IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Classification of Chronic Kidney Disease using Feature Selection Techniques

A. K. Shrivas [1*], Sanat Kumar Sahu[2]

[1*]Dept. of IT, Dr. C. V. Raman University, Bilaspur (C.G.), India
[2] Dept. of Computer Science, Govt. Kaktiya P.G. College, Jagdalpur (C.G.), India

[*]*Corresponding Author: akhilesh.mca29@gmail.com,*

Available online at: www.ijcseonline.org

Accepted: 26/Apr/2018, Published: 31/May/2018

***Abstract*—** Classification and features selection play very important role to develop robust and computationally efficient model. In this paper, we have compared different classification techniques for classification of chronic kidney disease data. Two supervised classification learning algorithms are used to develop classifiers as Multilayer Perceptron Network (MLPN) and Radial Base Function Network (RBFN). The main focus of this research work is to reduce the number of features using different feature selection technique. We have also used five different classification techniques for select the relevant feature subsets and improve the accuracy of the classification through the Feature Selection Technique (FST). The RBFN classifier achieved the highest average percentage of performance in terms of accuracy. The results shows that both classification techniques given satisfactory accuracy rate in each different selected feature subset.

***Keywords*—**MLP, RBFN,CKD, Feature Selection Techniques

I. INTRODUCTION

The problem of chronic diseases [1] is faced by human being for a long time. Basically a chronic disease is durable three months or more time. Chronic disease is leading causes of disability. This paper focuses on the most common chronic disease called kidney. Chronic kidney disease data classification system can help to reduce manual errors and can assist greatly in examination of data in less and accurate manner. Classification of chronic kidney disease data is beneficial to doctors, pharmacists, medical science, and healthcare personnel.

The task of data analysis, mainly classification [2] methods, is important to support better decision for personalized medicine. That is, decision-making with alertness for those patients can be classified into groups based on their personal characteristics and the samples. Here we have discussed the most common task of predictive analysis for healthcare problems: solving classification problems on clinical data using specialized pre-processing and specialized predictive algorithms. In this work we used machine learning based classification algorithms for identify and classification of chronic kidney disease. The selection of the most excellent medicine for the patients to diagnosis of disease is very challenging task. The medical science tries to better solution for giving better treatment with minimum expenses. In this research work we have used machine learning technique to develop the better classifier and prove the better treatment to the society. Classification is supervised learning techniques to predict the target class accurately for each in the dataset. Medical and clinical researches are used in supervised learning techniques to identify and diagnosis of diseases. In this research work we have used multi layer percpetron and radial basis function network for classification of disease. Feature selection techniques reduces the dimensionality of feature space, eliminate redundant, irrelevant, or noisy data. It also brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and thereof the performance of data mining, and increasing the comprehensibility of the mining results [3]. In this paper we use five ranking algorithms Chi squared, Gain Raito, Information Gain (Info Gain), Relief-F and Symmetric uncertainty for feature selection. In the study we evaluate the accuracy, sensitivity and specificity of classification techniques with different feature subset.

II. RELATED WORK

There are different researcher have used different techniques to identify and diagnosis of disease. [4] Used three different types of classifier such as Back Propagation Neural Network, Radial Basis Function and Random Forest for Chronic Kidney diseases data set. Radial basis function network gives the highest accuracy 85.3%. [5] Has worked on three classification algorithms i.e. naïve bayes, J48 and SMO in

WEKA environment where J48 performs the better to others. [1]used classifier for the predictions task of Chronic Kidney Disease dataset. The names of classifiers are Random Forest (RF) classifiers, Sequential Minimal Optimization (SMO), Naïve Bayes, Radial Basis Function (RBF) and Multilayer Perceptron Classifier (MLPC) and Simple Logistic (SLG). The Random forest performs better than other classifiers.

[6]have suggested two classification techniques like Support vector machine (SVM) and K-Nearest Neighbour (KNN) were used and observed that the performance of KNN classifier is better than SVM. [7] Have developed three prediction models in which decision tree (DT), multilayer perceptron (MLP) and general regression neural network (GRNN) were compared. These models were applied to a real clinical head injury data. The result shows that DT model gives 90.38% prediction average accuracy.

III. METHODOLOGY

This work we have used five ranking based algorithms namely Chi-Squared (CS) attribute evaluation, Gain Ratio (GR) attribute evaluation, Information Gain (IG) attribute evaluation, Relief-F (RF) attribute evaluation, and Symmetrical Uncertainty (SU) attribute evaluation and two classification algorithms namely, Multi layer perceptron, and the radial basis function (RBF) network for classification and computationally increase the performance of classifiers.

Feature Selection

Feature selection is very important technique to remove the irrelevant feature (s) from feature space and select the relevant feature; hence we computationally improve the performance of classifier. In this research work we have used following feature selection techniques:

•**Chi-Squared (CS) attribute evaluation:** Chi Square[3] Test is used in statistics to test the independence of two events. Given dataset concerning two events, be able to get the observed count O and the expected count E. Chi Square Score measures how much the expected counts E and observed Count O derivate from each other. In attribute selection, the two events are occurrence of the feature and occurrence of the class.

•**Gain Ratio (GR) attribute evaluation:** Gain Ratio [3] method has been developed to get the ratio. This method submits to application specific types of normalization with consider to the obtained information called split information.

•**Information Gain (IG) attribute evaluation:** Information gain [3] of an attribute notifies how much information among respect to the classification goal the attribute presents. That is, it measures the difference in information between the cases where you know the value of the attribute and where you don't know the value of the attribute. A common measure for the information is Shannon entropy, although any measure that allows quantifying the information content of a message will do. Information gain related on two things: how much information was available before knowing the attribute value, and how much was available after.

•**Relief-F (RF) attribute evaluation:** A main idea of Relief algorithm is to approximation the quality of attributes according to how fit their values discriminate between instances that are near to each other. Since algorithm Relief cannot handle data sets where there are missing values and noise in the data, and is restricted to problems involving two classes, their extension is created and it`s called ReliefF. ReliefF randomly selects an instance Ri and then searches for k of its nearest neighbors from the same class, called nearest hits and also k nearest neighbors from each of the different classes, called nearest misses. It updates the quality estimation for all features depending on their values for hits and misses [8].

•**Symmetrical Uncertainty (SU) attribute evaluation:** Information theoretic measure called symmetric uncertainty (SU) is used in order to evaluate the worth of constructed solutions. Symmetric uncertainty can be used to calculate the fitness of features for feature selection by calculating between feature and the target class .SU has a number of benefits i.e. it is symmetric in nature therefore SU(i,j) is same as that of SU(j,i) hence it reduces the number of comparisons required where i and j are two features[9].

Classification Technique

Classification is one the important application of data mining and based on supervised learning. Classification techniques are used to develop the classifiers and classify the data into different class level based on features. In this research work we have used two classifier for classification of chronic kidney disease.

•**Multi layer Perceptron:**Multilayer perceptron (MLP) [10], [11] is feed forward ANN model. In these techniques we have one input layer one or more hidden layer and one output layer. An MLP consists of many layers of nodes in a directed graph, among each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. This model output is depending on consequences input data. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

•**Radial Basis Function (RBF) Network:**Radial Basis Function Networks (RBFNs) [11] are type of ANN that has one input layer, one hidden layer and one output layer. The hidden layer calculates the norm of the input from the

neuron. It passes the norm through a non-linear activation function. The linear layer does the linear weighted addition of the outputs of the hidden neurons. This is given as the final output of the system. Each of the neuron in the hidden layer corresponds to a point in the input space. Further each of these neurons has a spread that determines the extent of its influence. Bias may be added as additional inputs. The various parameters of these networks are trained by a training algorithm that normally uses gradient descend rule for the training. This sets the various system parameters and the system is able to give high performance [12].

IV. RESULTS AND DISCUSSION

In this research work we have used Chronic Kidney Disease data set from UCI Machine learning repository benchmarks [13]. The dataset having 25 attributes with 400 instances. Having 11 numeric and 14 nominal (13 + 1 class) attributes. The description of data set as shown in Table 1. Table 2 shows that raking of features from higher to lower with different feature selection technique.

Table 1: Description of Attribute of Chronic Kidney Dataset

S.No.	Attribute	Full Form	Type	Description
1	age	Age	Numerical	age in years
2	bp	Blood Pressure	Numerical	bp in mm/Hg
3	sg	Specific Gravity	Nominal	Sg-1.005,1.010,1.015, 1.020,1.025)
4	al	Albumin	Nominal	al - (0,1,2,3,4,5)
5	su	Sugar	Nominal	su - (0,1,2,3,4,5)
6	rbc	Red Blood Cells	Nominal	rbc - (normal, abnormal)
7	pc	Pus Cell	Nominal	pc - (normal, abnormal)
8	pcc	Pus Cell clumps	Nominal	pcc - (present, notpresent)
9	ba	Bacteria	Nominal	ba - (present, notpresent)
10	bgr	Blood Glucose Random	Numerical	bgr in mgs/dl
11	bu	Blood Urea	Numerical	bu in mgs/dl
12	sc	Serum Creatinine	Numerical	sc in mgs/dl
13	sod	Sodium	Numerical	sod in mEq/L
14	pot	Potassium	Numerical	pot in mEq/L
15	hemo	Haemoglobin	Numerical	hemo in gms
16	pcv	Packed Cell Volume	Numerical	pcv
17	wc or wbcc	White Blood Cell Count	Numerical	wc in cells/cumm
18	rc or rbcc	Red Blood Cell Count	Numerical	rc in millions/cmm
19	htn	Hypertension	Nominal	htn - (yes, no)
20	dm	Diabetes Mellitus	Nominal	dm - (yes, no)
21	cad	Coronary Artery Disease	Nominal	cad - (yes, no)
22	appet	Appetite	Nominal	appet - (good, poor)
23	pe	Pedal Edema	Nominal	pe - (yes, no)
24	ane	Anemia	Nominal	ane - (yes, no)
25	class	Class	Nominal	class - (ckd, notckd)

Table 2: Ranked label of Chronic Kidney disease dataset

Attribute (id)				
CS	**GR**	**IG**	**RF**	**SU**
hemo (15)	sc (12)	hemo (15)	htn (19)	sc (12)
sc (12)	al (4)	sc (12)	dm (20)	hemo (15)
sg (3)	hemo (15)	pcv (16)	sg (3)	al(4)
pcv (16)	htn (19)	sg (3)	rbc (6)	sg (3)
al (4)	sg (3)	al (4)	pc (7)	16 pcv
htn (19)	dm (20)	htn (19)	al (4)	htn (19)
dm (20)	pcv (16)	dm (20)	hemo (15)	dm (20)
rbcc (18)	bu (11)	rbcc (18)	pcv (16)	bu (11)
bu (11)	bgr (10)	bu (11)	appet (22)	bgr (10)
bgr (10)	appet (22)	bgr (10)	pe (23)	rbcc (18)
sod (13)	pe (23)	sod (13)	ane (24)	bp (2)
bp (2)	bp (2)	bp (2)	rbcc (18)	appet (22)
pc (7)	pc (7)	appet (22)	su (5)	pe (23)
appet (22)	ane (24)	pc (7)	pcc (8)	pc (7)
pot (14)	rbcc (18)	pe (23)	wbcc (17)	sod (13)
age (1)	su (5)	pot (14)	cad (21)	rbc (6)
pe (23)	rbc (6)	rbc (6)	bgr (10)	su (5)
rbc (6)	pcc (8)	su (5)	age (1)	ane (24)
su (5)	cad (21)	age (1)	bp(2)	pot (14)
ane (24)	sod (13)	ane (24)	sc (12)	age (1)
wbcc (17)	wbcc (17)	wbcc (17)	bu (11)	wbcc (17)
pcc (8)	ba (9)	pcc (8)	sod (13)	pcc (8)
cad (21)	pot (14)	cad (21)	ba (9)	cad (21)
ba (9)	age (1)	ba (9)	pot (14)	ba (9)

We have applied the data set with and without feature selection technique in two different classifier like Multi Layer Perceptron and RBF network with 10-fold cross data partition. In 10 folds cross validations techniques dataset set is randomly sub divided into ten equal sized partitions. Along with the partitions nine of them are used as training set and the remaining one is used as a test set.

In this section we have presented the classification accuracy, sensitivity, specificity with and without feature selection. Table 3 shows that performance measures of Multi Layer Perceptron and RBF network in terms of Accuracy, Sensitivity and Specificity. We have compare the performance of classifier where performance is varying from one feature subset to others, but RBF network gives better accuracy both with and without feature selection. The RBF classifier gives 99.75% of accuracy in case of Chi Squared. Info Gain and Relief-F feature selection technique.

Table3: Performance measures of classifiers with and without feature selection technique

Multi Layer Perceptron					
S No	Number of feature selected	Name of Features Ranking Technique	Accuracy	Sensitivity	Specificity
1	All		97.75	96.40	100
2	Top 5	Chi Squared	96.75	97.20	96.00
		Gain Ratio	96.75	96.80	96.67
		Info Gain	96.75	97.20	96.00
		Relief-F	97.00	96.80	97.33
		Symmetric uncertainty	96.75	97.20	96.00
3	Top 10	Chi Squared	97.00	97.20	96.67
		Gain Ratio	97.50	97.20	98.00
		Info Gain	97.00	97.20	96.67
		Relief-F	98.75	98.00	100
		Symmetric uncertainty	97.00	97.20	96.67
4	Top 15	Chi Squared	97.00	97.20	96.67
		Gain Ratio	96.75	96.00	98.00
		Info Gain	97.50	97.20	98.00
		Relief-F	98.50	97.60	100
		Symmetric uncertainty	97.50	97.20	98.00
5	Top 20	Chi Squared	98.25	97.20	100
		Gain Ratio	98.25	97.20	100
		Info Gain	98.25	97.20	100
		Relief-F	98.50	97.60	100
		Symmetric uncertainty	98.25	97.20	100

Table4: Performance measures of classifiers with and without feature selection technique

RBF network					
S No	Number of feature selected	Name of Features Ranking Technique	Accuracy	Sensitivity	Specificity
1	All		99.00	98.40	100
2	Top 5	Chi Squared	99.00	98.80	99.33
		Gain Ratio	98.50	98.00	99.33
		Info Gain	99.00	98.80	99.33
		Relief-F	96.50	95.20	98.67
		Symmetric uncertainty	99.00	98.80	99.33
3	Top 10	Chi Squared	**99.75**	99.60	100
		Gain Ratio	99.50	99.60	99.33
		Info Gain	**99.75**	99.60	100
		Relief-F	**99.75**	99.60	100
		Symmetric uncertainty	**99.75**	99.60	100
4	Top 15	Chi Squared	99.50	99.20	100
		Gain Ratio	99.75	99.60	100
		Info Gain	99.50	99.20	100
		Relief-F	99.75	99.60	100
		Symmetric uncertainty	99.50	99.20	100
5	Top 20	Chi Squared	99.00	98.40	100
		Gain Ratio	99.25	98.80	100
		Info Gain	99.00	98.40	100
		Relief-F	99.25	98.80	100
		Symmetric uncertainty	99.00	98.40	100

V. CONCLUSION AND FUTURE SCOPE

In medical science identification and diagnosis of diseases are major role of every doctors and medical students. In this research work we have used five feature selection and two classification techniques to develop the robust classifier and improve the performance of model. Experimental results show that our proposed RBF Network gives satisfactory result for classification of chronic kidney disease with and without feature selection technique. The RBF classifier gives 99.75% of accuracy with Chi Squared. Info Gain and Relief-F feature selection technique in 10 feature subset. In future we will develop new hybrid model with our new feature selection technique to more efficient and improve the performance.

REFERENCES

[1] M. Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm," *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 2, pp. 24–33, 2016.

[2] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, Third. Elsevier, 2012.

[3] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugosl. J. Oper. Res.*, vol. 21, no. 1, pp. 119–135, 2011.

[4] D. N. R. S.Ramya, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, no. 1, pp. 812–820, 2016.

[5] M. Arora and E. A. Sharma, "Chronic Kidney Disease Detection by Analyzing Medical Datasets in Weka," *Int. J. Comput. Appl.*, vol. 6, no. 4, pp. 20–26, 2016.

[6] P. sinha; P. Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," vol. 4, no. 12, pp. 608–612, 2015.

[7] A. Marcano-Cedeño, P. Chausa, A. García, C. Cáceres, J. M. Tormos, and E. J. Gómez, "Data mining applied to the cognitive rehabilitation of patients with acquired brain injury," *Expert Syst. Appl.*, vol. 40, no. 4, pp. 1054–1060, 2013.

[8] D. Oreski and T. Novosel, "Comparison of Feature Selection Techniques in Knowledge Discovery Process," vol. 3, no. 4, pp. 285–290, 2014.

[9] S. I. Ali and W. Shahzad, "A feature subset selection method based on symmetric uncertainty and Ant Colony Optimization," pp. 1–6, 2012.

[10] Sivanandam and Deepa, *Principles of Soft Computing*, Second. wiley, 2014.

[11] S. Haykin, *Neural Networks and Learning Machines*, vol. 3. 2008.

[12] R. Kala, H. Vazirani, N. Khanwalkar, and M. Bhattacharya, "Evolutionary radial basis function network for classificatory problems," *Int. J. Comput. Sci. Appl.*, vol. 7, no. 4, pp. 34–49, 2010.

[13] "UCI Machine Learning Repository of machine learning databases," 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. [Accessed: 01-Jan-2016].

Authors Profile

Dr. Akhilesh Kumar Shrivas is working as Assistant Professor in Department of Information Technology, Dr. C.V. Raman University, Bilaspur, India. He obtained his Master's Degree in Computer Application from Guru Ghasidas Vishwavidyalaya, Bilaspur, India and Ph. D. in Computer Science from Dr. C.V. Raman University, Bilaspur, India. He has 6 year research experience and published more than 50 research papers in reputed journals and conference proceedings and attended workshop and conference at national and international level. His research interests are data mining, soft computing, big data and information security.

Mr. Sanat Kumar Sahu is working as Assistant Professor in Department of Computer Science, Govt. Kaktiya PG College, Jagdalpur (Bastar) Chhattisgarh, India. He obtained his Master's Degree in Computer Application from Guru Ghasidas Vishwavidyalaya, Bilaspur, India and M. Phil in Computer Science from Dr. C.V. Raman University, Bilaspur, India. He has more than 7 years teaching and 02 years research experience. He has published more than 12 research papers in reputed journals and attended workshop and conference at national and international level. His area of interest includes soft computing, machine learning, and data mining.

