IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-18, May 2019 E-ISSN: 2347-2693

Assessment of Chronic Kidney Disease using clustering techniques

Debolina Dalui[1], Priyabrata Karmakar[2], Dharmpal Singh[3*], Sonali Bhattacharyya[4]

[1,2,3,4] Dept. of Computer Science and Engineering, JIS College of Engineering, JIS University, Kalyani, India

Corresponding Author: dharmpal.singh@jiscollege.ac.in, Tel.: 7003058392

Available online at: www.ijcseonline.org

***Abstract*—** Data mining is the process of extracting hidden interesting patterns from massive database. It is used to extract the hidden information/knowledge /inference from the real-life database. In this paper an effort has been made to implement the concept of data mining in Chronic Kidney Disease. Chronic Kidney Disease contains heterogeneous data that can be mined properly to provide a variety of useful information for the physicians to detect a disease and predict the severity of the disease and above all survivability of the patients who have this disease. The concepts of clustering and data mining have been used to design the knowledge base for the prediction of chronic kidney disease based on the new data. The concept of factor analysis has been used for the selection of the best factor of this data set and there after the concept of clustering has been used to predict the output of new data element of same data set.

***Keywords*—** *Data mining, clustering, Hierarchical clustering, Chronic Kidney Disease, Clustering, Distance function*

.

I. INTRODUCTION

The Chronic Kidney Disease is termed as CKD. It is a serious disorder of Kidney that affects the functionality of kidney. Data mining is a technique which is very useful in this case. Nowadays, in many areas of medical science, Data mining and clustering techniques are applied. Data mining helps us to extract useful data or information from huge amount of data. There are many other terms for understanding data mining, for example mining of knowledge from databases, knowledge extraction, data analysis, and data archaeology. Basically, in knowledge discovery in database Data mining takes a very important role. There are many useful techniques of data mining that are used in medical domain today e.g. clustering, statistics, machine learning, decision trees, hidden Markova models, genetic algorithm, Meta learning and so on. In this paper, we have used the hierarchical clustering technique for prediction of CKD.

Clustering is basically a grouping technique. It divides a huge number of data set into similar kind of subsets. These subsets are called clusters. We can take decision depending upon those clusters. There are many kinds of clustering techniques like K-means clustering, hierarchical clustering, etc. The criteria of clustering is that the intra distance between two groups should be very large and the inter distance should be less.

K-means is the most familiar name in clustering technique. This centroid-based method takes the input which is k and creates k clusters consisting of n objects. The steps of K-means clustering are (a) divide the objects into some subsets, then (b) mark out the centroid of the clusters, next (c) calculate the distance of each point and (d) mark the centroid which is minimum and finally, (e)when the points are marked, we find the centroid of the new cluster.

Hierarchical clustering is a clustering technique where all data points are assigned to a cluster. Dendrogram, which is a pictorial technique with a tree-like structure, is used in hierarchical clustering.

II. RELATED WORK

Maryam SoltanpourGharibdousti et.al [1] have mainly on applying different machine learning classification algorithm to a dataset for diagnosis kidney disease. Here various classification techniques like e: Decision Tree, Linear Regressing, Super Vector Machine, Naive Bayesian and Neural Network. Obtaining correlation matrix were used. the performance measurements of different methods were calculated and compared to each other.

S.DilliArasu et.al [2] have performed research paper to find the drawbacks of the techniques which were introduced so far. the various data mining techniques were surveyed to predict kidney diseases and major problems were explained. SahanaB J et.al [3] The main objective is that for anticipates Chronic Kidney Disease (CKD) the arrangement methods like Naive Bayes was used and to anticipate the phase of Kidney illness utilizing the Artificial Neural Network (ANN) like C4.5.

SirageZeynu et.al [4] In This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts or practitioners for accurate Kidney disease diagnosis. This research paper intends to use data mining eature selection and data mining classification techniques like k-nearest neighbor (KNN), artificial neural network (ANN), and decision tree. The work also show that feature selection and classification based methods were improved the performance of the algorithm.

Tabassum S et.al [5] This study tries to assist healthcare specialists to early diagnose Kidney disease and assess related risk factors.researchers have the scope topredict kidney disease by using Data mining.

Guneetkauret.al [6] In this paper, KNN (K- nearest neighbor) and SVM (Support Vector Machine) data mining algorithms were used. There also used Big Data which is very effective to store a huge amount of structured data, unstructured data and semi-structured data.

Dr. S. Vijayarani et.al [7] In this paper the problem of distaining and abridging different algorithms of data mining used in the field of medical prediction are discussed. The main aim is on using different algorithms and combinations of several attributes for intelligent and efficient Kidney disease prediction using data mining.

It has been observed that the authors have used data mining concept in different real world optimization problem. Therefore in this paper the concept of data mining and soft computing will be used to form the knowledge base. This knowledge base predicts the value based on the new data. In section 1, implementation and literature work have been furnished where as in section 2 and 3 methodology and implementation have been discussed. In the section 4 result and conclusion have been discussed.

III. METHODOLOGY

1. Multivariate Data Analysis

Multivariate statistics is a form of statistics encompassing the simultaneous observation and analysis of more than one outcome variable.

Multivariate statistics concerns understanding the different aims and background of each of the different forms of multivariate analysis, and how they relate to each other. The practical implementation of multivariate statistics to a particular problem may involve several types of univariate and multivariate analysis in order to understand the relationship between variables and their relevance to the actual problem being studied.

In addition, multivariate statistics is concerned with multivariate probability distributions, in terms of both:

1. How these can be used to represent the distributions of observed data;
2. How these can be used as part of statistical inference, particularly where several different quantities are of interest to the same analysis.

Certain types of problem involving multivariate data, for example simple linear regression and multiple regression, are not usually considered as special cases of multivariate statistics because the analysis is dealt with by considering the (univariate) conditional distribution of a single outcome variable for the r given othe variables.

There are many different models, each with its own type of analysis:

1. Multivariate analysis of variance (MANOVA) extends the analysis of variance to cover cases where there is more than one dependent variable to be analyzed simultaneously.
2. Multivariate regression analysis attempts to determine a formula that can describe how elements in a vector of variables respond simultaneously to changes in others. For linear relations, regression analyses here are based on forms of the general linear model.
3. Principal components analysis (PCA) creates a new set of orthogonal variables that contain the same information as the original set. It rotates the axes of variation to give a new set of orthogonal axes, ordered so that they summarize decreasing proportions of the variation.
4. Factor analysis is similar to PCA but allows the user to extract a specified number of synthetic variables, fewer than the original set, leaving the remaining unexplained variation as error. The extracted variables are known as latent variables or factors; each one may be supposed to account for correlation in a group of observed variables.

1.1 Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called **factors**. In other words, it is possible, that variations in fewer observed variables mainly reflect the variations in total effect. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Computationally this technique is equivalent to low rank approximation of the matrix of observed variables. Factor analysis originated in psychometrics, and is used in behavioral sciences, social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data.

1.2 Clustering

Clustering deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The graphical example of cluster has shown in figure 1.9. In this case it is easily identified the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering.

Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if one object defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Clustering Algorithms

Clustering algorithms may be classified as listed below:

1. Exclusive Clustering
2. Overlapping Clustering
3. Hierarchical Clustering
4. Probabilistic Clustering

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the figure 1.10, where the separation of points is achieved by a straight line on a bi-dimensional plane. On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters. Finally, the last kind of clustering uses a completely probabilistic approach. The mostly used clustering algorithms are:

1. K-means
2. Fuzzy C-means
3. Hierarchical clustering

K-means is an exclusive clustering algorithm, Fuzzy C-means is an example of overlapping clustering algorithm, hierarchical clustering is obvious and lastly. Mixture of Gaussian is a probabilistic clustering algorithm. The distance measure plays important role in clustering which have been discussed in next section.

Distance Measure

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are in the same physical units it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances.

.

Clustering algorithm

K-Means Clustering

K-means (MacQueen, 1967) clustering algorithm is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) with a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point it is needed to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After these k new centroids have been obtained a new binding has to be done between the same data set points and the nearest new centroid. A loop has been formed. As a result of this loop it may be noticed that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, generally a squared error function. The objective function as

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

, where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j.

K-Means Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group of centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find

the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.K-means algorithm is a simple algorithm that has been adapted to many problem domains.

Hierarchical Clustering Algorithm

Hierarchical clustering technique is one of the important clustering methods. There are two basic approaches for generating a hierarchical clustering.

Agglomerative: Start with the points as individual clusters and, at each step, merge the closest pair of clusters. This requires defining a notion of cluster proximity.

Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, it is needed to decide which cluster to split at each step and how the splitting to be done.
Agglomerative hierarchical clustering techniques are by far the most common. A hierarchical clustering algorithm is often displayed graphically using a tree-like diagram called a dendrogram , which displays both the cluster, sub cluster relationships and the order in which the clusters are merged (agglomerative view) or splitted (divisive view). For sets of two-dimensional points, a hierarchical clustering can also be graphically represented using a nested cluster diagram..

Defining Proximity between Clusters

The cluster proximity defines the difference among the various agglomerative hierarchical techniques. Cluster proximity is typically defined with a particular type of cluster in mind, for an example, many agglomerative hierarchical clustering techniques, such as MIN, MAX, and Group Average which have come from a graph-based view of clusters. Min defines cluster proximity as the proximity between the closest two points that are in different clusters, or using graph terms, the shortest edge between two nodes in different subsets of nodes. Alternatively, max, takes the proximity between the farthest two points in different clusters to be the cluster proximity, or using graph terms, the longest edge between two nodes in different subsets of nodes
For similarities, however, where higher values indicate closer points, the names seem reversed. For that reason it is usually preferred to use the alternative names, single link and complete linkage respectively. Another graph-based approach, the group averagetechnique, defines cluster proximity to be the average pair wise proximities (average length of edges) of all pairs of points from different clusters. Figure 1.13 illustrates these three approaches
For a given set of N items to be clustered, and an N*N distance (or similarity) matrix is to be constructed, the basic process of hierarchical clustering is as follows:-

1. Start by assigning each item to a cluster, so that there N items, N clusters are available, each containing just one item. Let the distances (similarities) between the clusters be the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that one less number of clusters is available..
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

IV. IMPLEMENTATION

The available data contains the Chronic Kidney Disease based on the following attributes viz. Age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anaemia, class etc.

The purpose of this work is to correlate the items, so that based on any value of attributes , the value of disease can be estimated. Now it is necessary to check whether the data items are proper or not. If the data items are proper, extraction of information is possible otherwise the data items are not suitable for the extraction of knowledge. In that case pre-processing of data is necessary for getting the proper data.

Data Cleansing

The data cleansing techniques include the filling in the missing value, correct or rectify the inconsistent data and identify the outlier of the data which have been applied on the available data.

It has been observed that each data set does not contain any missing value. The said data item does not contain any inconsistent data i.e. any abnormally low or any abnormally high value. All the data values are regularly distributed within the range of that data items. Therefore the data cleansing techniques are not applicable for the available data.

Data Integration

The data integration technique has to be applied if data has been collected from different sources. The available data have been taken from a single source therefore the said technique is not applicable here.

Data transformation

The data transformation techniques such as smoothing, aggregation, normalization, decimal scaling have to be

applied to get the data in proper form. Smoothing technique has to be applied on the data to remove the noise from the data. Aggregation technique has to be applied to summarize the data. To make the data within specific range smoothing and normalization technique have to be applied. Decimal scaling technique has to be applied to move the decimal point to particular position of all data values. Out of these data transformation techniques, smoothing and decimal scaling techniques have been applied on the data

The methods of factor analysis has been applied on the available input data to make a decision to select the optimal model for the formation of association rule because Apriori Algorithm is not applicable for the numeric data.

Using the values of antecedent item, the correlation of coefficient of the items have been computed and these coefficients have been stored in table 1 termed as correlation matrix.

Table 1: Correlation matrix

1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.1586	-0.228	0.032	0.1837	-0.0521	0.0858	-0.1654	-0.0492	0.2224	0.191	0.1307	-0.045
0.1586	1	-0.199	0.075	0.1931	0.0911	0.1618	-0.0671	-0.118	0.1474	0.1829	0.1463	-0.089
-0.228	-0.199	1	-0.127	-0.14	0.1041	-0.1271	0.1203	0.0826	-0.2477	-0.3184	-0.317	0.3377
0.0318	0.0753	-0.127	1	0.2513	0.3424	0.5296	-0.3943	-0.3725	0.271	0.3128	0.1475	-0.194
0.1837	0.1931	-0.14	0.251	1	0.0447	0.194	-0.1536	-0.1128	0.625	0.119	0.0896	-0.105
-0.052	0.0911	0.1041	0.342	0.0447	1	0.3742	-0.1243	-0.1652	0.0855	0.1959	0.0874	-0.051
0.0858	0.1618	-0.127	0.53	0.194	0.3742	1	-0.5429	-0.3285	0.2567	0.3379	0.1591	-0.176
-0.165	-0.067	0.1203	-0.394	-0.154	-0.1243	-0.5429	1	0.2831	-0.197	-0.1879	-0.049	0.2209
-0.049	-0.118	0.0826	-0.373	-0.113	-0.1652	-0.3285	0.2831	1	-0.0856	-0.1458	-0.041	0.0392
0.2224	0.1474	-0.248	0.271	0.625	0.0855	0.2567	-0.197	-0.0856	1	0.1267	0.0746	-0.187
0.191	0.1829	-0.318	0.313	0.119	0.1959	0.3379	-0.1879	-0.1458	0.1267	1	0.5794	-0.112
0.1307	0.1463	-0.317	0.148	0.0896	0.0874	0.1591	-0.0491	-0.0405	0.0746	0.5794	1	-0.353
-0.045	-0.089	0.3377	-0.194	-0.105	-0.0511	-0.1755	0.2209	0.0392	-0.1866	-0.1115	-0.353	1
0.0501	0.0613	-0.002	0.096	0.1635	-0.0053	0.1448	0.0307	0.0047	0.037	0.3464	0.1963	0.1641
-0.174	-0.285	0.5514	-0.414	-0.147	-0.1428	-0.4071	0.2639	0.2018	-0.2648	-0.5354	-0.35	0.3004
-0.211	-0.298	0.5223	-0.407	-0.166	-0.1443	-0.4043	0.2685	0.1863	-0.271	-0.5223	-0.354	0.3117
0.064	-0.013	0.0247	0.16	0.1262	0.0607	0.1083	-0.1512	-0.1117	0.0812	0.0153	-0.044	0.0937
-0.149	-0.2	0.5448	-0.359	-0.142	-0.0761	-0.3161	0.2591	0.1475	-0.2153	-0.3929	-0.3	0.3798
-0.392	-0.269	0.46	-0.364	-0.248	-0.0352	-0.2902	0.1998	0.1016	-0.3741	-0.393	-0.281	0.1943
-0.359	-0.238	0.4573	-0.266	-0.431	-0.0218	-0.1829	0.1717	0.0697	-0.502	-0.3216	-0.21	0.2405
-0.241	-0.094	0.1978	-0.195	-0.212	-0.0676	-0.1595	0.1852	0.1674	-0.2102	-0.2266	-0.192	0.0809
0.1589	0.184	-0.244	0.298	0.0497	0.1133	0.2977	-0.1881	-0.1571	0.1705	0.2789	0.1622	-0.163
-0.095	-0.061	0.2747	-0.403	-0.11	-0.1124	-0.3637	0.105	0.1156	-0.0989	-0.3365	-0.173	0.0964
-0.052	-0.204	0.2871	-0.238	-0.029	-0.0772	-0.2765	0.1719	0.055	-0.1181	-0.4414	-0.236	0.1439

14	15	16	17	18	19	20	21	22	23
0.05	-0.1735	-0.211	0.064	-0.1489	-0.392	-0.359	-0.2413	0.1589	-0.095
0.061	-0.285	-0.298	-0.0132	-0.1997	-0.2686	-0.2381	-0.0939	0.184	-0.061
-0.002	0.5514	0.5223	0.0247	0.5448	0.46	0.4573	0.1978	-0.244	0.2747
0.096	-0.4139	-0.407	0.1599	-0.3585	-0.3635	-0.2657	-0.1954	0.2979	-0.403
0.164	-0.1467	-0.166	0.1262	-0.1417	-0.2479	-0.4307	-0.2118	0.0497	-0.11
-0.005	-0.1428	-0.144	0.0607	-0.0761	-0.0352	-0.0218	-0.0676	0.1133	-0.112
0.145	-0.4071	-0.404	0.1083	-0.3161	-0.2902	-0.1829	-0.1595	0.2977	-0.364
0.031	0.2639	0.2685	-0.1512	0.2591	0.1998	0.1717	0.1852	-0.188	0.105
0.005	0.2018	0.1863	-0.1117	0.1475	0.1016	0.0697	0.1674	-0.157	0.1156
0.037	-0.2648	-0.271	0.0812	-0.2153	-0.3741	-0.502	-0.2102	0.1705	-0.099
0.346	-0.5354	-0.522	0.0153	-0.3929	-0.393	-0.3216	-0.2266	0.2789	-0.337
0.196	-0.3503	-0.354	-0.0437	-0.3001	-0.2807	-0.21	-0.1921	0.1622	-0.173
0.164	0.3004	0.3117	0.0937	0.3798	0.1943	0.2405	0.0809	-0.163	0.0964

1	-0.0939	-0.117	-0.0695	-0.0784	-0.0533	-0.044	-0.0124	-0.029	-0.057
-0.094	1	0.8847	-0.0235	0.761	0.572	0.4668	0.2617	-0.408	0.3952
-0.117	0.8847	1	-0.0389	0.7604	0.5606	0.4649	0.2699	-0.422	0.4155
-0.07	-0.0235	-0.039	1	0.105	-0.0742	-0.1033	0.0028	0.1026	-0.067
-0.078	0.761	0.7604	0.105	1	0.4504	0.3901	0.2278	-0.372	0.3642
-0.053	0.572	0.5606	-0.0742	0.4504	1	0.6224	0.3319	-0.346	0.3796
-0.044	0.4668	0.4649	-0.1033	0.3901	0.6224	1	0.2795	-0.33	0.3048
-0.012	0.2617	0.2699	0.0028	0.2278	0.3319	0.2795	1	-0.155	0.1741
-0.029	-0.4082	-0.422	0.1026	-0.372	-0.3463	-0.33	-0.1545	1	-0.428
-0.057	0.3952	0.4155	-0.0674	0.3642	0.3796	0.3048	0.1741	-0.428	1
-0.097	0.5421	0.4934	-0.0175	0.3538	0.3553	0.1915	0.0436	-0.253	0.2093

The eigen value and Eigen vector (table 3) of the elements using the above correlation matrix have been calculated using Mat lab Tools. The contribution of eigen value of each item among all other items has been calculated and have been furnished in table 2.

Table 2 Eigen Value

Data Attribute	**Eigen value**	**Percentage contribution**	**Data Attribute**	**Eigen value**	**Percentage contribution**
1	0.1085	0.45	12	0.6359	2.65
2	0.208	0.87	13	0.6698	2.79
3	0.2523	1.05	14	0.7818	3.26
4	0.2703	1.13	15	0.8267	3.44
5	0.3025	1.26	16	0.9546	3.98
6	0.3746	1.56	17	1.011	4.21
7	0.3879	1.62	18	1.0619	4.42
8	0.4373	1.82	19	1.2074	5.03
9	0.4843	2.02	20	1.4214	5.92
10	0.5123	2.13	21	1.8633	7.76
11	0.591	2.46	22	1.9853	8.27

From the table2, it has been observed that percentage contribution of the item which have less than 2.5 % as compared to other items therefore that items have been ignored. The eigen vectors of the items have been furnished in table 3

Table 3: Eigen Vectors

1	2	3	4	5	6	7	8	9	10	11
-0.038	0.0032	-0.0552	0.0927	0.0577	-0.095	0.0253	0.056	0.448	0.0527	0.0622
-0.016	-0.064	-0.0341	-0.0325	0.1573	-0.103	0.0122	-0.032	-0.002	0.0445	-0.262
-0.047	-0.099	-0.0626	-0.0841	0.0473	0.4521	-0.041	-0.404	0.225	0.5006	0.1156
0.0175	0.0387	-0.0753	0.2692	0.1599	-0.58	0.0039	-0.053	0.005	0.4906	-0.232
-0.01	0.1328	0.255	0.0877	-0.482	-0.054	0.25	-0.156	0.33	-0.112	0.0215
-0.006	0.0233	-0.0399	-0.216	-0.138	-0.033	0.1872	0.162	-0.071	-0.145	0.1274
0.0307	-0.136	0.0858	0.5031	-0.1	0.2855	-0.406	0.167	-0.022	-0.281	-0.144
0.0215	-0.121	0.1292	0.3615	0.0518	0.1137	-0.252	0.233	0.244	0.1291	0.0216
-0.015	0.0334	0.0288	0.1291	0.026	-0.223	-0.028	-0.059	-0.017	0.0209	-0.089
-0.009	-0.092	-0.1447	-0.2762	0.4829	-0.049	-0.391	0.015	0.007	-0.165	0.217
0.0289	-0.108	0.5522	0.0673	0.3925	-0.004	0.2177	-0.294	-0.106	-0.171	0.0448
-0.01	0.0011	-0.4284	-0.0266	-0.258	-0.09	-0.288	-0.223	0.014	-0.047	-0.02

0.0283	-0.046	-0.2247	0.0606	-0.166	-0.245	-0.157	-0.349	-0.147	-0.301	0.0469
-0.034	0.0389	-0.0973	-0.1756	0.007	0.0618	0.0528	0.41	-0.115	0.22	0.1323
0.7424	-0.357	0.0272	-0.0745	-0.037	-0.128	0.0933	0.136	-0.075	-0.035	-0.099
-0.658	-0.46	0.059	-0.0082	-0.067	-0.166	0.0977	0.16	-0.152	-0.069	-0.126
-0.013	-0.127	0.0207	-0.0426	0.0479	0.0545	-0.043	0.127	0.082	0.0437	0.0345
-0.026	0.718	0.0106	0.0911	0.134	0.0553	-0.015	0.201	-0.16	-0.061	-0.18
-0.022	0.0254	-0.3555	0.3536	0.3447	-0.083	0.3405	-0.031	0.311	-0.297	0.3157
-0.025	0.1047	0.3533	-0.3308	-0.042	-0.34	-0.392	0.095	0.468	-0.092	0.0379
0.0001	0.0378	0.0261	-0.0562	-0.066	-0.011	-0.041	-0.241	-0.043	-0.033	-0.046
-0.012	0.0603	0.0904	0.0312	-0.138	-0.156	-0.018	0.091	-0.126	0.0566	0.6342
0.0426	-0.017	0.1526	0.2925	-0.125	-0.14	-0.149	-3E-04	-0.294	0.2529	0.4237
-0.054	0.1467	0.1929	0.0078	0.1112	0.0052	-0.215	-0.3	-0.217	-0.022	0.0351

12	13	14	15	16	17	18	19	20	21	22	23
0.0961	0.3983	-0.145	0.405	0.056	-0.21	-0.4292	-0.2093	0.0793	-0.309	0.1106	-0.1308
-0.248	-0.22	-0.283	-0.107	-0.34	-0.643	0.1697	0.0285	0.0949	-0.07	0.0694	-0.1374
0.0523	-0.213	0.0235	0.204	-0.09	0.01	0.0235	-0.0288	0.1741	0.0409	-0.3063	0.2335
0.0891	0.0721	0.0144	-0.068	-0.03	0.193	0.0608	0.0246	-0.0115	0.0665	-0.3741	-0.2258
-0.005	-0.129	-0.044	-0.151	-0	0.096	0.2448	0.2213	0.2019	-0.478	-0.1207	-0.1409
-0.09	0.3801	0.3012	0.183	-0.29	0.053	0.0824	0.1555	0.0953	0.1638	-0.3598	-0.0841
-0.088	-0.061	-0.137	0.261	0.024	-0.043	0.0067	0.0896	0.0365	0.1192	-0.3936	-0.2217
0.1508	0.1899	0.0931	-0.322	-0.31	0.215	0.2324	-0.1504	0.1739	0.0143	0.3383	0.1633
-0.424	-0.221	0.152	0.584	0.14	0.252	0.2706	-0.0257	0.0599	-0.041	0.3446	0.1142
0.1985	-0.046	0.0854	0.056	-0	0.096	0.2767	0.2045	0.0324	-0.459	-0.0839	-0.1789
0.1566	-0.032	0.0362	-0.014	0.148	0.04	-0.1408	-0.0034	0.3899	0.2084	0.0659	-0.2482
0.0906	-0.104	-0.147	-0.164	0.229	0.156	-0.2732	0.2826	0.2558	0.1613	0.2001	-0.1801
0.0375	0.0319	0.2449	-0.022	-0.18	-0.142	0.0479	-0.5167	0.3632	-0.032	-0.1113	0.1526
-0.254	0.0943	-0.224	0.003	0.05	0.048	0.063	0.0419	0.667	0.07	0.0227	-0.0581
0.1101	-0.092	-0.209	0.068	0.015	0.096	-0.1115	0.033	0.0727	-0.157	-0.1111	0.3276
0.1946	-0.101	-0.154	0.049	0.025	0.08	-0.0941	0.0394	0.0612	-0.137	-0.1057	0.3271
-0.346	0.0185	0.1193	-0.333	0.605	-0.057	0.0654	-0.3902	-0.0303	-0.144	-0.2558	-0.033
0.2902	-0.173	-0.051	0.047	0.096	-0.027	-0.0619	-0.0898	0.1716	-0.143	-0.15	0.2874
-0.163	-0.059	-0.111	-0.12	0.029	8E-04	0.0398	0.1997	0.0047	0.1631	-0.1163	0.2809
-0.066	-0.124	0.015	-0.015	0.011	-0.102	-0.0761	0.0891	0.0189	0.3407	-0.1189	0.2522
0.2052	0.4746	-0.44	0.111	0.295	-0.109	0.5431	0.0063	-0.0423	0.1779	0.0165	0.1535
0.0901	-0.284	-0.403	0.08	-0.14	0.105	0.0479	-0.3399	-0.1656	0.0765	-0.042	-0.2066
0.0187	0.0885	0.2638	0.002	0.178	-0.447	-0.0582	0.3575	0.0303	-0.118	0.061	0.2059
-0.476	0.2956	-0.304	-0.16	-0.22	0.266	-0.2721	0.0702	-0.0631	-0.239	-0.0882	0.2021

The major factors have been calculated as per the formula (√eigen value × eigen vector) using the selected eigen value as furnished in table 2 and eigen vector as furnished in table 3. The major factors have been furnished in table 4

Table 4: Cumulative Effect Value of Items

1	A	0.950081	13	M	0.897164
2	B	0.878449	14	N	0.854671
3	C	0.869754	15	O	0.701793
4	D	0.869907	16	P	0.717143
5	E	0.907753	17	Q	0.895845
6	F	0.710776	18	R	0.750706
7	G	0.858651	19	S	0.832696
8	H	0.858746	20	T	0.842711

9	I	0.930361	21	U	0.95645
10	J	0.893516	22	V	0.895782
11	K	0.834006	23	W	0.914607
12	L	0.774567	24	X	0.91561

Now a relation has been formed by using the cumulative effect value of all the elements to produce total effect value.
Total effect value = (0.32) × A + (1.17) × B + (1.22) × C. ...Now using the relation, a resultant total effect value formed (due to size of data it has not shown here)

Implementation by Clustering:
The Concept of hierarchical clustering has been used using the Matlab command to form the cluster on the total effect to select the optimal cluster for the formation of knowledge base.
The command is furnished as follows:
Step 1 :
X = [total effect value];
Step 2 :
Use this we can define the whole data set as a matrix. Then pass it to pdist.
Y = pdist(X)
Step 3:
The pdist is a function which calculates the distance between every object.
Z = linkage(Y)
The linkage function is a function which generates a hierarchical cluster tree, then the linkage information is returning in a matrix Z. After that when hierarchical cluster tree is created we will partition the data into the clusters.
Step 4 :
At first, we create two cluster, then 4,8,16onwards.
T = cluster(Z,2)
T = cluster(Z,4)
```
T = cluster(Z,8)
```

V. RESULTS AND DISCUSSION

Then we calculated the total distance of each cluster and we found that the total distance for the 4th cluster (T = cluster(Z,16)) is less as furnished in table 5

Table 5. Distance of cluster based on number of cluster

Cluster no	Total distance
2	893.3095
4	450.0309
8	262.3342
16	115.5645

After that based on the cluster number 16, we have related the item with the sixteen cluster and then predict the output based on it. The error has been found as 1.45 %.

VI. CONCLUSION AND FUTURE SCOPE

This research work is based on clustering in health care which has been developed by using data mining techniques and clustering techniques. The concept of hierarchical Clustering for prediction of kidney disease. has been used here and prediction has been done based on it. It has been observed that, hierarchical clustering has given 1.5 % as error.

The concept of K-means and fuzzy c means will be used to minimize the error in our future work

ACKNOWLEDGMENT

Authors are grateful towards JIS College of Engineering for providing lab and related facilities for carrying out the research activities.

REFERENCES

[1] Maryam SoltanpourGharibdousti, Kamran Azimi, Saraswathi Hathikal, Dae H Won, "Prediction of Chronic Kidney Disease Using Data Mining Techniques", In the Proceedings of Industrial and Systems Engineering Conference, in the year 2017.
[2] S.DilliArasu, Dr.R.Thirumalaiselvi , "Review of Chronic Kidney Disease based on Data Mining Techniques", International Journal of Applied Engineering Research , Volume 12, Number 23 (2017) pp. 13498-13505 , ISSN 0973-4562
[3]Sahana B J, Dr Minavathi, "Kidney Disease Prediction Using Data Mining Classification Techniquesand ANN", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 4, April 2017.
[4] SirageZeynu, Shruti Patil, "Survey on Prediction of Chronic Kidney Disease Using Data Mining Classification Techniques and Feature Selection"International Journal of Pure and Applied Mathematics, Volume 118, No. 8 2018, 149-156, ISSN: 1311-8080
[5] Tabassum S, Mamatha Bai B G, Jharna Majumdar, "Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques", International Journal of Engineering Research in Computer Science and Engineering, Vol 4, Issue 9, September 2017, ISSN (Online) 2394-2320
[6] GuneetKaur, Ajay Sharma, "PREDICT CHRONIC KIDNEY DISEASE USING DATA MINING ALGORITHMS IN HADOOP", International Journal of Advances in Electronics and Computer Science, Volume-5, Issue-4, Apr.-2018, ISSN: 2393-2835
[7]Dr.S.Vijayarani,Mr.S.Dhayanand, "DATA MINING CLASSIFICATION ALGORITHMS FOR KIDNEY DISEASE PREDICTION", International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015
[8] D. P., J. P. Choudhury and M. De, "An Enhance DE algorithm for analysis in data set", International Journal of Data Science (IJDS), 2016 (Accepted).
[9] Dharmpal Singh, Abhishek Banerjee, Gopal Purkait , "Assessment of Heart DiseaseTypes Using Clustering Techniques Under the Domain

of Data Mining", International Journal of Artificial Intelligence and Knowledge Discovery, Vol.6 Issue 03 pp. 10-16, Print ISSN: 2231-2021 e-ISSN: 2231-0312 .
[10] D. P. Singh, S. Sahana, SK Saddam Ahmed, "A Comparative Study to Assess the Crohn's Diseasetype using Statistical and Fuzzy Logic Methodology", International Journal of Computer Sciences and Engineering, Volume-04, Issue-06, Page No (41-46), Aug -2016, E-ISSN: 2347-2693
[11] D. P. Singh, J. P. Choudhury and M. De, "An effort to select a preferable meta heuristic model for knowledge discovery in Data mining", International Journal of mataheuristics, Vol. 4 No. 1, pp.57-90, September, 2015. 1755-2184, ISSN print: 1755-2176 DBLP Index
[12] Sk. Saddam and D. P.Singh, "Genotype based classification of Crohn's disease using various BPN training algorithms", International Journal of Artificial Intelligence and Knowledge Discovery, Vol. 5 No. 2, pp.22-29, July, 2015, Print ISSN: 2231-2021 e-ISSN: 2231-0312.
[13] D. P. Singh, J. P. Choudhury and M. De, "An Effort to Compare the Clustering Technique on Different Data Set Based On Distance Measure Function in the Domain of Data Mining", International Journal of Artificial Intelligence and Knowledge Discovery, Vol. 5, No. 1, pp. 1-8,January, 2015, Print ISSN: 2231-2021 e-ISSN: 2231-0312.
[14] D. P. Singh, J. P. Choudhury and M. De, "A Comparative Study to Select a Soft Computing Model for Developing the Knowledge Base of Data mining with Association Rule For mation by Factor Analysis", International Journal of Artificial Intelligence and Knowledge Discovery, Vol. 3, No. 3, pp.18-23, October, 2013, Print ISSN: 2231-2021 e-ISSN: 2231-0312.
[15] D. P. Singh, J. P. Choudhury and M. De, "An Effort to Developing the Knowledge Base in Data Mining by Factor Analysis and Soft Computing Methodology", International Journal of Scientific & Engineering Research (IJSER), Vol. 4, No. 9, pp. 1912-1923, September, 2013, ISSN 2229-5518.
[16] D. P. Singh, J. P. Choudhury and M. De, "A comparative study on the performance of Fuzzy Logic, Bayesian Logic and neural network towards Decision Making" International Journal of Data Analysis Techniques and Strategies (IJDATS), Vol. 4, No. 2, pp. 205-216, April, 2012. SSN online: 1755-8069 ISSN print: 1755-8050, Scopus Index
[17] D. P. Singh, J. P. Choudhury and M. De, "A Comparative Study to Select a Soft Computing Model for Knowledge Discovery in Data Mining", International Journal of Artificial Intelligence and Knowledge Discovery, Vol. 2, No. 2, pp. 6-19, April, 2012.
[18] D.P. Singh, J.P. Choudhury and M. De, "A Comparative Study on the performance of Soft Computing models in the domain of Data Mining," International Journal of Advancements in Computer Science and Information Technology, Vol. 1, No. 1, pp. 35-49, September, 2011, ISSN 2277-9140.
[20] D.P. Singh, J.P. Choudhury and M. De, "Optimization of Fruit Quantity by comparison between Statistical Model and Fuzzy Logic by Bayesian Network", PCTE, Journal of Computer Sciences, Vol. 8, No.1, Punjab, pp. 91-95, June-July, 2010.
[21] D. P. Singh, J. P. Choudhury and M. De, "Prediction Based on Statistical and Fuzzy Logic Membership Function", PCTE, Journal of Computer Sciences, Vol. 8, No. 1, Punjab, pp. 86-90, June-July, 2010.
[22] D.P. Singh, J. P. Choudhury and M. De, "Performance Measurement of Neural Net Work Model Considering Various Membership Functions under Fuzzy Logic", International Journal of Computer and Engineering, Vol. 1, No. 2, pp. 1-5, 2010, ISSN-0976-9587.
[23] D. P. Singh, J. P. Choudhury, "Assessment of Exported Mango Quantity by Soft Computing Model", International Journal of Information Technology and Knowledge Management, Kurukshetra University, Vol. 2, No. 2, pp. 393-395, June-July, 2009, ISSN: 0973-4414.
[24] D.P Singh, "A Modified Bio Inspired BAT algorithm," International Journal of Applied Metaheuristic Computing (IJAMC), Vol. 9 No. 1, pp. 60-77, 2018. Scopus Index.
[25] D. P. Singh, J. P. Choudhury and M. De, "A modified ACO for classification on different data set" International Journal of Computer Application, Vol.123, No. 6, pp-39-52, August 2015, ISSN 0975-8887.
[26]D. P. Singh, J. P. Choudhury and M. De, "Performance measurement of Soft Computing models based on Residual Analysis" International Journal for Applied Engineering and Research, Vol. 6, No. 5, Delhi, India, pp. 823-832, Jan-July, 2011, Print ISSN 0973-4562. Online ISSN 1087-1090. Scopus Index.

Authors Profile

Dharmpal Singh received his Bachelor of Computer Science and Engineering and Master of Computer Science and Engineering from West Bengal University of Technology. He has done his PhD in year 2015. He has about 12 years of experience in teaching and research. At present, he is with JIS College of Engineering, Kalyani, and West Bengal, India as an Associate Professor and Head of the department. He has published 32 papers in referred journal and conferences index by **Scopus, DBLP and Google Scholar and editorial team and senior membe**r of many reputed journal index by **SCI, Scopus, DBLP and Google Scholar**.
He has organized **seven national levels Seminar/Workshop, published three patent** and has applied for the **AICTE Project in year of 2019.**
He is a member of the Computer Science Teachers Association (CSTA), Computer Society of India (CSI) and also a member of International Association of Computer Science and Information Technology (IACSIT).

Debolina Dalui is a 2nd year student, persuing Master of Technology in the department of Computer Science and Engineering, from a renowned engineering college of West bengal.

Sonali Bhattacharyya received her Bachelor of Computer Science and Engineering and Master of Computer Science and Engineering from a renowned engineering college of West bengal. At present, she is with JIS College of Engineering, Kalyani, and West Bengal, India as an Associate Professor. she is a member of the Computer Science Teachers Association (CSTA), Computer Society of India (CSI)

Priyabrata Karmakar is a 2nd year student, persuing Master of Technology in the department of Computer Science and Engineering, from a renowned engineering college of West Bengal.

