A Survey on Chronic kidney Disease Early Stage Prediction using Classification Techniques

Ankur Sharma[1] Ms. Sonal Arora[2]
[2]Department of Computer Science & Engineering
[1,2]DPGITM,Maharshi Dayanand University, Haryana, India

Abstract—*Chronic kidney disease is one of major cause of death in India.In India every year 7 lac people suffers from Chronic Kidney Disease.They died because of late detection of this disease.Data mining has the capability to analyse and predict this from past data. In this Survey paper we discuss various classification techniques - SVM, ANN, Naive Bayes, Decision Tree etc. and analyse their performance on basis of accuracy, processing time, error etc. We process these technique for prediction of Chronic Kidney Diseases in early disease so treatment could be possible.*

Keywords—CKD, Data Mining, SVM,Classification Techniques, Naive Bayes, ANN

I. INTRODUCTION

Kidney is the important part of our body. Its excrete waste, toxic, and unwanted elements from our body passed by blood. Its main function is to filter the blood and remove these unwanted elements. Kidney helps in regulating blood pressure. It produce some hormones like ErethroProtien (helps in making blood cells). Vitamin D also produces in kidney that helps in absorbing other vitamins, calcium etc for body growth.

Chronic Kidney disease is a type of Kidney disease in which kidney stops functioning properly. Its filtering rate decreases in period of months. Its symptoms are swelling in Feet,blood pressure fluctuates abnormally, deficiency in Vitamin D in body. These symptoms does not appear in starting not even in life of patient. Mostly symptoms are visible when upto 95% kidney damaged. That increase the risk of life of patient. After this there would be only two options left

1) Dialysis
2) Kidney Transplant

Both of these methods are painful and costly in India. In India approx & lac patient suffered from these disease because of no early detection of this disease.

Data Mining is rapidly growing field in computer science. It is used in many domains e.g. financial forecasting, weather forecasting, health care etc. In Health care large volume data is present from which we need to extract useful information so can predict about diseases, discover disease patterns, record disease outbreaks over the world, enhance quality of service and also provide services in area where it's difficult to provide medical diagnosis services. Data Mining is suitable for this task due its cost and efficiency of large data analysis[8].

Rest of our paper includes section of Data mining Classification techniques (section II), current work done in this field (section III),our methodology (section IV) and later discussion and conclusion in section V.

II. CLASSIFICATION TECHNIQUES IN DATA MINING

Classification is a process of separating data into predefined categories(classes). It is supervised learning model. In this class models are created on the basis of training data for which classes are predefined. When model is created it is used for class prediction of unknown data. There are following efficient techniques in classification for CKD prediction.

A. Support Vector Machine (SVM)

SVM is a classifier which separate data in to classes. In supervised learning SVM produce optimal line (in 2d separation) and could produce multi curve plane depends on the nature of data set[8].this line has maximum margin from all classification so there remain space for new cases for learning and classifying without any ambiguity.SVM is implemented in weka by SMO algorithm. SMO algo based on solving series of small quadratic problems,in each iteration only two variable are selected in working set [8]. Fig 1. Shows SVM classification example.

1) Advantage:

- SVM is a deterministic Algorithm
- It separating classes using maximum Marginal hyperplane so more instances can be added and less ambiguous in classification.
- Can learn complex functions easily

2) Disadvantage:

- Computationally extensive
- Suitable for Binary classification

B. Naive Bayes

Naive bayes classifier is based on Bayes theorem, it works on conditional probability. The main assumption for this theorem is that data set attributes must be independent.Navie bays classifier process all features of data set independently.

$$p(ci|d) = p(d|ci)\ p(ci)/\ p(d)$$

Where
p(ci|d) - probability of attribute d that it lies in class ci
p(d|c) - probability of instance d given class is ci

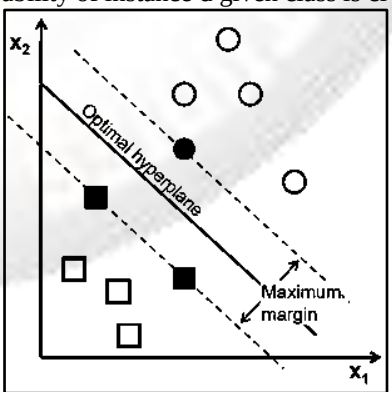

Fig. 1:Line separation between labels by SVM

p(ci) - probability of occurrence of class ci in classes
p(d) - probability of instance d occurring

1) Advantage:

- Yields optimum prediction
- High-speed

– Handle discrete and numeric values
– Easy to compute

2) Disadvantage:

– Computationally intractable
– Independent features assumption violated in real life cases and decrease efficiency

C. Artificial Neural Network (ANN)

This classifier is inspired from human biological nerve system. ANN is a collection of artificial neurons layer that are connected to each other. The training data used to creating mapping from input layer to output layer.

ANN contains mainly two layers 1) Input layer, 2) Output layer. It might be contains hidden layer that mapped the input data with output. Some ANN does not connect output to input called feedforward network. Others are called feed backward networks.

ANN learn and minimise output mapping so predict and provide accurate results. Fig 2 shows an multilayer neural network.

1) Advantage:

– Adaptive Learning
– Self-Organised network
– Fast prediction
– Can handle complex relationships and noisy data

2) Disadvantage:

– All inputs need to be translated to Numeric types
– Training is slow
– Overfitting

D. Genetic Algorithm

Genetic Algorithm based on natural selection and natural genetic. They gives robust searching in complex problems space. In Genetic algorithms Initial population created that represent string presentation of rules. This population than assigned a fit value. After that fittest rules are taken and rest are removed so fittest rules are left as offsprings(new population). Offsprings are creating using genetic operators - mutation and crossovers. GA is used for classification and optimisation problems.

1) Advantage:

– Easy to understand
– Parallelism supported

2) Disadvantage:

– These are slow
– Fitness function should be accurate

E. Decision Tree

Decision tree are broadly used classification technique. It is best suitable for binary classification. It is a tree based data structure contain root, branches, leaf node. Here each node denotes a rule or test for an attribute. Branch denotes the output of test and leaf nodes denotes the label of class in which the data set lies. In decision tree an attribute is chosen so that whole data set can efficiently split and categorised.Some popular decision tree algos are - ID3, C4.5, CART. These algos select attributes on the basis of Gini Index, Information Gain and gain ratio. C5 is successor of C4.5 is also used for decision tree processing. When a new data set is tested, process starts from root node and travers towards leaf node the class label of leaf node is the class of that instance [1]. Fig 3 shows Decision Tree.

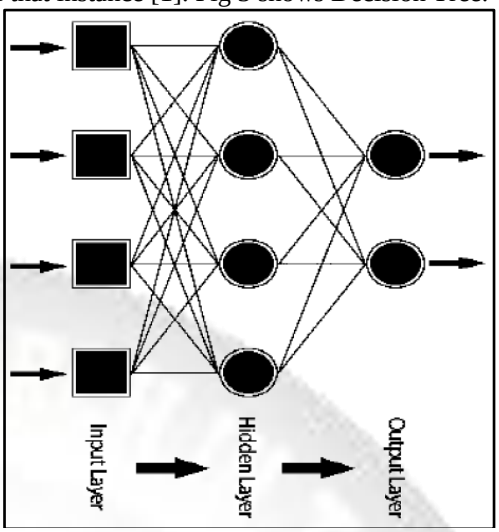

Fig. 2: Feedforward Artificial Neural Network

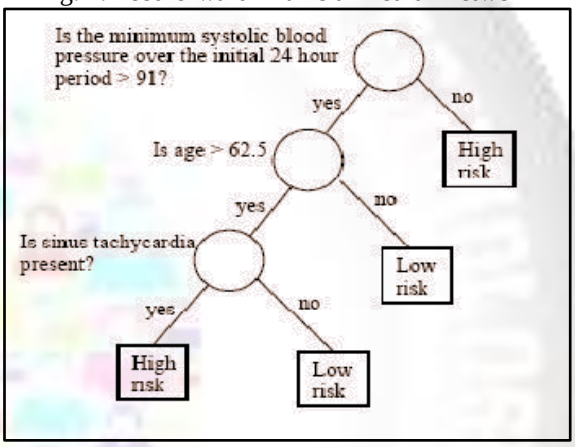

Fig. 3: Decision Tree

1) Advantage:

– Fast learning and prediction
– Less memory required
– Handle high dimension easily
– Easy to interpret
– Can handle numerical as well as categorical values

2) Disadvantage:

– Sometimes tree parts replicate
– Numeric attributes leads to large branching factors
– Restricted one output attribute

III. LITERATURE SURVEY

In March 2018, Mohammad Ashraf Ottom , Khalid M. nahar[8] experiment with Navie bayes classification algo for prediction of CKD, they preprocessed data and use feature selection techniques corelationAttributeEval, CfsSubsetEval. Experience results shows that effficiency of Navie bayes algo increased due to refinment of features using feature selection algos. They used UCI Machine learning Repo data set for their experiment.

In March 2018, W.H.S.D. Gunarathne , K.D.M Perera, K.A.D.C.P. Kahandawaarachchi[9] consider only 14 attributes of CKD data set which are independent of each. They create classification model using Decision Forest tree, Navie Bayes, SVM techniques and found that Multiclass

decision forest algo models giving 99.1% of accuracy with reduced data attribute sets in CKD early stage prediction.

In Nov 2017, S.Dilli Arasu, Dr. R. Thirumalaiselvi[10] have analyze various data mining techniques on Chronic kidney dataset containing 400 instances and having 24 features. They find that preprocessing of data set is necessary. Data mining algo gives less accuracy when trained with original data as compare to preprocessed training data set. They preprocessed data by replacing empty or void values with default values based on observation. The results are also varied according to the tools and techniques used.

A different study[11] carried out with UCI dataset of CKD, They used decision tree, Linear regressing, SVM, Naive Baysen and Neural networks techniques for comparison, they first create two data sets - original data set, preprocessed(filled missing values). Classification algorithms then processed these data set retaining performance criteria Accuracy and sensitivity of algorithms.

They used LR and SVM as feature selection. They used 70% data for training and evaluate test on rest of data. Results state that prediction accuracy of CKD for retaining 8-10 attribute is same for using 24 attributes so we can easily remove dependent features to reduce data storage cost.

IV. METHODOLOGY

A publicly available Chronic Data set is used from UCI Machine learning repository. Table 1 describe all attributes, their types. This is a real Data Set, consist 400 instances, 25 features and donated for research purpose in 2015.

Tools we are using Weka 3.8 best suitable and implements many Data mining algorithms. The attribute and its description are given in table 1.

First we preprocess data set and replace the empty values with default observational value. This preprocessed data set is used for classification experiment.

In Table 2 result show the performance of various classification techniques.Classification techniques - SVM (SVO implementation), ANN (multilayer perceptron), Decision Tree (J48), Naive Bayes are used with default parameters in Weka, all attributes are used for classification.

Data set is divided into two sets 60% are training data sets and rest are training data set.

Attribute Name	Description	Type / Values
Age	Age	Numerical
BP	Blood Pressure (mm/Hg)	Numerical
SG	Specific Gravity	Nominal (1.005, 1.010, 1.015, 1.020, 1.025)
AL	Albumin	Nominal (0-5)
SU	Sugar	Nominal (0-5)
RBC	Red Blood Cell	Nominal (normal, abnormal)
PC	Pus Cell	Nominal (normal, abnormal)
PCC	Pus Cell Clumps	Nominal (present, not present)
BA	Bacteria	Nominal (present, not present)
BGR	Blood Glucose (mgs/dl)	Numerical
BU	Blood Urea (mgs/dl)	Numerical
SC	Serum Creatinine (mgs/dl)	Numerical
SOD	Sodium (mEq/dl)	Numerical
POT	Potassium (mEq/dl)	Numerical
HEMO	Haemoglobin (g)	Numerical
PCV	Packed Cell Volume	Numerical
WC	White Blood cell count (cell/cumm)	Numerical
RC	Red Blood cell count (millions/ cmm)	Numerical
HTN	Hypertension	Nominal (Y/N)
DM	Diabetes Mellitus	Nominal (Y/N)
CAD	Coronary artery disease	Nominal (Y/N)
APPET	Appetite	Nominal (good, poor)
PE	Pedal Edema	Nominal (Y/N)
ANE	Anaemia	Nominal (Y/N)
CLASS	Classification group	Nominal (CKD/notCKD)

Table 1: Attributes of CKD Dataset

Performance Measure	SVM	Naive Bayes	ANN	Decision Tree
Accuracy	96.875	95	98.125	100
Time Taken to Build Model(sec)	0.02	0	2.53	0.02
Precision	0.971	0.956	0.982	1.0
Recall	0.969	0.95	0.981	1.0
F-Measure	0.969	0.951	0.981	1.0

Table 2: Performance Measure with Original Data

In table 3 performance measure with attribute selection algorithm are presented. We use ClassifierAttributeEvaluation feature selection algorithm with best search searching algorithm. Result shows that accuracy and other performance measure are increased using feature selection.

But there is raise in Model building time in SVM and Naive Bayes techniques. Feature selection algorithm finds 16 features independent and rest are dependent on these. So if we remove these than computation process is reduced and also performance in testing new data set increases.

Performance Measure	SVM	Naive Bayes	ANN	Decision Tree
Accuracy	97.5	96.25	98.75	100
Time Taken to Build Model(sec)	0.03	0.02	1.28	0.01
Precision	0.977	0.966	0.988	1.0
Recall	0.975	0.963	0.988	1.0
F-Measure	0.975	0.963	0.988	1.0

Table 3: Performance Measure with Feature Selection

V. CONCLUSION

CKD is a fatal disease by its nature. It causes the other fatal disease like Hypertension, Heart related diseases and toxic

effects in body. Data Mining has much significant in medical domain and steps are taken to predict early stage prediction of CKD so kidney failure can be prevented. Various research is done by different peoples were studied. This paper evaluates Chronic kidney disease prediction using data mining supervised algorithms by WEKA tool.

In this work mainly four algorithms - Decision tree, naive bayes, SVM, ANN are studied and results are analysed. It shows that Decision tree outperforms follows that ANN,SVM , Naive bayes for prediction of this disease.

Time taken to train the model is least in Naive Bayes without feature selection and in Naive Bayes with feature selection. While Artificial Neural network takes high time for training but its accuracy is nearest to Decision tree.

REFERENCES

[1] Sujata Joshi, Mydhili K. Nair, Survey of Classification based prediction techniques in healthcare, April 2018
[2] D.Sindhuja,R. Jemina Priyadarsini, A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder, May 2016
[3] Jyoti Soni,Ujma Ansari ,Dipesh Sharma,Sunita Soni, Predictive Data mining for medical diagnosis: An Overview of heart disease prediction, March 2011
[4] K.Lakshmi, D.Iyajaj Ahmed,G.Siva Kumar , A smart Clinical decision support system to predict Diabetes diseases using Classification Teachniques, 2018
[5] Durga Kinge, S.K. Giakwad, Survey on Data Mining Techniques for diseases prediction, Jan-2018
[6] Angelina Prasanna G, P.C Sakthipriya , A Survey on Cancer Prediction Using Data Mining Techniques, Jan - 2018
[7] Pushpa M. Patil , Review on prediction of chronic disease using data mining techniques, May 2016
[8] Mohmmad Asharf Ottom, Khalid M. Nahar, Computer Aided Diagnosis for Chronic Kidney Diseases using data Mining, April 2017
[9] W.H.S.D. Gunarathne, K.D.M Perera, K.A.D.C.P. Kahandawaarachchi, Performance Evaluation on Machine learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease(CKD), Jan 2018
[10] S.Dilli Arasu, Dr. R. Thirumalaiselvi, Review of chronic kidney disease based on data mining Techniques, Nov 2017
[11] M.S. Gharibdousti, Kamran Azimi, Saraswati Hathikal ,Dae H Won, Prediction of chronic kidney disease Using data mining techniques, Oct 2017