ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Machine learning Approach of Chronic Kidney Disease Prediction using Clustering Technique

S.Gopika [1], Dr.M.Vanitha [2]

Research Scholar, PG and Research Department of Computer Science, J.J College of Arts and Science (Autonomous), Pudukottai, TamilNadu, India[1]

Assistant Professor, Department of Computer Applications, Alagappa University, Karaikudi, TamilNadu, India[2]

ABSTRACT: Chronic Kidney Disease (CKD) is a gradual decrease in renal function over a period of several months or years. Diabetes and high blood pressure are the most common causes of chronic kidney disease. The main objective of this work is to determine the kidney function failure by applying the clustering algorithm on the test result obtained from the patient medical report. The aim of this work is to reduce the diagnosis time and to improve the diagnosis accuracy using clustering algorithms. The proposed work deals with clustering of different stages in chronic kidney disease according to its severity. The experiment is performed on different algorithms like k-means, k-medoids and Fuzzy C Means. The experimental results show that the Fuzzy c means algorithm gives better result than the other clustering algorithms and produces 87% accuracy.

KEYWORDS: Chronic Kidney Disease (CKD), Clustering, Data mining, Machine Learning (ML), Fuzzy C Means

I. INTRODUCTION

Data Mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

This has led to the emergence of Knowledge Discovery in Databases (KDD) which is responsible for transforming low-level data into high-level knowledge for decision making. Knowledge discovery in databases consists of the list of iterative sequence steps of processes and data mining is one of the KDD processes. Data mining is the application of algorithms for extracting patterns from large volume of data. There is a wealth of data available within the healthcare systems. The healthcare environment is information rich yet knowledge poor. Hence, for healthcare research, data driven statistical research has become a complement. As with the use of computers powered with automated tools the large volumes of healthcare data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for healthcare researchers to identify and exploit patterns and relationships among large number of variables, and also made them able to predict the outcome of a disease using the historical cases stored within datasets.

Clustering is an important area of application for a variety of fields including data mining, knowledge discovery, statistical data analysis, data compression and vector quantization. Clustering has been formulated in various ways in machine learning, pattern recognition, optimization and statistics literature. Clustering is the most common form of

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

unsupervised learning. According to the rule of the unsupervised learning, clustering does not require supervision. No supervision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. The notion of what constitutes a good cluster depends on the application and there are many methods for finding clusters subject to various criteria. These include approaches based on splitting and merging such as ISODATA, randomized approaches such as CLARA, CLARANS, and methods based on neural nets, and methods designed to scale to large databases, including DBSCAN, BIRCH and ScaleKM. Among clustering formulations that are based on minimizing a formal objective function, perhaps the most widely used and studied is partition based algorithms like K-Means, and Fuzzy C-Means clustering. In a partitioned algorithm, given a set of n data points in real d-dimensional space, and an integer k, the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. This measure is often called the squared-error distortion and this type of clustering falls into the general category of variance based clustering.

Chronic kidney diseases have become a major public health problem. Chronic diseases are a leading cause of morbidity and mortality in India. Chronic kidney diseases account for 60% of all deaths worldwide. Eighty percentage of chronic disease deaths worldwide occur in low - and middle-income countries [2]. The National Kidney foundation determines the different stages of chronic kidney disease based on the presence of kidney damage and glomerular filtration rate (GFR), which is measure a level of kidney function. There are five stages of chronic kidney disease.

The remaining paper is organized as follows: Section II discusses literature survey of the research. In section III describes the methodologies used for clustering chronic kidney disease are discussed. Section IV deals with the experiments and its results for parameter measures used in clustering, and the result obtained in each clustering algorithms. Section V describes the conclusion of the proposed work along with the feature enhancement.

II. LITERATURE REVIEW

Mohammed Abdul Khaleel et al. This paper presents a performance analysis on various clustering algorithm namely K-means, expectation maximization, and density based clustering in order to identify the best clustering algorithm for microarray data. Sum of squared error, log likelihood measures are used to evaluate the performance of these clustering methods. This paper conducted an empirical study on various clustering algorithms in order to observe their performance on gene expression data in terms of sum of squared error and log likelihood. In this empirical study, the performance of the clustering algorithms namely density based clustering, expectation maximization clustering and K-means clustering are evaluated on various gene expression data. From this evaluation, it is observed that the performance of expectation maximization clustering algorithm is comparatively better than the density based clustering algorithm in terms of log likelihood [1].

Veerappan et al. This paper analyze the three major clustering algorithms: K-Means, Farthest First and Hierarchical clustering algorithm and compare the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. The result analysis shows that K-means algorithm performs well without inserting the principle component analysis filter as compared to the Hierarchical clustering algorithm and Farthest first clustering since it have less instances of incorrectly clustered objects on the basis of class clustering. Hierarchical clustering as compared to Farthest fast clustering gives better performance. Also this algorithm performs better after merging principle component analysis filter with it. Farthest first clustering though gives a fast analysis when taken an account of time domain, but makes comparatively high error rate. [2]

Lambodar Jena et al. This paper mainly presents an overview of types of clustering techniques and some of the applications of data mining where clustering techniques can be applied. The main goal of clustering is to produce a good and high quality clusters that depends mainly on the similarity measure which has the ability to discover some or all hidden patterns and also make the analysis of data easy. The quality of clusters produced by clustering method is

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

measured by its ability to discover some or all of the hidden patterns. It has been observed that, the most common type of clustering technique that has been used by different applications of data mining is the k-means clustering technique [3].

Abeer et al. In this paper we take diabetes and heart datasets relate with their matching fields then apply the classification algorithm in kidney dataset in software tool finding weather people affected by diabetes are getting chance to get kidney disease or not, output are evaluated as Tested Negative (No Diabetes), Tested Normal(Not affected), Tested High(affected). A new approach for efficiently predicting the diabetes, kidney disease from some medical records of patients. Dataset has designed with matching attributes applied in classification algorithms like J48, Random Tree, Random Forest, REP, Naïve Bayesian algorithm [4].

Jerlin Rubini et al. In this study, we have made a comprehensive comparative analysis of 14 different classification algorithms and their performance has been evaluated by using 3 different cancer data sets. The results indicate that none of the classifiers outperformed all others in terms of the accuracy when applied on all the 3 data sets. Most of the algorithms performed better as the size of the data set is increased. We recommend the users not to stick to a particular classification method and should evaluate different classification algorithms and select the better algorithm. This study focuses on finding the right algorithm for classification of data that works better on diverse data sets. However, it is observed that the accuracies of the tools vary depending on the data set used. It should also be noted that classifiers of a particular group also did not perform with similar accuracies [5].

III. METHODOLOGY

A. Cluster Analysis

The objective of cluster analysis is the classification of objects according to similarities among them, and organizing of data into groups. Clustering techniques are among the unsupervised methods, they do not use prior class identifiers. The main potential of clustering is to detect the underlying structure in data, not only for classification and pattern recognition, but for model reduction and optimization. Various definitions of a cluster can be formulated, depending on the objective of clustering. Generally, one may accept the view that a cluster is a group of objects that are more similar to one another than to members of other clusters. The term "similarity" should be understood as mathematical similarity, measured in some well-defined sense. In metric spaces, similarity is often defined by means of a distance norm.

The proposed system compare K-Means, K-Means++, Fuzzy C-Means clustering algorithms for the chronic kidney disease from UCI machine repository , in these algorithms group the dataset with its matrix values to calculate the PC (Partition Coefficient),CE (Classification Entropy), SC (Partition Index), S (Separation Index), XB (Xie and Beni's Index), DI (Dunn's Index), ADI(Alternative Dunn Index). To compare those algorithms with various validity indices.

B. K-means Algorithm

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.

Algorithm steps

Given the data set X, choose the number of clusters $1 < c < N$.

Initialize with random cluster centers chosen from the data set. Repeat for $l = 1; 2;$

Step 1 Compute the distances

$$D_{ik}^2 = \left(x_k - v_i\right)^T \left(x_k - v_i\right), \qquad 1 \le i \le c, \qquad 1 \le k \le N.$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Step 2 Select the points for a cluster with the minimal distances, they belong to that cluster.

Step 3 Calculate cluster centers

$$v_i^{(l)} = \frac{\sum_{j=1}^{N_i} x_i}{N_i}$$

Until

$$\prod_{k=1}^{n} max \left| v^{(l)} - v^{(l-1)} \right| \neq 0$$

Ending Calculate the partition matrix

C. K-Medoids Algorithm

The *k*Medoidsalgorithm is clustering algorithm related to the *k*-means algorithm and the medoid shift algorithm. Both the *k*-means and *k*-medoids algorithms are partition (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the *k*-means algorithm. A medoids can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. I.e. it is a most centrally located point in the cluster.

Algorithm Steps:

1: Arbitrarily choose k data items as the initial medoids.

2: Assign each remaining data item to a cluster with the nearest medoid.

3. Randomly select a non-medoid data item and compute the total cost of swapping old medoid data item with the currently selected non-medoid data item.

4. If the total cost of swapping is less than zero, then perform the swap operation to generate the new set of k-medoids.

5. Repeat steps 2, 3 and 4 till the medoids stabilize their locations.

D. Fuzzy C-Means (FCM)

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. Straightly speaking, this algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one.

Algorithm steps

1. Randomly select cluster centre
2. Initialize U=[u_{ij}] matrix, $U^{(0)}$

$$u_{ij} = \frac{1}{\sum_{k-1}^{c} \left(\frac{|| xi - cj ||}{|| xi - ck ||} \right)^{\frac{2}{2\pi - 1}}}$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

3. Calculate the u_{ij} using:

4. At k-step: calculate the centre vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i-1}^{N} u_{ij}^{m} \; xi}{\sum_{i-1}^{N} u_y^{m}}$$

5. Update $U^{(k)}$,$U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k-1}^{c} \left(\frac{|| xi - cj ||}{|| xi - ck} \right)^{2m-1}}$$

6. If $|| U^{(k+1)} - U^{(k)} ||< \varepsilon$ or the minimum J is achieved, then STOP; otherwise return to step 2.

7. In the end, you will get a result like:

E. Validity Indices

Different scalar validity measures have been proposed in the literature, none of them is perfect by oneself, and therefore we used several indexes in our Toolbox, which are described below:

1. Partition Coefficient (PC): measures the amount of "overlapping" between clusters.

$$PC(c) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} (\mu_{ij})2$$

Where μ_{ij} is the membership of data point *j* in cluster *i*. The disadvantage of PC is lack of direct connection to some property of the data themselves. The optimal number of cluster is at the maximum value.

2. Classification Entropy (CE): it measures the fuzzyness of the cluster partition only, which is similar to the Partition Coefficient.

$$CE(c) = -\frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij} \log(\mu_{ij})$$

3. Partition Index (SC): is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster.

$$SC(c) = \sum_{i=1}^{c} \frac{\sum_{j=1}^{N} (\mu_{ij})^{m} \; || x_{j-} v_i ||^2}{N_i \; \sum_{k=1}^{c} || v_{k-} v_i ||^2}$$

SC is useful when comparing different partitions having equal number of clusters. A lower value of *SC* indicates a better partition.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

4. Separation Index (S): on the contrary of partition index (SC), the separation index uses a minimum-distance separation for partition validity.

$$S(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} \left(\mu_{ij}\right)^{2} \| x_{j-} v_{i} \|^{2}}{N \min_{i,k} \| v_{k-} v_{i} \|^{2}}$$

5. Xie and Beni's Index (XB): it aims to quantify the ratio of the total variation within clusters and the separation of clusters.

$$XB(c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{N} \left(\mu_{ij}\right)^{m} \| x_{j-} v_{i} \|^{2}}{N \min_{i,j} \| x_{j-} v_{i} \|^{2}}$$

The optimal number of clusters should minimize the value of the index.

6. Dunn's Index (DI): this index is originally proposed to use at the identification of "compact and well separated clusters". So the result of the clustering has to be recalculated as it was a hard partition algorithm.

$$DI(c) = min_{i\epsilon c}\{min_{j\epsilon c, i\neq j}\{\frac{min_{x\epsilon C_i, y\epsilon C_j} d(x,y)}{max_{k\epsilon c}\{max_{x,y\epsilon C^d(x,y)}\}}\}\}$$

The main drawback of Dunn's index is computational since calculating becomes computationally very expansive as *c* and *N* increase.

7. Alternative Dunn Index (ADI): the aim of modifying the original Dunn's index was that the calculation becomes simpler, when the dissimilarity function between two clusters ($min_{x\epsilon Ci, y\epsilon Cj} d(x, y)$) is rated in value from beneath by the triangle-non equality:

$$d(x, y) \geq |d(y, v_j) - d(x, v_j)|$$

Where *vj* is the cluster center of the *j*-th cluster.

$$ADI(c) = min_{i\epsilon c}\{min_{j\epsilon c, i\neq j}\{\frac{min_{x_i\epsilon C_i, x_j\epsilon C_j | d(y,v_j) - d(x_i,v_j)|}}{max_{k\epsilon c}\{max_{x,y\epsilon C^d(x,y)}\}}\}\}$$

IV. EXPERIMENTAL RESUTS

The dataset for diagnosis of chronic kidney disease is obtained from medical reports of the patients collected from UCI machine learning repository. There are 400 instances with 25 different attributes related to kidney disease like PID (patients ID), Age, Gender, Weight, Serum - albumin, Serum - sodium, Blood urea nitrogen, Serum creatinine, Serum uric acid, Sodium urine, Urine urea nitrogen, Urine creatinine, Urine uric acid ,EGFR and Kidney failure. The main contributing attribute to identify the chronic kidney disease is EGFR. Based on the value of the EGFR, the instances are classified as Low, Mild, Moderate, Normal and Severe.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Table 1: Description of Dataset

Dataset	Number of Objects	Number of Attributes	Number of Clusters
Chronic Kidney Disease	400	25	2

A. Validity Measure

Table 2: Clustering Validity Measure for Chronic Kidney Disease Dataset and Algorithm

Dataset	Algorithm	PC	CE	SC	S	XB	DI	ADI
Chronic Kidney Disease	**K-Means**	1	NaN	0.7265	0.0014	3.1784	0.6481	0.5636
	K-Means++	1	NaN	0.7237	0.0014	Inf	0.6481	0.5501
	FCM	0.8088	0.3311	0.9548	0.0018	2.6115	0.64808	0.5471

Table 3: Performance Measure for Chronic Kidney Disease Dataset and Algorithm

Performance Measure	K-Means	K-Medoids	FCM
Sensitivity	0.89	0.73	0.92
Specificity	0.83	0.80	0.87
Accuracy	86%	76.5%	89%

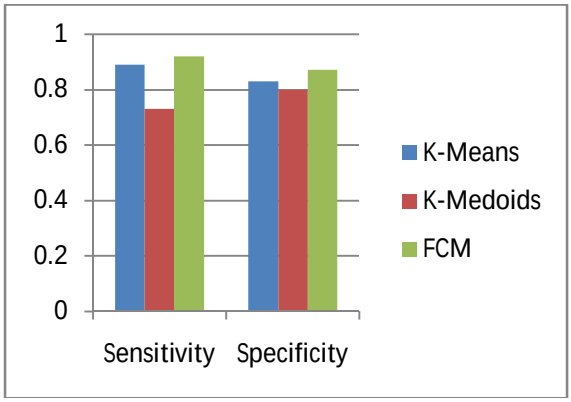

Fig 1.Performance Measure for sensitivity and Specificity

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

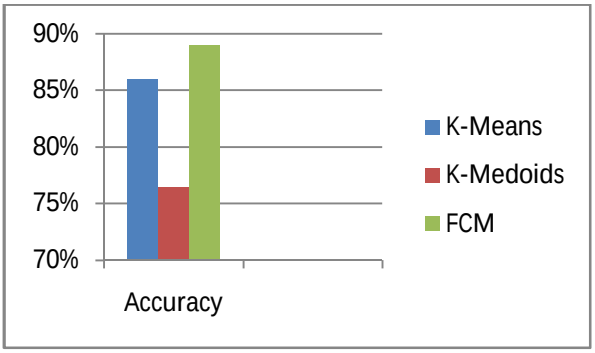

Fig 2.Performance Measure for Accuracy

Prediction of chronic kidney disease is one of the essential topics in medical diagnosis. The proposed work is to cluster the different stages of chronic kidney disease according to its severity. The clustering algorithms that have been considered for predicting chronic kidney disease are k-means, k-medoids and FCM. The models are evaluated with four different measures like Accuracy, Sensitivity and Specificity. From the experimental result, the FCM Function is the better accuracy for predicting chronic kidney disease and it attains the accuracy of 89%.

V. CONCLUSION

Cluster analysis is one of the major tasks in various research areas. The clustering aims at identifying and extract significant groups in underlying data. This is based on a certain clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters. Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed and are available in literature. In the proposed system to analysis the major clustering algorithms such as K-Means, K-Medoids and Fuzzy C-Means with Euclidean distance measure by using Chronic Kidney Disease UCI dataset. It illustrates the efficiency of clustering algorithm with its validity measures. It shows the Fuzzy C-Means clustering algorithm had better than other clustering algorithms. The experimental result shows the performance of the Fuzzy C-Means algorithm was improved significantly.

REFERENCES

[1]Mohammed Abdul Khaleel and Sateesh Kumar Pradham, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases", International Journal of Advanced Research in Computer Science and Software Engineering, 2013, Vol.3, No. 8, pp. 149-153.
[2]Veerappan, Ilangovan and Abraham Georgi," Chronic Kidney Disease: Current Status, Challenges and Management in India", Indian Journal of Public Health Research and Development, 2014, Vol.6, No.1, pp. 1694-1702.
[3]Lambodar Jena and Narendra Ku. Kamila, "Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease", International Journal of Emerging Research in Management and Technology, 2015, Vol. 4, No. 11, pp. 110-118.
[4]Abeer, Ahmad and Majid," Diagnosis and Classification of Chronic Renal Failure Utilizing Intelligent Data Mining Classifiers", International Journal of Information Technology and Web Engineering, 2014, Vol.9, No.4, pp. 1-12.
[5]Jerlin Rubini and Eswaran," Generating Comparative Analysis of Early Stage Prediction of Chronic Kidney Disease", International Journal of Modern Engineering Research, 2015, Vol. 5, No. 7, pp. 49-55.
[6]Ruey Kei Chiu and Renee Yu - Jing," Constructing Models for Chronic Kidney Disease Detection and Risk Estimation", Proceedings of 22nd IEEE International Symposium on Intelligent Control, Singapore, 2011, pp. 166-171.
[7]Jayalakshmi and Santhakumaran, "Improved Gradient Descent Back Propagation Neural Networks for Diagnoses of Type II Diabetes Mellitus", Global Journal of Computer Science and Technology, 2010, Vol. 9, No. 5, pp. 94-97.
[8]Koushal Kumar and Abhishek, "Artificial Neural Networks for Diagnosis of Kidney Stones Disease", International Journal Information Technology and Computer Science, 2012, Vol. 7, No. 3, pp. 20-25.
[9]Rajalakshmi, Neelamegam and Bharathi, "Diagnosis and Classification of Level of Kidney function Using Associative Neural Network and Polynomial Neural Network", Research Journal of Pharmaceutical, Biological and Chemical Sciences, 2013, Vol. 4,No. 4, pp. 724-738.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

[10]Vijayarani and Dhayanand, "Kidney disease prediction using Support Vector Machine and Artificial Neural Network algorithms", International Journal of Computing and Business Research (IJCBR), 2015, Vol. 1, No. 3, pp.1765-1771.
[11]Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2012, Vol. 2, No. 2, pp. 55-66.
[12]Abhinandan Dubey, "A Classification of CKD Cases Using Multi Variant K-Means Clustering", International Journal of Scientific and Research Publications, 2015, Vol. 5, No. 8, pp. 1-5.
[13]Yashpal Singh and Alok Singh Chauhan," Neural Networks in Data Mining", Journal of Theoretical and Applied Information Technology, 2005, Vol.5, No.6, pp. 37-42.
[14]Baylor, Konukseven and Koku," Control of a Differentially Driven Mobile Robot Using Radial Basis Function Based Neural Networks", World Scientific and Engineering Academy and Society Transactions on Systems and Control, 2008, Vol. 3, No. 12, pp. 1002-1013.
[15]Shubham Bind, Arvind Kumar Tiwari and Anil Kumar Sahani, " A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction", International Journal of Computer Science and Information Technologies, 2015,Vol. 6, No. 2, pp.1648-1655.