International Journal of Research
Available at https://edupediapublications.org/journals
Special Issue on Conference Papers
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 06
March 2018

A Predictive Classifier Studying Behaviour on Chronic Kidney Disease

Dr. M.Sujatha[*1], K. Krishna Chaitanya Rao[#2], YV.Samhitha Chandra[#3], B.Madhav[#4]

[1] *Associate Professor, Computer Science Engineering Dept. , Jyothishmathi Institute of Technology and Science, Karimnagar*
[2,3,4]*B. Tech Student, Computer Science Engineering Department, Jyothishmathi Institute of Technology and Science, Karimnagar*

[1]Dr.Sujathamadugula@gmail.com

***Abstract*— The worldwide population of 10% is influenced on chronic kidney disease. Every year millions die because of major risk factors for kidney disease includes diabetes, high blood pressure, and family history of kidney failure [1]. The medical data has huge amount of patient's information having missing values. Initially, the chronic kidney disease dataset has to be preprocessed for reducing the impurities in the data. This is done by K-NN for imputing missing values. In this paper, t-test statistical method is applied for analyzing behaviour of medical dataset using the SVM classifier.**
Experiments are conducted based on medical dataset for Chronic kidney disease. The results analyzed the behaviour of chronic kidney disease on SVM classifier.

***Keywords*— data preprocessing, missing values, K-NN, SVM.**

I. INTRODUCTION

Preprocessing is the significant issue for extracting knowledge from medical data. Data preprocessing has the data reduction techniques, which removes missing, redundant, irrelevant and noisy values from the data. Missing value exist if any data value is not available for an attribute in the records. A proper approach is to be selected for imputing the missing values to get better quality of database. Imputation is the process of substituting values for missing data [2]. A researcher presents a comparison of imputation techniques such as Mean\Mode, Expectation Maximization, Hot-Deck, and C5.0 for missing data [3].T-test is used to identify In the first Step, a score based on the t-test (named t-score or TS) is calculated for each attribute. In the second step, all the genes are rearranged according to their TSs. The attribute with the largest TS is put in the first place of the ranking list, followed by the gene with the second largest TS, and so on. Finally, only some top genes in the list are used for classification. By applying K-NN algorithm for replacing missing values with a point value might be measured by the values of the points which are near to it, based on remaining attributes [4]. SVM classifier use linear separating hyper plane for misclassification on medical datasets [5].

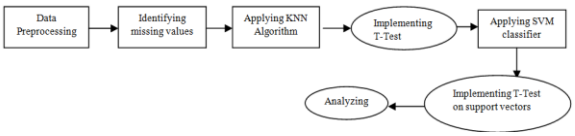
II. RELATED WORK

Welch [6] proposed statistical method called t-test is used to measure large datasets which is a fairly common activity in data analytics is a comparison between two variables. Andrzej Walczak and Michał Paczkowski [7] analyzed medical data on data preprocessing. During preprocessing, medical data is transformed from descriptive, semantic into parameterized form. Andreas Holzinger [8] stated the issues machine learning in data protection, safety, security, privacy and social implications. Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq has studied learning algorithm on C4.5 classifier for predicting the symptoms of the patients on medical data set [9]. Sarab AlMuhaideb and Mohamed El Bachir Menai [10] explained the data preprocessing on medical dataset having missing values, incompleteness and noisy data. He addressed the problem how to select the preprocessing techniques to handle dirty data.

III. METHODOLOGY

Real world data is often highly missing values, *noisy* and *inconsistent* (dirty) as they originate from multiple, heterogeneous sources. In order to go through the solution finding, traditional way is not always suitable because it's a time taking process. Accelerate the data by creating a data supply chain built on a hybrid technology environment. The cause that incurred on Chronic Kidney Disease doesn't come through in a single (instance/attribute). There are many different elements in play so CDK case has to be investigated in a way that total instances has to be consider for prior conditions later on the strength of the attributes are to be preferred. Quality decisions can be drawn only from clean medical data. Hence

data has to be preprocessed to set it free from dirt for knowledge discovery. The missing values are imputed by K-NN algorithm. Then T-test is used to identify or compare the performance between two variables. It can help us to decide whether the difference between the conditions is 'real' or whether it is due merely to chance of fluctuations from one time of testing another. To classify Chronic Kidney Disease patients symptoms by constructing a non-linear classifier using support vector machine. The relevant attribute subset selections are consider for analyzing the behaviour of SVM classifier shown in Fig.1.



A. Fig 1: overview of predictive classifier behaviour *K-NN (Nearest Neighbor)*

In pattern recognition, the k-nn algorithm is instance based learning method used to classify objects based on their closest training examples in the feature space. An object is classified by a majority vote of its neighbors, i.e., the object is assigned to the class that is most common amongst its k-nearest neighbors, where k is a positive integer. In the k-nn algorithm, the classification of a new test feature vector is determined by the classes of its k-nearest neighbors. Here, the k-nn algorithm was implemented using Euclidean distance metrics to locate the nearest neighbor. The Euclidean distance metrics d(x, y) between two points x and y is calculated. Where N is the number of features such that x= {x1,x2,x3,...xN} and y= {y1,y2,y3,...yN}. The number of neighbors(i.e., k) used to classify the new test vector was varied in the range of 1 to 10, and its effects on the classification performance were determined in the form of classification accuracy with standard deviation.

B. *Support Vector Machine (SVM)*

SVM is one of the most effective algorithms in machine learning used for multiclass and biclasification. SVM is a supervised learning model used to analyze the medical data by classification. The goal of SVM is to find the best possible separating hyper plane where the margin separates. The effectiveness of the medical data set is verified by SVM algorithm. Clustered data is then used on the SVM algorithm to classify and split into number of small datasets.SVM has a large margin linear classifier i.e. , linear separable and non-linear separable. Linear and non-linear ways use kernel with limited set of points in many directions. SVM tends to be very good because it should be able to find the linear separation that should exist. SVM classification is based on the pattern matrix; future the kernel function is used. SVM builds a model, each class is mapped by the decision boundary and a hyper plane is specified to separate the different classes. The classification accuracy is obtained by increasing hyper plane margin which increases the distance between classes. The dimensionality of the data set is done effectively using kernel function. The advantage of SVM is one can explicitly control the classifier complexity and error, instead of feature vectors nontraditional data like trees can be used as input to SVM . But to make use of advantages provided, we have to choose the good kernel function.

IV. EXPERIMENTAL

In this paper, the benchmark medical dataset Chronic Kidney Disease (CKD) is used [11]. This CKD dataset has initially 4.7% missing values. During data preprocessing, imputing missing values by using K-NN algorithm on CKD dataset. The Imputed missing values on CKD dataset is applied to SVM classifier for learning algorithm to make prediction on CKD dataset to classify patients into two categories :suffering from disease (*cdk*) and not suffering from disease (*notcdk*).

A. *Description of the Dataset Used*

The CKD has 400 records and 23 attributes containing two classes' cdk and notcdk. The description of benchmark medical dataset CKD is given in Table [2]. The class distribution of CKD dataset is analyzed in Table [1].

TABLE 1.

chronic kidney disease class distribution			
S. No	Class	Distribution	Instances
1	cdk	62.50%	250
2	notcdk	37.50%	150

TABLE 2.

Abbreviations and description of chronic kidney disease			
Attribute number	**Name of attribute**	**Abbreviated attribute name**	**Description**
01	Age	Age	Age
02	Blood pressure	Bp	Mm/hg
03	Specific gravity	Sg	1.005,1.010,1.015,1.020,1.025
04	Albumin	Al	0.1.2.3.4.5
05	Sugar	Su	0.1.2.3.4.5
06	Red blood cells	Rbc	Normal, abnormal
07	Pus cell	Pc	Normal, abnormal
08	Pus cell count	Pcc	Present, not present
09	Bacteria	Ba	Present, not present
10	Blood glucose random	Bgr	Mgs/dl
11	Blood urea	Bu	Mgs/dl
12	Serum creatinine	Sc	Mgs/dl
13	Sodium	Sod	Meq/L
14	Potassium	Pot	Meq/L
15	Hemoglobin	Hemo	Gms
16	Packed cell volume	Pcv	
17	White blood cell count	Wc	Cells/cumm
18	Red blood cell count	Rc	Millions/cmm
19	Hypertension	Htn	Yes, no
20	Diabetes mellitus	Dm	Yes, no
21	Coronary artery	Cad	yes, no
22	appetite	appet	Good, poor
23	class	cdk	notcdk

B. Experimental Analysis

SVM classifier is experimented on CKD medical dataset. In the below Fig 2 gives the highest supporting vectors by considering γ (gamma) values along with supporting vectors with respective to their attributes.

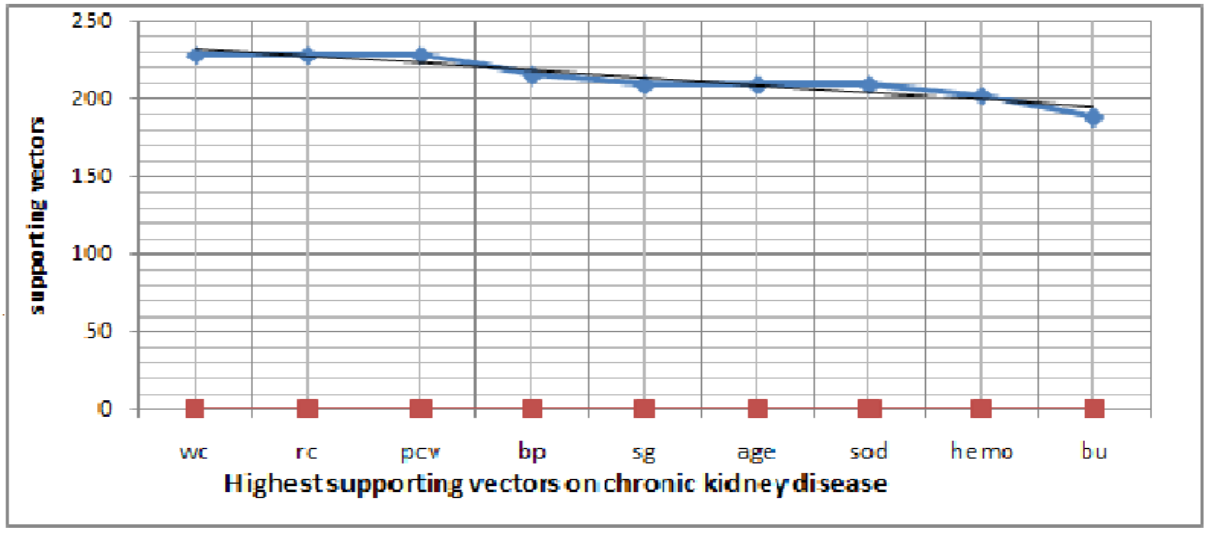


Fig 2: Highest supporting vectors on chronic kidney disease

Fig 3 : Before Training Chronic Kidney Disease Data Set on SVM Classifier

The Fig 3 represents before training of the data, we applied SVM methodology to sort out the highest impacting factors. As it is very much complicated to high dimensional data in order to sort out the best impact factors on ckd SVM uses kernel functions as it creates kernel matrix which summarizes all the data. It scales relatively well to find high supporting vectors, ultimately this analysis lead to attributes of selection.

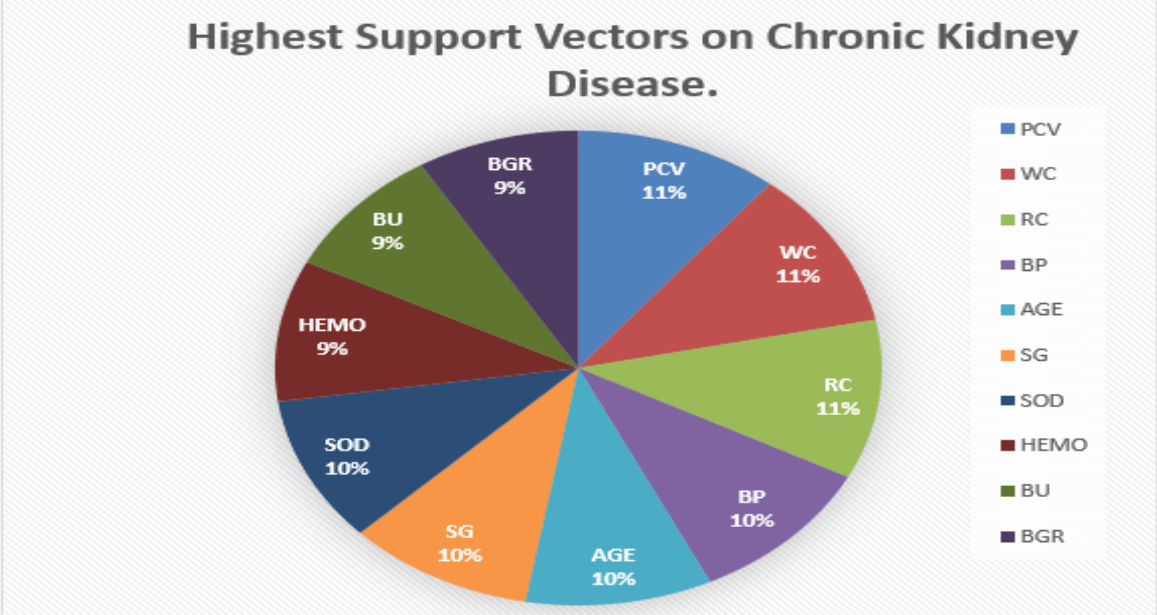


Fig 4 : Highest Support Vectors on Chronic Kidney Disease Along with their Impact Percentages

A. T-Test Analysis

The Fig 5 shows the performance of various evolution parameters. The X-axis denotes Attributes, these are predicted by using SVM algorithm (attribute sub selection), whereas the Y-axis denotes the T-values, those are analyzed by using T-test.

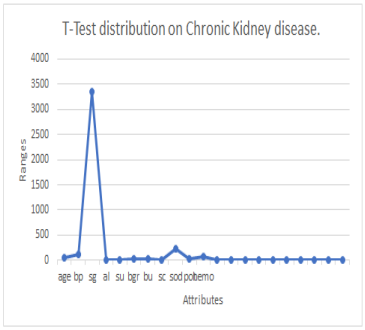


Fig 5 : T-Test Distribution on Chronic Kidney Disease

V. CONCLUSION

In this paper, how t-test distribution is done on Chronic Kidney Disease dataset before training and after training. Handling missing values is one of challenging in data preprocessing on benchmark medical data. Missing values are imputated by applying K-NN algorithm on medical data.Support Vector Machine deals with Chronic Kidney Disease dataset. SVM efficietively perform a non-linear classification on medical data. It is critical to solve the challenges on big data technology in medical data set. Thus to improve patient disease and to cut down wastage of resources in medical data, which might be real scenario for big data studies.

REFERENCES

[1] World Kidney Day: Chronic Kidney Disease. 2015; http://www.worldkidneyday.org/faqs/chronic-kidney-disease/ .
[2] M. Mostafizur Rahman and D.N.Davis, "Machine Learning-Based Missing Value Imputation Method for Clinical Datasets",IAENG Transactions on Engineering Technologies, Pp 245-257.
[3] Tahani Aljuaid and Sreela Sasi, "Proper imputation techniques for missing values in data sets", Data Science and Engineering, 2016
[4] Mai Shouman, Tim Turner and Rob Stocker, "Applying k-

nearest neighbor in diagnosing heart disease patients", International Journal of Information and Education Technology, Vol. 2(3), Pp. 220-223, 2012.

[5] Jianxin Chen, Yanwei Xing, Guangcheng Xi and Jing Chen, "A Comparison of Four Data Mining Models: Bayes, Neural Network, SVM and Decision Trees in Identifying Syndromes in Corona ry Heart Disease," Advances in Neural Networks -ISNN 2007, Lecture Notes in Computer Science, Vol.4491, pp. 1274-1279, 2007.

[6] Welch BL, "The generalization of student's problem when several different population are involved",Biomethika, Vol.34, PP. 28–35, 1947

[7] Andrzej Walczak and Michał Paczkowski ,"Medical data preprocessing for increased selectivity of diagnosis", Bio-Algorithms and Med-Systems, vol.12(1), 2016.

[8] Andreas Holzinger, "A Introduction to machine learning and knoweldge extraction (MAKE)", Machine learning knoweldge extraction,Vol1 (1), pp 1-20, 2017.

[9] Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq ,"Predictive analytics in healthcare system using data mining Techniques",Computer Science and Information Technology-CSCP, Pp. 01-09, 2016.

[10] Sarab AlMuhaideb and Mohamed El Bachir Menai,"An individualized preprocessing for medical data classification",Symposium on Data Mining Applications, Procedia Computer Science Vol. 82 ,Pp.35 – 42, 2016.

[11] https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease#