European Journal of Molecular & Clinical Medicine

ISSN 2515-8260 Volume 07, Issue 11, 2020

Performance evaluation of triumvirate clustering algorithms for heart disease prediction

Dr.V.Poornima[1] &Sarala Devi. U[2]

[1] Assistant Professor,New Prince ShriBhavani Arts and Science College, Chennai

[2]Assistant Professor,PrinceShriVenkateshwara Arts and Science College, Chennai.

E-mail id: poornimasudhaagar@yahoo.com

Abstract

Now a day's heart disease is the dominant reason for deaths far and wide.Vast number of people annually suffers from heart malfunction worldwide. A heart patient shows several symptoms and it is very tough to attribute them to the heart disease in so many steps of disease progression. Data mining, as an answer to extract a hidden pattern from the clinical dataset, are applied to a database in this analysis. Clustering is an important means of data mining based on separating data categories by similar features. This study showed that how to obtain the clusters and how to determine the new centroid using K-means,EM and Farthest first algorithms.The k-means algorithm is one of thewidely recognized clustering tools that are applied in a varietyof scientific and industrial applications. The EM technique is similar to the K-Means technique.Instead of assigning examples to clusters to maximize the differences in means for continuous variables, the EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions.Farthest first algorithm is suitable for the large dataset but it creates the non-uniform cluster. Finally this study examined the performance of these three algorithms.The dataset of 303 people were collected from Cleveland dataset of UCI machine learning.

Keywords: Data Mining Algorithm, Clustering Techniques, Heart Disease Prediction, Weka Tool

1. Introduction

Among all harmful disease, heart attacks diseases are considered as the most universal. Medical practitioners conduct so many surveys on heart diseases and collect information of heart patients, their disease progression and symptoms. World Health Organization (WHO) estimated 17.5 million people deceased from cardiac diseases worldwide in 2012 [1]. Prediction of HD can reduce the calamitous rate of human. Information technology plays a crucial role in Health field [2].Data mining is the technique of withdrawing hidden information from a large set of database. It helps researchers gain both profound insights of unprecedented understanding and novel of large medical datasets[3].The process of data mining is composed of selecting,analyzing, preparing, applying, interpreting and evaluating theresults[4]. Among many data mining techniques, clustering technique is considered as one of the efficient and popular data mining techniques and it clusters or groups the data items based on their similarity [5]. This paper is an attempt to present the detailed study about the three different clustering algorithms such as using

K-means, EM and Farthest first algorithms were analyzed to predict the heart disease. The Cleveland dataset is employed to these three algorithms and their clusters are estimated.To examine the outcome of these algorithms, it was implemented using weka tool.

2. Related work

Clustering is one of the important tasks for data analysis exploration which aims to find data structures which have intrinsic state by modifying the data objects into similar groups and the representation of data in classes, for this reason it is called unsupervised classification or learning performed by observation [6]. A cluster is therefore a collection of objects which are 'similar' between each other and are 'dissimilar' to the objects belonging to other clusters[7].The main goal of clustering analysis is to group both similar and different objects in the same clusters and different clusters respectively. In clustering, objects in a cluster are identical to one another yet dissimilar to object in other clusters. The semantic of the classes is not known beforehand in clustering techniques[8].

K-Means Clustering is a method of cluster analysiswhich aims to partition n observations into k clusters in whicheach observation belongs to the cluster with the nearest mean.Each cluster is assigned a random target number of clusters-kand started from a random initialization. The proposedtechnique classifies the group of the objects based onattributes into K number of groups. The grouping is done byminimizing the sum of squares of distances between datausing Euclidean distance formula and the correspondingcluster centroid. The research result shows that the integrationof clustering gives promising results with highest accuracyrate and robustness[9].EM and K-means are similar in the sense that they allow model refining of an iterative process to find the best congestion. However, the K-means algorithm differs in the method used for calculating the Euclidean distance while calculating the distance between each of two data items; and EM uses statistical methods. The EM algorithm is often used to provide the functions more effectively.

3. Clustering Algorithms

The categorization of objects into various groups or thepartitioning of data set into subsets so that the data in each ofthe subset share a general feature, frequently the proximitywith regard to some defined distance measure [10], is knownas Clustering. The clustering problem has been identified innumerous contexts and addressed being proven beneficial inmany medical applications. Clustering the medical data intosmall with meaningful data can aid in the discovery ofpatterns by supporting the extraction of numerous appropriatefeatures from each of the clusters thereby introducingstructure into the data and aiding the application ofconventional data mining techniques [11]. Several similarity and dissimilarity measures are applied to find the relationships and patterns which exist in those data items. Many types of clustering algorithms are available; they are, hierarchical algorithms, partitioning algorithms, density based, grid based and distance based algorithms.

3.1 K Means Clustering Algorithm

The k-means algorithm is one of thewidely recognized clustering tools that are applied in a varietyof scientific and industrial applications[12]. k-means groups thedata in accordance with their characteristic values into kdistinct clusters. Data categorized into the same clusterhave identical feature values. k, the positive integer denotingthe number of clusters, needs to be provided in advance. Thesteps involved in a k-means algorithm are given subsequently:
Prediction of heart disease using K – Means clusteringtechnique

1. K points denoting the data to be clustered are placed intothe space. These points denote the primary groupcentroids.
2. The data are assigned to the group that is adjacent to thecentroid.
3. The positions of all the K centroids are recalculated assoon as all the data are assigned.
4. Steps 2 and 3 are reiterated until the centroids stop movingany further. This results in the segregation of data intogroups from which the metric to be minimized can bedeliberated [13].

The preprocessed heart disease data is clustered using the Kmeansalgorithm with the K values. Clustering is a type ofmultivariate statistical analysis also known as cluster analysis,unsupervised classification analysis, or numerical taxonomy.K-Means clustering generates a specific number of disjoint,flat (non-hierarchical) clusters. It is well suited to generatingglobular clusters. The K-Means method is numerical,unsupervised, non-deterministic and iterative.

3.2 EM Algorithms

The EM (expectation maximization) technique is similar to the K-Means techniqueThe concept of the EM algorithm stems from the Gaussian mixture model (GMM). The GMM method is one way to improve the density of a given set of sample data modeled as a function of the probability density of a single-density estimation method with multiple Gaussian probability density function to model the distribution of the data.The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data (distribution). In general, to obtain the estimated parameters of each Gaussian blend component if given a sample data set of the log-likelihood of the data, the maximum is determined by the EM algorithm to estimate the optimal model. Principally, the EM clustering method uses the following algorithm:

Input: Cluster number *k*, a database, stopping tolerance.

Output: A set of *k*-clusters with weight that maximize log-likelihood function.

1. Expectation step: For each database record *x*, compute the membership probability of *x* in each cluster $h = 1,\ldots, k$.
2. Maximization step: Update mixture model parameter (probability weight).
3. Stopping criteria: If stopping criteria are satisfied stop, else set $j = j + 1$ and go to (1).

In the analytical methods available to achieve probability distribution parameters, in all probability the value of the variable is given. The iterative EM algorithm uses a random variable and, eventually, is a general method to find the optimal parameters of the hidden distribution function from the given data, when the data are incomplete or has missing values.[14]
The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data (distribution). Put another way, the EM algorithm attempts to approximate the observed distributions of values based on mixtures of different distributions in different clusters. The results of EM clustering are different from those computed by k-means clustering. The latter will assign observations to clusters to maximize the distances between clusters. The EM algorithm does not compute actual assignments of observations to clusters, but classification probabilities. In other words, each observation belongs to each cluster with a certain probability.

3.3 Farthest First Algorithms

The farthest-first traversal k-center (FFT) is a fast and greedy algorithm [15]. In this algorithm k points are first selected as cluster centers. The first center is select randomly. The second center is greedily select as the point furthest from the first. Each remaining center is determined by greedily selecting the point farthest from the set of already chosen centers, and the remaining points are added to the cluster whose center is the closest. In the following, we present a method for selecting the first point deterministically.
For each $Xi = [xi,1, xi,2, \ldots, xi,m]$ in D that is described by m categorical attributes, we use $f(x_{i,j}|D)$ to denote the frequency count of attribute value xi,j in the dataset. Then, a scoring function is designed for evaluating each point, which is defined as:
Score $(X_i) = \Sigma^{n} f(x_{i,j}|D)$.

1. Farthest first traversal(D: data set, *k*: integer) {
2. randomly select first center;
3. //select centers
4. for (I= 2,…,*k*) {
5. for (each remaining point) { calculate distance to the current center set; }
6. select the point with maximum distance as new center;}
7. //assign remaining points
8. for (each remaining point) {
9. calculate the distance to each cluster center;
10. put it to the cluster with minimum distance;}}

In farthest first algorithm, the clusters are not uniform or the objects are grouped in a single cluster massively.

4. Experimental Results

Three representatives of the clustering algorithms are the K Means, Expectation Maximization (EM) algorithmand Farthest first Algorithms.These three algorithms were applied for heart disease prediction dataset by using Waikato Environment for Knowledge Analysis (WEKA) [16]

and the performance of these algorithms were estimated. Here, for experimentation, the dataset given in the UCI machine learning repository such as, Cleveland is subjected to analyze the performance of the system.Experimental data for Cleveland dataset is consist of 303 instancesand 14 attributes[17]and each attribute is defined in Table 1.The dataset of cleveland has been converted as arff file and it has been sent to the Weka tool to predict the HD using different clustering algorithms.

Table 1:Data set

S no	Input variables	Description	Options
1	Age	Age in years	Continuous value
2	Sex	1 = male, 0= female	Male, female
3	Cp	Chest pain type	Chest pain type. Values from 1 to 4. 1: typical angina. 2: atypical angina. 3: non-anginal pain. 4: asymptomatic.
4	Trestbps (blood pressure)	Resting blood pressure in mmHg	Continuous value in mmHg
5	Chol (cholesterol)	Serum cholesterol in mm/dL	Continuous value in mm/dL
6	Fbs (fasting blood sugar)	Fasting blood sugar in mg/dL	Fasting blood sugar attributes value "1" for greater than 120 mg/dL, else the attribute value is 0 (false). Value 1 = true. Value 0 = false.
7	Restecg (ECG)	Electrocardiographic results (ECG result)	Resting electrocardiographic results value ranging from 0 to 2. 0: normal. 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV). 2: showing probable or definite left ventricular hypertrophy
8	Thalach (heart rate)	Maximum heart rate	Continuous value categorized into normal and abnormal
9	Exang	Exercise Induced angina	Exercise induced angina. Values from 0 to 1. Value 1 = yes. Value 0 = no.
10	Old peak	ST depression induced by exercise relative to rest	Continuous values
11	Slope	Slope of the peak exercise ST segment	Measure of slope for peak exercise. Values can be 1, 2, or 3. Value 1: up sloping. Value 2: flat. Value 3: down sloping.
12	Ca	Number of major vessels colored by fluoroscopy	Number of major vessels from 0 to 3
13	Thal	Heart rate of patient	Represents heart rate of the patient. It can take values 3, 6, or 7. Value 3 = normal. Value 6 = fixed defect. Value 7 = reversible defect.
14	Class	Class labels (predicted outcome)	Contains a numeric value between 0 and 1. Each value represents heart disease or absence of disease. Value 0: absence of heart disease. Value 1: presence of heart disease.

K Means

Number of iterations: 6

Within cluster sum of squared errors: 367.8302453852061

Initial starting points (random):

ISSN 2515-8260 Volume 07, Issue 11, 2020

Cluster 0: 65,0,3,140,417,1,2,157,0,0.8,1,1,3,0
Cluster 1: 61,1,4,148,203,0,0,161,0,0,1,1,7,2
Final cluster centroids:

		Cluster#	
Attribute	**Full Data**	**0**	**1**
	(303.0)	(159.0)	(144.0)
Age	54.4389	52.8931	56.1458
Sex	0.6799	0.478	0.9028
chestpaintype	3.1584	2.8302	3.5208
Restbloodpressure	131.6898	130.0503	133.5
serum	246.6931	247.7799	245.4931
Fastingbloodsugar	0.1485	0.1195	0.1806
Reselectro	0.9901	0.8931	1.0972
Maxheartrate	149.6073	157.7673	140.5972
exercise	0.3267	0.1258	0.5486
Oldpeak	1.0396	0.6038	1.5208
Slope	1.6007	1.4025	1.8194
Majorvessels	0.6854	0.3628	1.0417
thal	4.7342	3.099	6.5398
Class	0.9373	0.2201	1.7292

Time taken to build model (full training data): 0.05 seconds
Clustered Instances
0 159 (52%)
1 144 (48%)

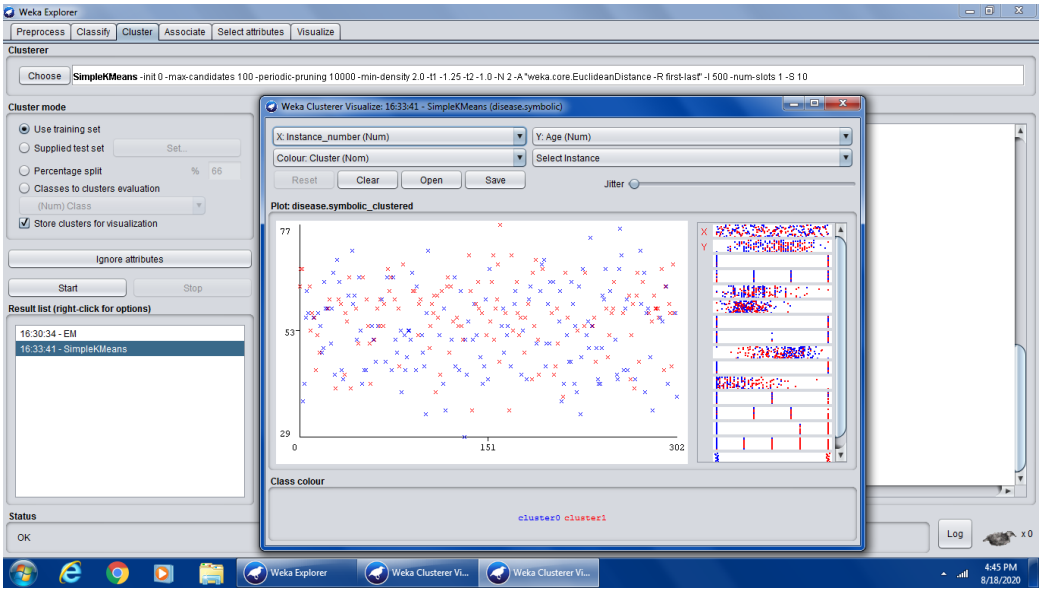

Figure 1: K Means clustering result plot

EM
==

Number of clusters selected by cross validation: 3
Number of iterations performed: 0

	Cluster		
Attribute	0	1	2
	(0.34)	(0.37)	(0.29)
Age			
mean	56.6311	52.1786	54.75
std. dev.	7.9212	9.2719	9.3799
Sex			
mean	0.8641	0.5893	0.5795
std. dev.	0.3444	0.4942	0.4965
Chestpaintype			
mean	3.835	2.8482	2.7614
std. dev.	0.487	0.8925	1.0394
Restbloodpressure			
mean	134.233	128.4286	132.8636
std. dev.	18.461	16.7023	17.2335
serum			
mean	251.1748	236.1607	254.8523
std. dev.	51.4801	45.9076	57.2363
Fastingbloodsugar			
mean	0.1748	0.125	0.1477
std. dev.	0.3816	0.3322	0.3569
Reselectro			
mean	1.2136	0	1.9886
std. dev.	0.9666	0.995	0.1066
Maxheartrate			
mean	134.6214	157.6518	156.9091
std. dev.	20.8689	20.3756	19.3127
exercise			
mean	0.7573	0.0982	0.1136
std. dev.	0.4308	0.2989	0.3192
Oldpeak			
mean	1.8019	0.6009	0.7057
std. dev.	1.2986	0.787	0.9293
Slope			
mean	1.9126	1.3839	1.5114
std. dev.	0.5259	0.5734	0.625
Majorvessels			
mean	1.2427	0.4079	0.3864
std. dev.	1.0334	0.7643	0.7339
thal			
mean	6.444	3.9375	3.747
std. dev.	1.2536	1.6781	1.5082
Class			
mean	2.233	0.25	0.2955
std. dev.	1.1132	0.5615	0.6095

Time taken to build model (full training data) : 2.4 seconds

Clustered Instances

0 108 (36%)

1 106 (35%)

2 89 (29%)

Log likelihood: -27.82606

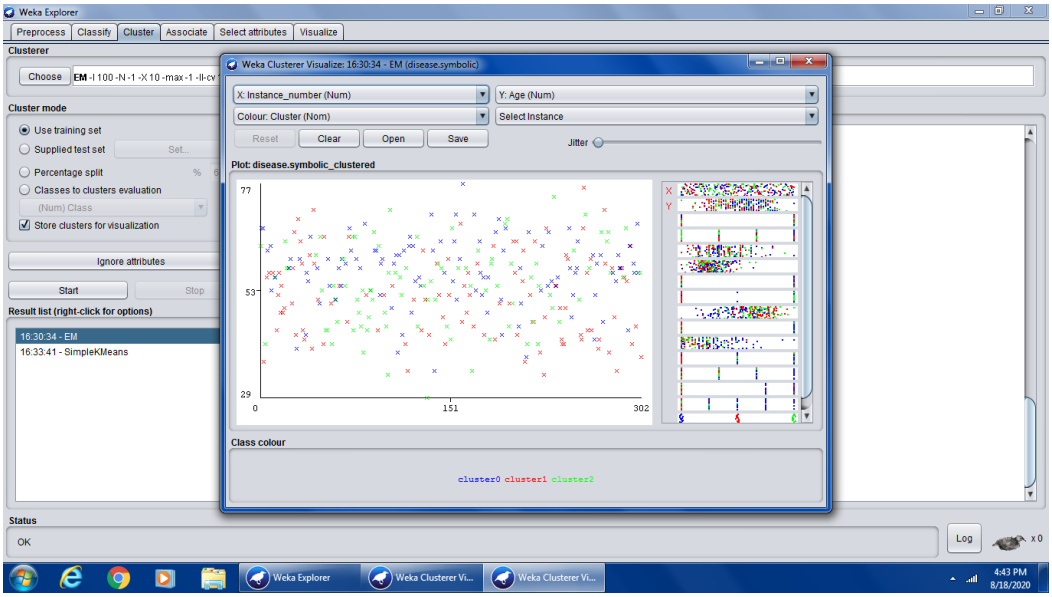

Figure 2: EM clustering result plot

FarthestFirst

==========

Cluster centroids:
Cluster 0
50.0 0.0 2.0 120.0 244.0 0.0 0.0 162.0 0.0 1.1 1.0 0.0 3.0 0.0
Cluster 1
56.0 0.0 4.0 200.0 288.0 1.0 2.0 133.0 1.0 4.0 3.0 2.0 7.0 3.0

Time taken to build model (full training data) : 0 seconds

Clustered Instances

0 228 (75%)

1 75 (25%)

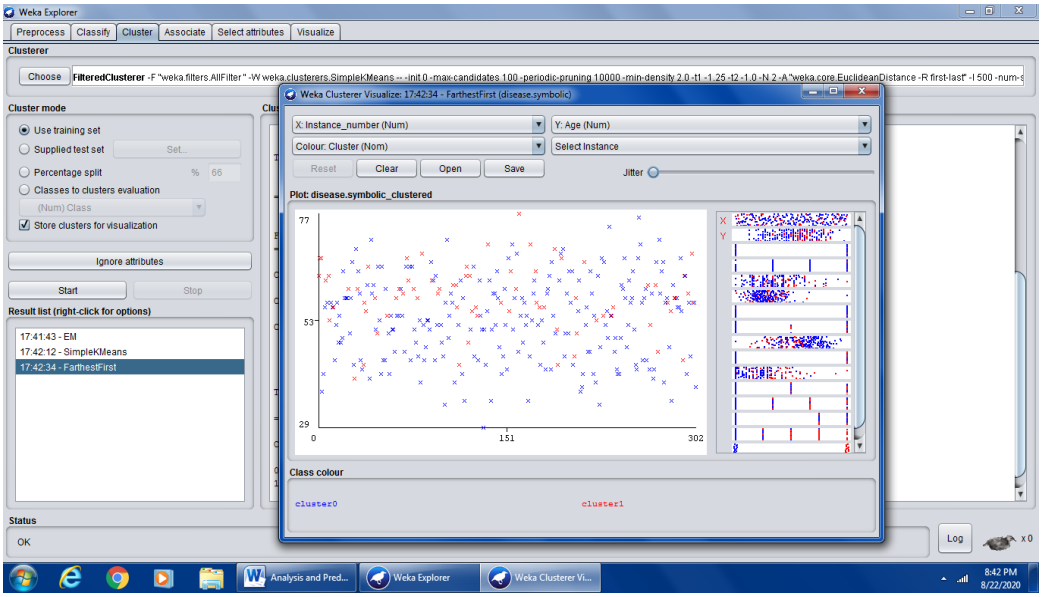

Figure 3: Farthest First clustering result plot

Performance of K Means, EM clustering and Farthest First clustering has been analyzed based on the following factors.

Table 2: Performance of K Means ,EM and FFT

Clustering Techniques used	No. of clusters	Time taken
K Means	2	0.05 seconds
EM	3	2.4 Seconds
Farthest First	2	0 Seconds

A Cleveland dataset is applied to weka tool and the results for the number of clusters and time taken for K Means,EM clustering and farthest first is noted. Farthest First clustering algorithm has performed well when compared with other two algorithms.

5. Conclusions

The primary objective of this analysis is to predict the diseases from the medical data sets. The main aim of this paper is to identify the most appropriate data mining technique to predict the heart disease at an early stage by analyzing different clustering techniques.In this paper the heart disease dataset of Cleveland is taken and subjected to various clustering algorithms using Weka.Clustering is considered as an unsupervised learning technique based on observations.The main goal of clustering analysis is to group both similar and different objects in the same clusters and different clusters respectivelyK-means, EM clustering and Farthest first methods were compared in terms of number of clusters of the classification results and speed. This study showed that how to obtain the clusters and how to determine the new centroid using K-means, EM and Farthest first algorithms.These clusters obtained from three algorithms can be further used as input to classification to get an optimum accuracy for heart disease prediction.

References

[1] World Health Organization. (2016). Hearts: technical package for cardiovascular disease management in primary health care.

[2]Devi, M. R. (2016). Analysis of various data mining techniques to predict diabetes mellitus,*International Journal of Applied Engineering Research,* 11(1). 727-730

[3] Reetu Singh and E.Rajesh," Prediction of Heart Disease by Clustering and Classification Techniques" International Journal of Computer Sciences and Engineering Open Access Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

[4] Arun K. Pujari, ―Data Mining Techniques‖, Universities Press (India) Ltd, 2001.

[5] S. Vijayarani* and S. Sudha," An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples , Indian Journal of Science and Technology, Vol 8(17), DOI: 10.17485/ijst/2015/v8i17/52123, August 2015

[6] Jonathan.C.Prather, M.S. "Medical Data Mining: Knowledge Discovery in a clinical Data warehouse ", 1995.

[7] Madhulatha TS. An overview on clustering methods. IOSR J Eng. 2012;2(4):719–725.
[8] Alkadhwi Ali Hussein Oleiwi and AdelajaOluwaseunAdebayo ,"Data Mining Application Using Clustering Techniques (K-Means Algorithm)In The Analysis Of Student's Result", Journal of Multidisciplinary Engineering Science Studies (JMESS),ISSN: 2458-925X,Vol. 5 Issue 5, May – 2019
[9] BalaSundar V, Devi .T and Saravanan.N "Development of a Data Clustering Algorithm for Predicting Heart",International Journal of Computer Applications (0975 – 888),Volume 48– No.7, June 2012
[10] ZakariaNouir, BernaSayrac, BenoîtFourestié, WalidTabbara, and Françoise Brouaye, "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering," Journal of Communications,Vol. 2, No. 6, pp. 30-37, November 2007.
[11] F. H. Saad, B. de la Iglesia, and G. D. Bell, ― A Comparison of Two Document Clustering Approaches for Clustering Medical Documents, Proceedings of the 2006 International Conference on Data Mining (DMIN-06), 2006.
[12] C. Ordonez, ― Programming the K-Means Clustering Algorithm in SQL, Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, pp. 823-828, 2004.
[13] Shantakumar, B. Patil and Y.S Kumaraswamy.,―Intelligent and Effective Heart attack Prediction System Using Data Mining and Artificial Neural Network, Eurp Journals Publishing Inc. ISSN 1450-216X Vol.31 No.4 2009, pp.642-656, 2009.
[14] Bilmes JA. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models. Berkeley: CA: International Computer Science Institute; 1998
[15]Mr. Rinal H. Doshi Dr. Harshad B. Bhadka Ms. Richa Mehta" Development Of Pattern Knowledge Discovery Framework Using Clustering Data Mining Algorithm" International Journal of Computer Engineering & Technology (Ijcet) Volume 4, Issue 3, May-June (2013), Pp. 101-112.
[16] Data mining in bioinformatics using Weka.Frank E, Hall M, Trigg L, Holmes G, Witten IH Bioinformatics. 2004 Oct 12; 20(15):2479-81
[17] Datasets from "(http://archive.ics.uci.edu/ml/datasets.html)"
[18] Dr.G.Suresh, Dr.A.Senthil Kumar, Dr.S.Lekashri, Dr.R.Manikandan. (2021). Efficient Crop Yield Recommendation System Using Machine Learning For Digital Farming. International Journal of Modern Agriculture, 10(01), 906 - 914. Retrieved from http://www.modern-journals.com/index.php/ijma/article/view/688
[19] Dr. R. Manikandan, Dr Senthilkumar A. Dr Lekashri S. Abhay Chaturvedi. "Data Traffic Trust Model for Clustered Wireless Sensor Network." INFORMATION TECHNOLOGY IN INDUSTRY 9.1 (2021): 1225–1229. Print.
[20] Dr.A.Senthil Kumar, Dr.G.Suresh, Dr.S.Lekashri, Mr.L.Ganesh Babu, Dr. R.Manikandan. (2021). Smart Agriculture System With E – Carbage Using Iot. International Journal of Modern Agriculture, 10(01), 928 - 931. Retrieved from http://www.modern-journals.com/index.php/ijma/article/view/690