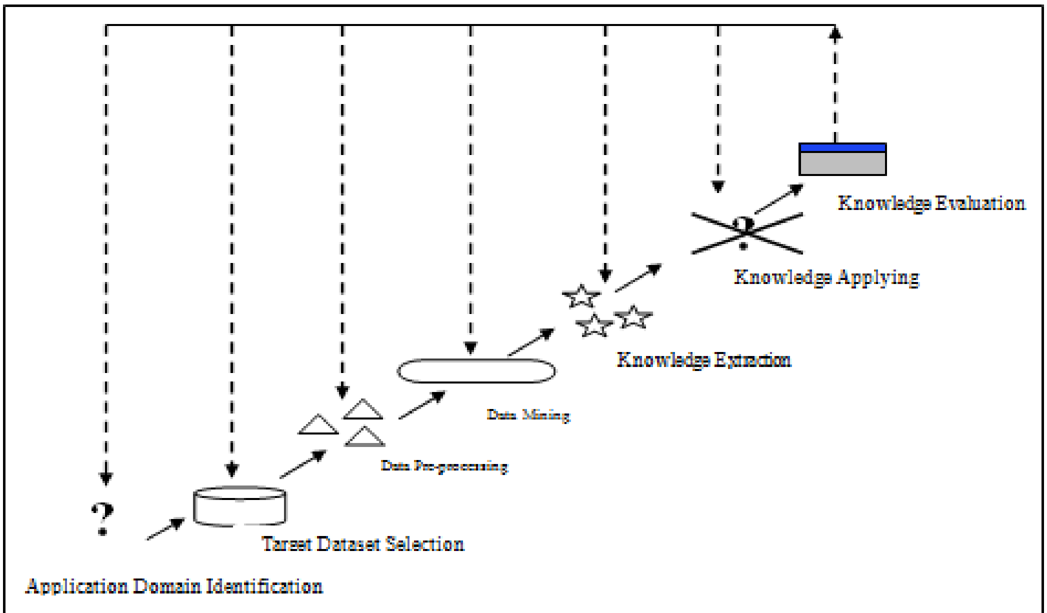


Fig.:3. Data mining phases (Han & Kamber, 2001)

III. DATASET

In UCI repository,all type of dataset was available.All medical dataset such as breast,diabetes,Thyroid,lung cancer and CKD have two classes. In order to predict CKD ,the data obtained from UCI Machine Learning Repository was utilized.CKD dataset consist of nominal and numerical attributes. It contained 24 attributes and 1 class attributes.Dataset implemented using Weka 8.1.The following information is about the CKD dataset .

Title: Chronic Kidney Disease
Source Information

Creator	:	L.Jerlin Rubini
Guided By	:	P.Eswaran
Date	:	July 2015

Number of Instances: 400
Number of Attributes: 25
Class: {CKD, NOTCKD}
Missing Attribute Values: yes
Class Distribution: [63% for CKD] [37% for NOTCKD]
Attribute Information: Table1

S.No	Major Attributes	Data Type
1	Age, Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packaged Cell Volume, WBC count, RBC count.	Numerical
2	Specific Gravity, Albumin, Sugar, RBC,Pus cell, Pus cell clumps, Bacteria,Hypertension, Diabetes Mellitus,Coronary Artery Disease, Appetite, Padal Edema, Anemia and class.	Nominal

Table:1 Data type of attribute

IV. RESULTS

We introduced the concept of confusion matrix, which is presented in Table 2. Where TP is the number of true positives, which means that some cases with 'positive' class (with ckd) is correctly classified as positive; FN, the number of false negatives, which means that some cases with the 'positive' class is classified as negative; TN, the number of true negatives, which means that some cases with the 'negative' class (with notckd) is correctly classified as negative; and FP, the number of false positives, which means that some cases with the 'negative' class is classified as positive.

CONFUSION MATRIX		
	Actual positive (CKD)	Actual negative (not-CKD)
Predicted positive (CKD)	True positive(TP)	False positive(FP)
Predicted negative (Not-CKD)	False negative(FN)	True negative(TN)

Table:2 Format of Confusion matrix for ckd prediction

For example:Table3, In a sample of 400 instances of the Chronic Kidney Disease, 249 (TP) were correctly labeled ,1(FN) were incorrectly labeled, 0(FP) were incorrectly labeled and150(TN) were correctly labeled by MLP classifier. The resulting confusion matrix could look like as below:

	Actual positive (CKD)	**Actual negative (Not-CKD)**
Predicted positive (CKD)	249(TP)	1(FP)
Predicted negative (Not-CKD)	0(FN)	150(TN)

Table:3 Sample format of Confusion matrix for ckd prediction

4.1Performance Analysis Factors

Using 10-fold cross-validation, we can evaluate which model is the most appropriate model for ckd prediction, i.e. provides the highest prediction accuracy. The purpose of this evaluation is to find a suitable method for dataset. In addition to prediction accuracy,Type I error, Type II error, Type I error rate, Type II error rate, Sensitivity, Specificity, F-score and Kappa are also examined as the performance analysis factors. These factors are used to evaluate each classifiers for ckd prediction analysis. The Factors are as follows:

Type I Error

Type I error measures the proportion of non-bankrupt cases which are incorrectly identified as bankrupt ones. Type I error = FP

Type II Error

Type II error measures the proportion of bankrupt cases which are incorrectly identified as non-bankrupt ones. Type II error = FN

Type I Error rate

Type I error rate is the Sensitivity of failed. Type I error rate is important because the misclassification of a failed concern is considered more costly than the misclassification of a non-failed firm. It is calculated using Type I error rate = $\frac{FN}{FN+TN}$

Type II Error rate

Type II Error rate is the Sensitivity of non-failed. Type II error rate is important because the misclassification of a non-failed concern is considered more costly than the misclassification of a failed concern. It is calculated using Type II error rate = $\frac{FP}{FP+TP}$

Sensitivity

It is a statistical measure of the performance of a binary classification test, also known in statistics as classification function. Sensitivity (also called the true positive rate, or the recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Specificity

It (sometimes called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition). These two measures are closely related to the concepts of type I and type II errors.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Accuracy

Accuracy is the percentage of correctly classified instances. It is one of the most widely used classification performance metrics.Overall predictive accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

F-score

It is important to calculate the F-score, defined as the weighted harmonic mean of precision and recall. the best F-score, which was chosen because this is an average measure. Special care was taken against overtraining, because some algorithms can be affected by this effect. Overtraining can be detected if the training has higher accuracy than the prediction.

$$\text{F-score} = \frac{2TP}{2TP+FP+TN}$$

Kappa Test

Measuring the validity with sensitivity, specificity and reliability are done by calculating kappa. To measure agreement between two rater (The expert's analysis and data mining techniques analysis) classifying the same set of cases can be calculated by using of Kappa. Cohen's kappa defines the measure of agreement as the ratio of the percentage of agreement minus the chance agreement to the largest possible non chance agreement. This measure, thus, takes into account the classifications that could match merely by chance. The chance agreement actually depends upon the percentage of matches in each class, and it reduces as the number of classes' increases. Using the above definition, a kappa value of 1 indicates a perfect agreement and a kappa value of 0 indicates that agreement is no better than chance. The formula for calculating Kappa is: Kappa Test = $\frac{\text{Observed Agreement} - \text{Expected Agreement}}{100 - \text{Expected Agreement}}$

where, Observed Agreement = %(Overall Accuracy)

Expected Agreement = (%(TP+FP)* %(TP+FN)) + (%(FN+TN)* %(FP+TN))

V. DISCUSSION

The datasets analyzed by us in this experiment, which is chronic kidney disease. It is a very important issue to accurately predict in medical decision-making. Chronic kidney disease prediction has regarded as a critical topic in medial. Data mining process is the step to select and extract more valuable information in the massive related materials.Table1 shows the content of these datasets. Each dataset are divided into training and testing data by using 10-fold cross validation. The classifier can be measured by the following performance factors such as, overall prediction accuracy, type I error, Type II error, type I error rate, Type II error rate, sensitivity, specificity, F-score and kappa were used .The performance three classifiers is given below table 4:

DATA-MINING CLASSIFIERS	CONFUSION MATRIX		TYPE I ERROR	TYPE II ERROR	TYPE ERRO RATE	TYPE II RRO RATE	SENSITIVITY	SPECIFICITY	ACCURACY	F-SCORE	KAPPA
RBF NETWORK	**244**	**6**	**0**	**2.4**	**1.5**	**0**	**96.15**	**100**	**98.5**	**98.03**	**0.96**
	0	**150**									
MLP	**249**	**1**	**0**	**0.4**	**0.25**	**0**	**99.33**	**100**	**99.75**	**99.66**	**0.99**
	0	**150**									
LOGISTIC REGRESSION	**241**	**9**	**0.25**	**3.6**	**2.25**	**0.66**	**94.30**	**99.58**	**97.5**	**96.7**	**0.95**
	1	**149**									

Table :4 Performance Analysis of Chronic Kidney Disease dataset

VI. CONCLUSION

Chronic kidney disease is one of the leading cause of death and early prediction of chronic kidney disease is important. The computer aided ckd prediction system help the patients as a tool for diagnose of kidney disease. New dataset of chronic kidney disease can be published in the UCI (University of California

Irvine)machine learning repository and evaluated with three classifiers in this paper .There are many kinds of artificial neural networks that have been reported very successful. Among them multilayer perceptron are reported to have better performance than other neural network.The performance of the classifier evaluated and their results are analyzed.Finally,multilayer perceptron classifier gave good accuracy.In the future,ckd can integrate with data preprocess,rules and feature selection with classifiers.

REFERENCES

[1]. G. Baylor, E.I. Konukseven, A.B. Koku, Control of a Differentially Driven Mobile Robot Using Radial Basis Function Based Neural Networks, WSEAS Transcations on Systems and Control, vol. 3, issue 12, pp. 1002-1013, 2008.

[2]. A. Esposito, M. Marinaro, D. Oricchio, S. Scarpetta, Approximation of Continuous and Discontinuous Mappings by a Growing Neural RBF-based Algorithm, Neural Networks, Vol. 13, No. 6, pp. 651-656, 2000.

[3]. O. Buchtala, M. Klimek, B. Sick, Evolutionary Optimization of Radial Basis Function Classifiers for Data Mining Applications, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 35, No. 5, pp. 928-947, 2005

[4]. D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning Internal Representation by Error Propagation, In Parallel Distributed Processing, vol. 1, Rumelhart, D.E., McClelland, J.L. Eds., The MIT Press, 1986.

[5]. J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, In Proceedings of National Academic Science, vol. 79, pp. 2554-2588, 1982. [4] G.A. Carpenter, S. Grossberg, ART-3: Hierarchical Searching Using Chemical Transmitters in SelfOrganizing Pattern Recognition Architectures, Neural Networks, vol. 3, pp. 129-152, 1990.

[6]. G.E. Hinton, T.J. Sejnowski, D.H. Ackley, Boltzmachines: Constraint satisfaction networks that learn, Carnegie-Mellon University, Technical Report CMU-CS-84-119, 1984. [6] K. Fukushima, S. Miyake, T. Ito, Neocognitron: A neural network model for a mechanism of visual pattern recognition, IEEE Transactions on Systems, Man and Cybernetics, vol. 3, no. 5, pp. 826-834, 1983.

[7]. L. Nikolaos, Radial basis Function Networks to Hybrid Neuro-Genetic RBFNs in Financial Evaluation of Corporations, International Journal of Computers, vol. 2, issue 2, pp. 176-183, 2008. [8] A. Hofmann, B. Sick, Evolutionary Optimization of Radial Basis Function Networks for Intrusion Detection, Proceedings of the International Joint Conference on Neural Networks, Vol. 1, pp. 415420, 2003.

[8]. P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison Wesley, 2006