International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(2), pp.408-418
DOI: http://dx.doi.org/10.21172/1.102.65
e-ISSN:2278-621X

IMPLEMENTATION OF CLUSTERING MECHANISM FOR WEB BASED DATA WAREHOUSE

Renu Bala[1], Gurdev Singh[2]

Abstract: Big data is a data set that is big in size. It is much complicated so traditional data processing application software is not capable to handle them. There are several challenges such as capture of data, storage of data, searching of data, & transfer of data. Some challenges are related to visualization & querying of data. Scientist has faced several challenges in e-Science such as meteorology, complicated physics simulation & environmental researches. Lot of challenges has been faced due to big data in case of biology & genomics. The problems with existing system were search, sharing, storage, transfer, visualization, querying-updating. These problems can be reduced by using proposed algorithm. In this paper we have explain clustering and proposed algorithm is discussed.
Keywords: Clustering, K-Mean, Data mining, Big data

1. INTRODUCTION

By examining one or more attributes or classes, you may group individual pieces of data value together to form a structure opinion. At a simple stage, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is valuable to identify dissimilar info since this correlates with other examples so you may see where similarities & ranges agree. Clustering[7] may work both ways. You may assume that there is a cluster at a certain point & then use our credentials criteria to see if you are correct. Data Mining includes use of complicated data analyzing tools to find out former unknown and valid patterns & relationships in huge data sets. These type of tools can include various mathematical algorithms, statistical models, & machine learning methods such as neural networks or like decision trees. According to it, data mining includes more than collecting & managing, analysis & calculation of the data. The Objective of data mining is to find potentially useful, valid, novel, understandable correlations & patterns in the present data. To Find useful patterns in the data is known as various names.

Process

The Knowledge Discovery in Databases (KDD) process is commonly defined with stages:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) Data Mining
(5) Interpretation/Evaluation

It exists, however, in many variations on this theme, such as Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

(1) Business Understanding
(2) Data Understanding
(3) Data Preparation
(4) Modeling
(5) Evaluation
(6) Deployment

Or a simplified process such as (1) pre-processing, (2) data mining, & (3) results validation.

Polls conducted in 2002, 2004, & 2007 show that CRISP-DM methodology is leading methodology used by data miners. Only other data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models, & Azevedo & Santos conducted a comparison of CRISP-DM & SEMMA in 2008.

2. PROPOSED WORK

Performance Improvement of Web Usage Mining by Using Learning Based K-Mean Clustering through Neural Network

Due to increasing amount of data available online, World Wide Web had becoming one of most valuable resources for information retrievals & knowledge discoveries. Web removal technologies are correct solutions for information discovery

[1] M.Tech Student Department of CSE, Jind Institute of Engineering & Technology, Jind (Haryana)
[2] Assistant Professor, Department of CSE, Jind Institute of Engineering & Technology, Jind (Haryana)

on Web. Knowledge extracted from Web could be used to raise performances for Web information retrievals, question answering, & Web based data warehousing. In present – day work, we proposition a new technique to increase learning capabilities and reduce calculation intensity of a competitive learning multi-layered neural network using K-means clustering algorithm.

2.1. Proposed algorithm

The main aim is to eliminate limitations of K-mean clustering algorithm, we would customize algorithm as follow.

1. Initialization: In this first step data set, number of clusters & centroid should be calculated automatically according to size of data.

2. Classification: distance is calculated for each data point from centroid & data point having minimum distance from centroid of a cluster is assigned to that particular cluster.

3. Centroid Recalculation: Clusters generated previously, centroid is again reputably calculated means recalculation of centroid.

4. Convergence Condition: Some convergence conditions are given as below:

   4.1 Stopping when reaching a given or defined number of iterations.

   4.2 Stopping when there is no exchange of data points betweenclusters.

   4.3 Stopping when a threshold value is achieved.

5. If all of above conditions are not satisfied, then go to step 2 & whole process repeat again, until given conditions are not satisfied.

6. Elimination of Empty Clusters: Clustersgenerated previously are rechecked

Clusters where no data points are allocated to a cluster under consideration during assignment phase are eliminated.

1) No need of predefined cluster center
2) There would be no Empty clusters at end

7.Make communication between Clients and Server.

8. Read data from cluster on client end before sending.

9. Trasmit the data using IP address.

In this first step data set, number of clusters & centroid should be calculated automatically according to size of data.

Suppose we had following data set

2	5	6	8	12	15	18	28	30

Suppose K=3

C1=2

C2=12

C3=30

2	5	6	8	12	15	18	28	30
C1				C2				C3

So cluster according to distance are as follow

12-5>5-2

So cluster for data point 5 is C1

6-2>12-6

So cluster for data point 6 is C1

In same way cluster would be assigned

2	5	6	8	12	15	18	28	30
C1	C1	C1	C2	C2	C2	C2	C3	C3

Data member of C1 are 2, 5, 6

Data Member for C2 are 8,12,15,18

Data Member for C3 are 28, 30

Clusters generated previously, centroid is again repeatly calculated means recalculation of centroid.

So mean of cluster C1 is (2+5+6)/3=4.3

So mean of cluster C2 is (8+12+15+18)/4=13.25

So mean of cluster C3 is (28+30)/2=29

Now distance would be recalculated within new mean & cluster of data point would be changed according to new distance

2	5	6	8	12	15	18	28	30
C1	C1	C1	C2	C2	C2	C2	C3	C3

For example take 8 from C2 cluster

But problems within existing [10] system were analysis, capture, search, sharing, storage, transfer, visualization, querying updating. One more problems within K-means clustering is that empty clusters are generated during execution, if within case no data points are allocated to a cluster under consideration during assignment phase. Proposed algorithm overcome these problem. K-mean clustering proposed algorithm [14] as follow.

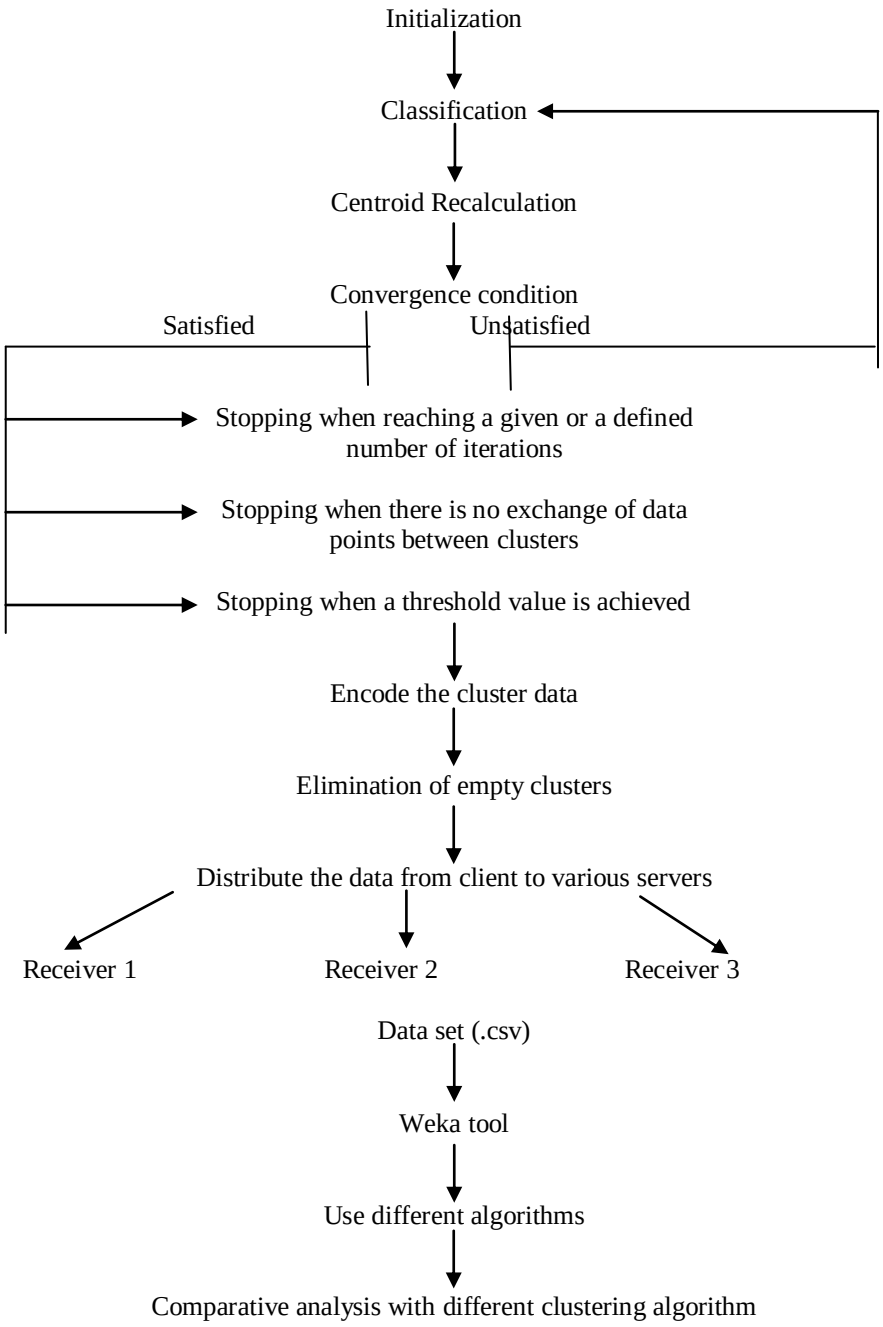

Figure 2 Purposed Model

3. IMPLEMENTATION

3.1 Implementation of K-mean clustering on dataset in Java

When we are applying clustering mechanism on the database named 'Sugar mill' it creates 3 clusters named cluster 1, cluster 2, and cluster 3.

These clusters are automatically created with the code run as javac clustering1.java

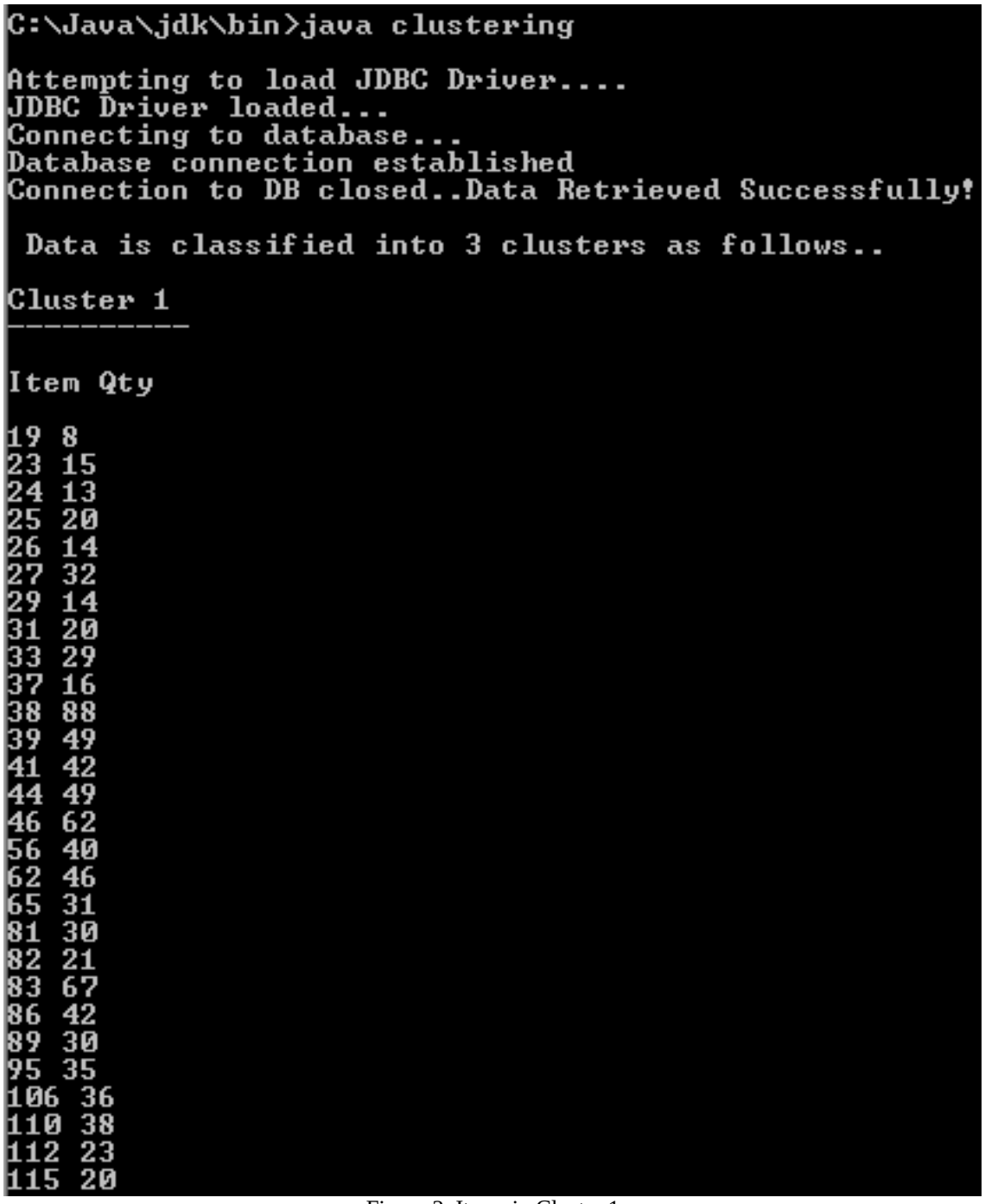

Figure 2 Items in Cluster 1

Above are the items in cluster1. Clusters are created on the basis of the centoid. The points or the data near to the centroid calculated are put on the corresponding cluster.

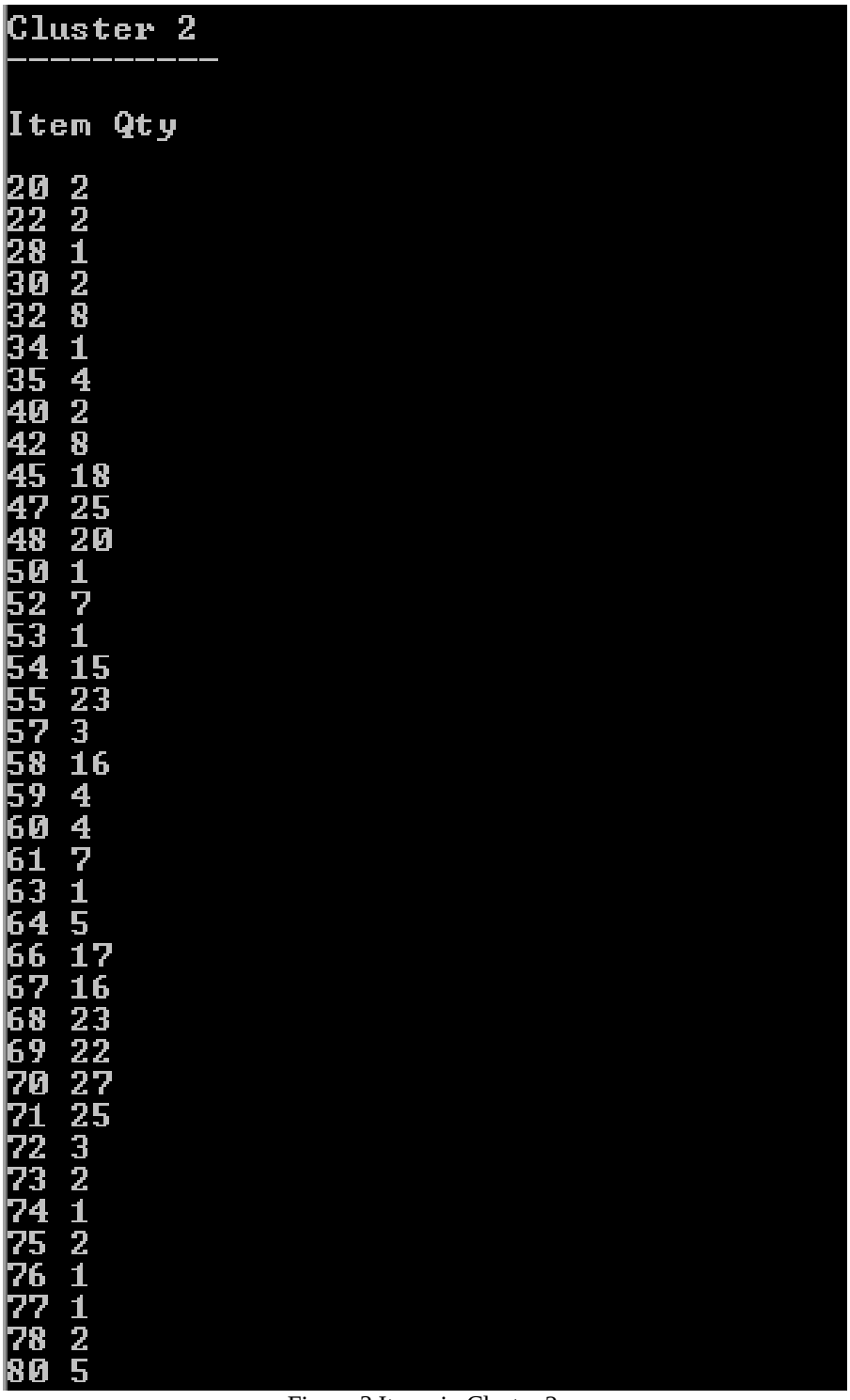

Figure 3 Items in Cluster 2

In the cluster 2 text file, we can see the items with their corresponding quantities. Same as above the centroid calculated and the items near to that centroid point are put into the cluster2.

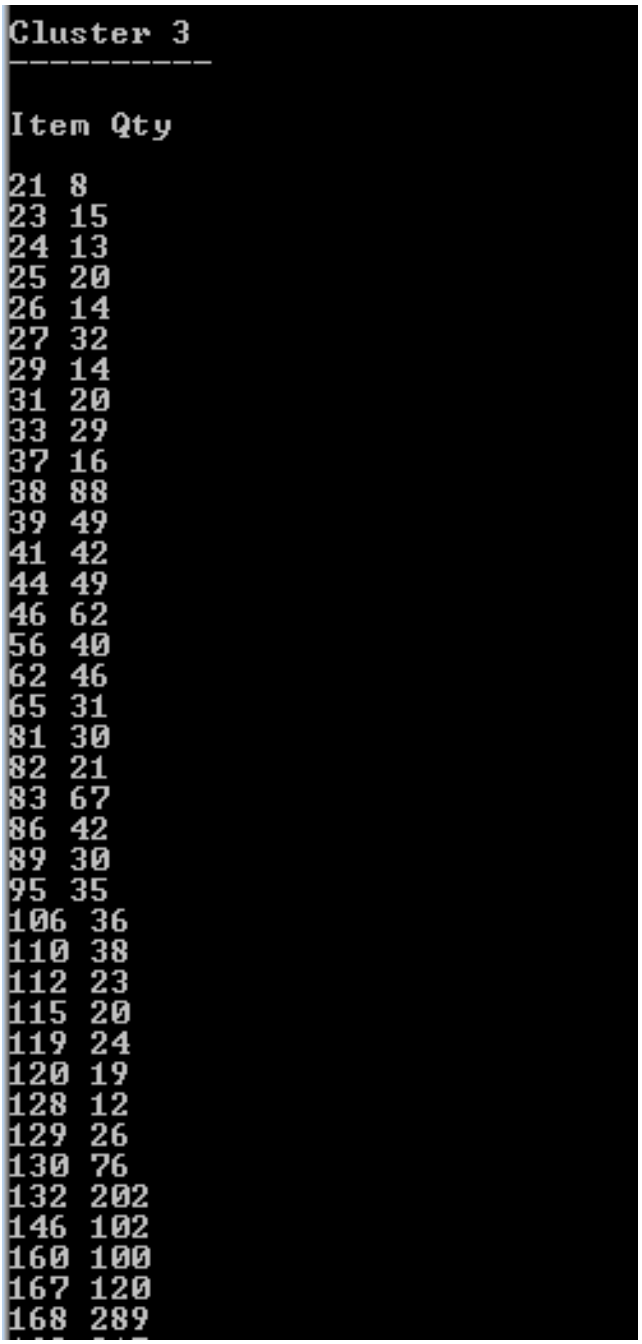

Figure 4 Items in Cluster 3

3.2 Removal of empty clusters

We have successfully created cluster from existing dataset after that vacant cluster have been removed by checking the items stored in cluster if item stored in cluster is 0 then file is removed. We used following code to perform this.

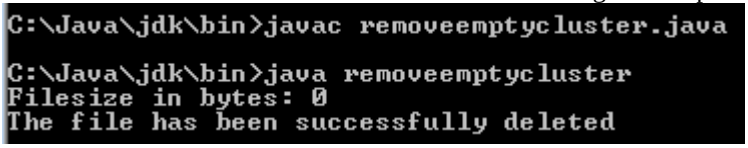

Figure 5 Remove empty clusters

3.3 Implementation of dataset in WEKA

We create the database with the extension of csvThen open that database in the WEKA tool. Then select the attributes of that particular database. And then check for the different clusters algorithms. Save the output of the various algorithms.
While comparing the various clusters algorithm we check for the following errors:

1. Mean absolute error
2. Root mean square error
3. Relative absolute error
4. Root relative squared error

Table 1 Dataset used in WEKA tool

	A	B	C	D
1	id	name	age	salary
2	1	Naresh	32	20000
3	2	Ram	35	20000
4	3	Pooja	32	40000
5	4	Dimple	32	40000
6	5	Priyanka	32	40000
7	6	Renu	32	40000
8	7	Nandini	32	40000
9	8	Ramveer	32	20000
10	9	Param	32	20000
11	10	Purva	34	20000
12	11	Neha	34	20000
13	12	Rishabh	34	30000
14	13	Gouri	34	30000
15	14	Khushi	34	30000
16	15	Suman	34	30000
17	16	Gautam	34	30000
18	17	Santosh	34	32000
19	18	Aastha	34	32000
20	19	Radhe	34	32000
21	20	Sonu	34	32000
22	21	Sanjay	35	30000
23	22	Deepu	35	30000
24	23	Kiran	35	20000
25	24	Laddu	35	20000

dataset

We get the following conclusion from the above dataset.

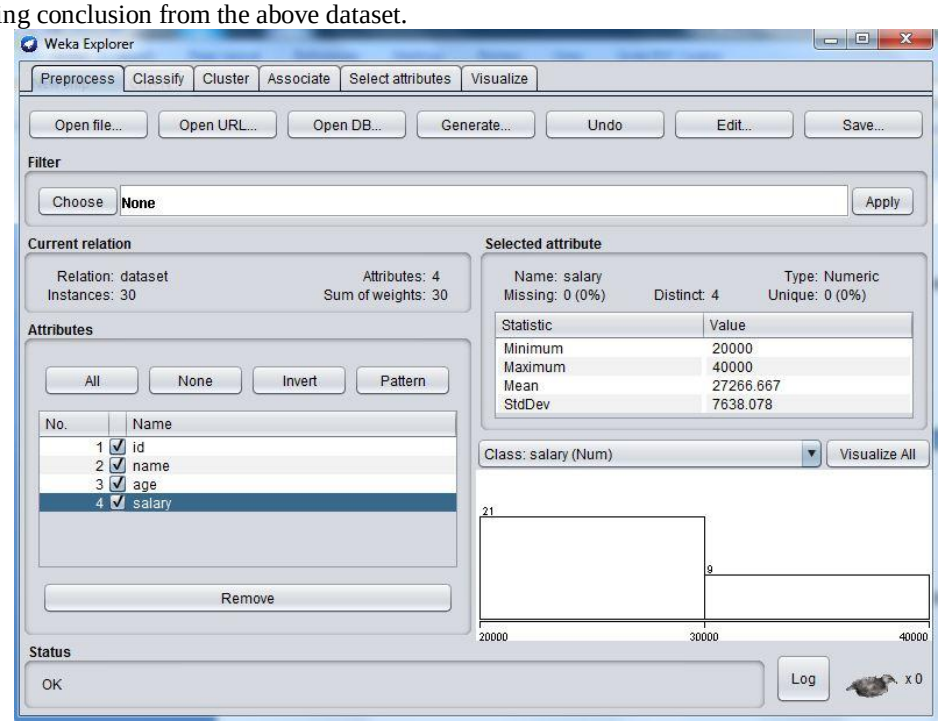

Figure 6 Conclusions from Dataset

There are four attributes in the dataset.
There are total 30 instances in this dataset.

The mean is 27266.667
The standard deviation is 7638.078

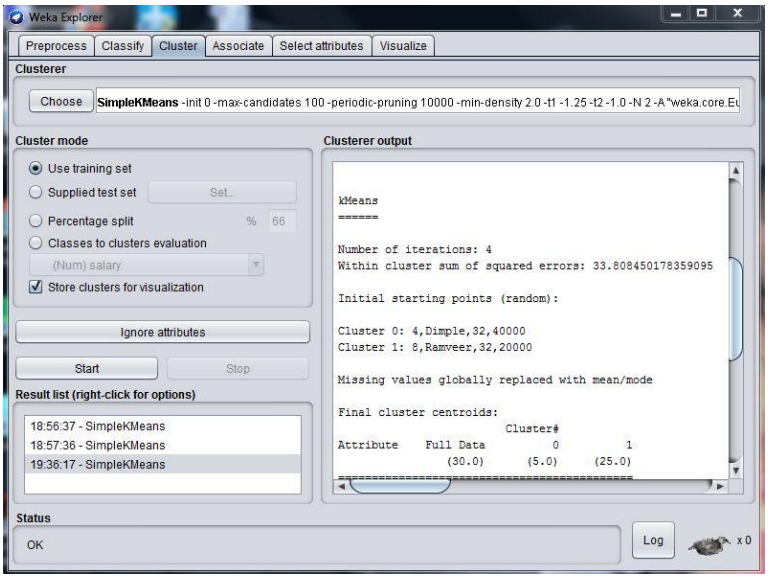

Figure 7 K Means Clustering

When we are using Kmeans clustering then in the following dataset there are 4 iterations and the total number of squared errors are 33.8.8450178359095.

Now we check forDensity Based Clustering

Density-based spatialclusteringis a data clustering algorithm in which given a set of points in some space, it groups together points that are closely packed together marking as outliers points that lie alone in low-density regions DBC is one of the most common clustering algorithms and also most cited in scientific literature.

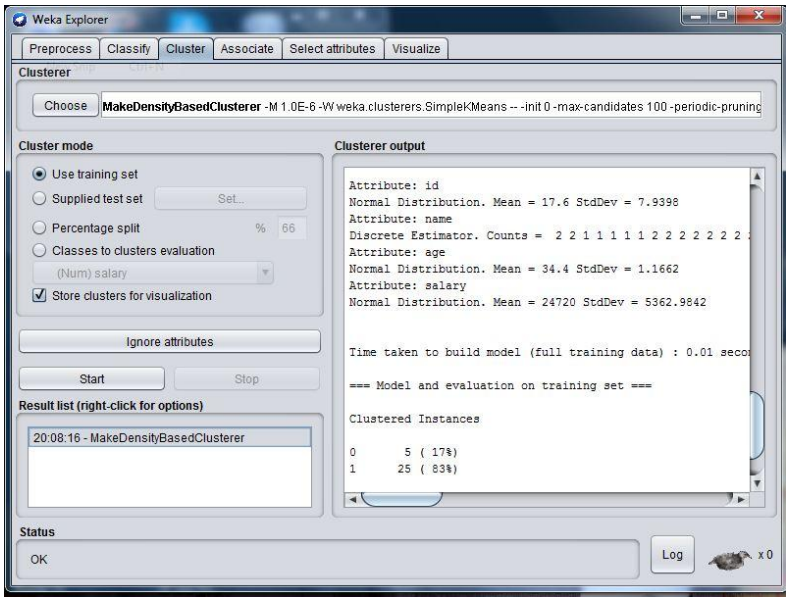

Figure 8 Density Based Clustering

In the make density based cluster algorithm, on the basis of different attributes we get the different mean and the standard deviations

Attribute: id
Normal Distribution. Mean=17.6 Std Dev=7.9398
Attribute: name
Discrete Estimator. Counts= 2 2 1 1 1 1 1 1 2 2 2 2 2
Attribute: salary
Nomal Distributioin. Mean=24720 StdDev=5362.9842

3.4 Now check for Hierarchical Clustering Algorithm

In data mining hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

1. Agglomerative: This is a bottom up approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
2. Divisive: This is a top down approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

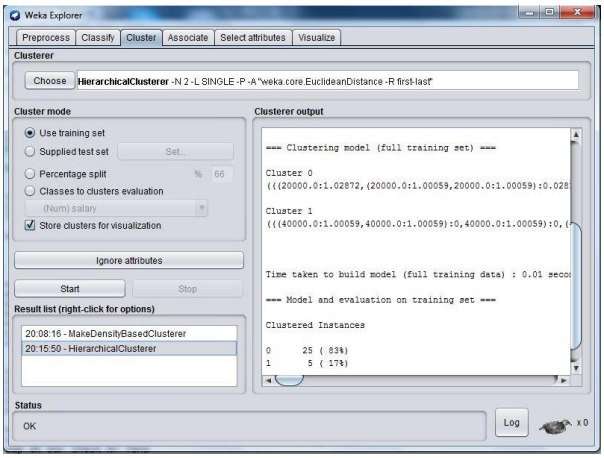

Figure 9 Hierarchical Clustering Algorithm

While using the previous data set, with the help of Hierarchical Clustering, two clusters are created and the respective instances are 83% with the first cluster and 5% with the second cluster.

Now check for farthest first clustering algorithm

The farthest first clustering of a bounded metric space is a sequence of points in the space, where the first point is selected arbitrarily and each successive point is as far as possible from the set of previously-selected points. The same concept can also be applied to a finite set of geometric points, by restricting the selected points to belong to the set or equivalently by considering the finite metric space generated by these points. For a finite metric space or finite set of geometric points, the resulting sequence forms a permutation of the points, known as the greedy permutation.

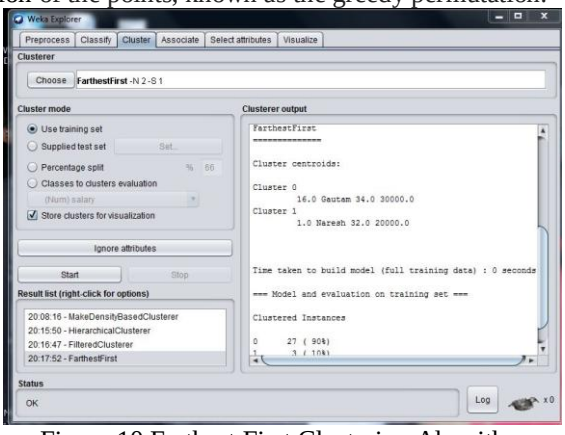

Figure 10 Farthest First Clustering Algorithm

In the farthest first clustering, the respective cluster instances are 90% in the first cluster. 10% with the second cluster.

In this way, we can calclucate mean, standard deviation and instances with respective clustering algorithms.

Summary of different clustering meechanisms

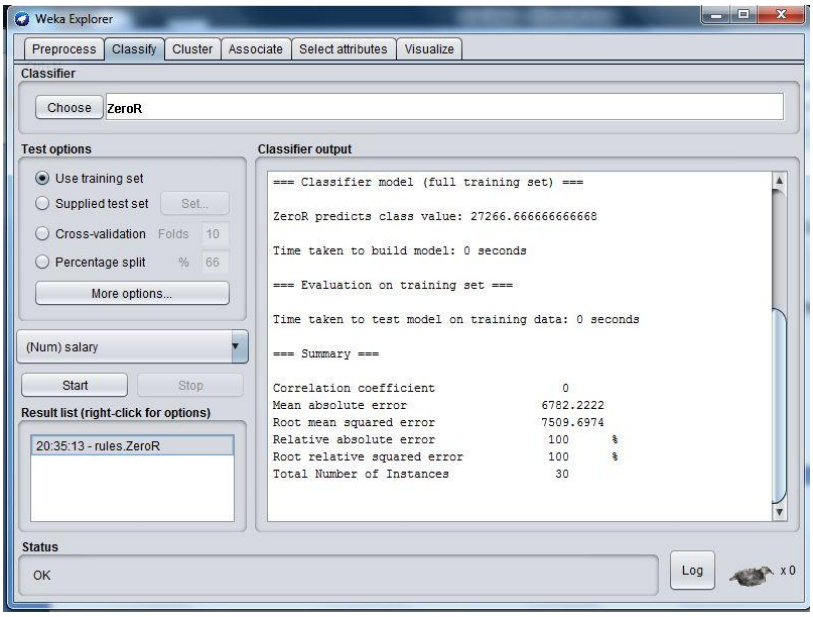

Figure 11 Summary

This figure provides with the summary of the classification done yet.
It will determine the total percentage of mean absolute error, root mean squared error, relative absolute error and total number of instances occurred.
We can also cross validate the result come from this WEKA tool by using Test option of Cross Validation.

4. CONCLUSION

Clustering is process of grouping objects that belongs to same class. Similar objects are grouped in one cluster & dissimilar objects are grouped in another cluster. We have explain comparative analysis of number of clusters formed in case of existing K mean clustering & proposed K mean clustering[14] . The number of vacant clusters had been removed in case of proposed clustering algorithm so number of clusters get reduced in case of proposed algorithm.

5. REFERENCES

[1] Fahim A.M. (2006) wrote on "An Efficient enhanced k-means clustering algorithm" International Journal of Database Theory & Application
[2] Hong Liu 1, & Xiaohong Yu (2009) wrote on "Application Research of k-means Clustering Algorithm in Image Retrieval System" International Journal of Database Theory & Application
[3] Bhagyashree Umale (2011) "Overview of K-means & Expectation Maximization Algorithm for Document Clustering" International Conference on Quality Up-gradation in Engineering, Science & Technology David Jensen & Jennifer Neville (2012) "Data Mining in Social Networks" Computer Science Department faculty publication.
[4] Vishal Shrivastava (2012) "A Study of Various Clustering Algorithms on Retail Sales Data of Computing, Communications & Networking" International Journal of Advanced Research in Computer Science & Software Engineering
[5] Kalyani M Raval (2012) "Data Mining Techniques" International Journal of Advanced Research in Computer Science & Software Engineering
[6] Navjot Kaur (2012) wrote on "Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining" International Journal of Research
[7] Neelamadhab June (2012) " Survey of Data Mining Applications & Feature Scope" International Journal of Computer Science, Engineering & Information Technology
[8] Atul Kumar Pandey (2013) "Data Mining Clustering Techniques in Prediction of Heart Disease using Attribute Selection Method" International Journal of Science, Engineering & Technology Research (IJSETR)
[9] Bhoj Raj Sharma (2013) "Clustering Algorithms: Study & Performance Evaluation Using Weka Tool" International Journal of Current Engineering & Technology
[10] Nikita Jain1, Vishal Srivastava2 Nov (2013) Data mining techniques: A Survey paper IOSR Journal of Computer Engineering (IOSR-JCE)
[11] Aarti Sharma (2014) "Application of Data Mining – A Survey Paper" International Journal of Trend in Research & Development
[12] Shweta Srivastava (2014) "Clustering Techniques Analysis for Microarray Data" International Journal of Computer Science & Mobile Computing A Monthly Journal of Computer Science & Information Technology IJCSMC
[13] Muhammad Husain Zafar (2015) "A Clustering Based Study of Classification Algorithms" International Journal of Database Theory & Application
[14] D. Asir Antony (2016) "Performance Analysis on Clustering Approaches for Gene Expression Data" International Journal of Advanced Research in Computer & Communication Engineering