CLUSTERING BIOLOGICAL DATA: AN ENHANCED K-MEANS APPROACH

Sunila[1] , Rishipal Singh[2], Sanjeev Kumar[3]

[1,2,3] Dept. of CSE, Guru Jambheshwar University Sci. & Tech. Hisar, (India)

ABSTRACT

Clustering is one of the unsupervised learning method in which a set of essentials is separated into uniform groups. The k-means method is one of the most widely used clustering techniques for various applications However, there are some shortcomings of it, such as it requires a user to give out the number of clusters at first, and its sensitiveness to initial conditions, and its easily getting to the trap of a local solution. Cluster analysis for medical data sets has proved to be a useful tool for identifying co-expressed genes, biologically relevant groupings of genes and samples. But it has a heavy computational load.

In this paper, we propose a new version of the K-means algorithm. The outstanding features of our algorithm are its superiority in execution time and accuracy. In this paper, the most delegate algorithms K-Means ,Hierarchical and Proposed Enhanced K-Means are examined and analyzed using WEKA machine learning tool. Experimental results on Biological data sets show that the Proposed Enhanced K-Means algorithm make clusters in minimum time and have good performance than K-Means and Hierarchical clustering. Imbalanced problems of Biological domains are relaxed by optimizing distance measure and performing ensembles using Boosting Approach.

Keywords:* *Data Mining, Clustering, K-Means Clustering, Hierarchical Clustering, Distance Measure.

I INTRODUCTION

Clustering is a method of un-supervisory learning and a common technique for data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar between them- selves and dissimilar compared to objects of other groups. Cluster analysis is a very important technology in Data Mining. It divides the datasets into several meaningful clusters to reflect the data sets' natural structure. Cluster is aggregation of data objects with common characteristics based on the measurement of some kind of information. The following diagram represents the clustering process[14]:

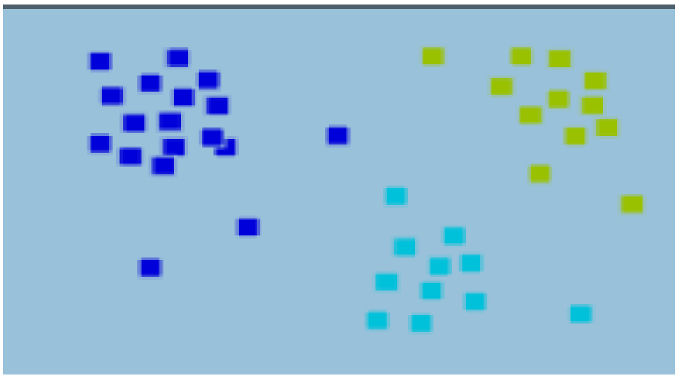

Fig 1: Result of the Cluster Analysis

There are several commonly used clustering algorithms, such as K-means, Density based ,Hierarchical and so on [2]. Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups.[3] As its well known that each clustering algorithm ,sometimes even the same clustering algorithm applied several times on initial dataset, can result in different partitions. The selection of subset of attributes for Clustering as well as number of clusters also effects performance of clustering.

Clustering of a set forms a partition of its elements chosen to minimize some measure of dissimilarity between members of the same cluster .Clustering algorithms are often useful in various fields like data mining, pattern recognition, learning theory etc[18].

The rest of the paper is organized as follows: Section 2 reviews the previously work done on related algorithms. Section 3 describes Existing clustering techniques. Section 4 describes Proposed Modified Algorithm. Section 5 describes the performance evaluation of various clustering techniques. Conclusions are remarked in section 6.

II RELATED WORK

Osama Abu Abba et al. [2] paper is intended to study and compare different data clustering algorithms. The algorithms under investigation are: k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm and expectation maximization clustering algorithm. All these algorithms are compared according to the following factors: size of dataset, number of clusters, type of dataset and type of software used. Some conclusions that are extracted belong to the performance, quality, and accuracy of the clustering algorithms.

Manish Verma et al. [3]. This paper reviews six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DB Scan clustering, Density Based Clustering, Optics, EM Algorithm. A.K. Jain et al. [4] presented an overview of pattern clustering methods from a statistical pattern recognition perspective. D.Napoleon et al. [5] present that K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, apply PCA on original data set and obtain a reduced dataset containing possibly uncorrelated variables. In this paper principal component analysis and linear transformation is used for dimensionality reduction and initial centroid is computed, then it is applied to K-Means clustering algorithm.

Kehar Singh, Dimple et al. [6] K-means is very popular because it is conceptually simple and is computationally fast and memory efficient but there are various types of limitations in k means algorithm that makes extraction somewhat difficult. In this paper we are discussing these limitations and how these limitations will be removed.

N.S.Chandolikar et al. [7] this paper evaluate performance to two well known classification algorithms for attack classification. Bayes net and J48 algorithm are analyzed The key ideas are to use data mining techniques efficiently for intrusion attack classification.

Adil M. Bagirov et al. [8] developed a new version of the global k-means algorithm, the modified global k-means algorithm. Clustering algorithms based on global optimization techniques are not applicable to even relatively large data sets. Algorithms which are applicable to such data sets can locate only local minima of the function and these local minima can differ from global solutions significantly as the number of clusters increases. The number of clusters, as a rule, is not known in advance. So an incremental approach used to locate a local solution gave solution close to global one.

Christopher M. et al. [9] present a hierarchical visualization algorithm which allows the complete data set to be visualized at the top level, with clusters and sub clusters of data points visualized at deeper levels. The algorithm is based on a hierarchical mixture of latent variable models, whose parameters are estimated using the expectation-maximization algorithm.

M.Awad et al. [10] proposed an extension of the CFA algorithm (the Enhanced CFA algorithm, ECFA) which uses the real output of the net that they were using to make the approximation, since clusters should not be located in the same place when using classical RBFs as when using normalized RBFs. ECFA migrates clusters to the zones of the input space where the approximation error is bigger, thus trying to homogenously distribute the total distortion in every cluster, producing a better share-out of clusters for the data input space. This paper shows that ECFA outperforms the CFA algorithm not only with respect to the final approximation error but also with respect to the execution time.

Fasahat Ullah Siddiqui et al. [11]. This paper presents an improved version of the Moving K Means algorithm called Enhanced Moving K-Means (EMKM) algorithm. In the proposed EMKM, the moving concept of the conventional Moving K-Means (i.e. certain members of the cluster with the highest fitness value are forced to become the members of the clusters with the smallest fitness value) is enhanced. Two versions of EMKM, namely EMKM-1and EMKM-2 are proposed. The qualitative and quantitative analyses have been performed to measure the efficiency of both EMKM algorithms over the conventional algorithms (i.e. K-Means, Moving K-Means and Fuzzy C-Means) and the latest clustering algorithms (i.e. AMKM and AFMKM). It is investigated that the proposed algorithms significantly outperform the other conventional clustering algorithms.

Mu-Chun Su et al. [12] present a modified version of the K-means algorithm to cluster data. The proposed algorithm adopts a novel non metric distance measure based on the idea of point symmetry. This kind of point symmetry distance can be applied in data clustering and human face detection. Several data sets are used to illustrate its effectiveness.

Tapas Kanungo et al. [13] this paper present a simple and efficient implementation of Lloyd's k means clustering algorithm, which is called filtering algorithm. This algorithm is easy to implement, requiring a tree as only major data structure. They establish the practical efficiency of the filtering algorithm in two ways. First, they present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, they present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

Malay et al. [14].This paper presents a modified version of the k-means algorithm that efficiently eliminates this empty cluster problem. The proposed algorithm is semantically equivalent to the original k-means and there is no performance degradation due to incorporated modification.

Narendra Sharma et al. [15] presents the study of various clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. WEKA tool was used for comparisons.

Bradley et al. [16] present a technique for initializing the K-means algorithm. They begain by randomly breaking the data into10, or so, subsets. They then performed a K-means clustering on each of the10 subsets, all starting at the same set of initial seeds, which are chosen randomly. The result of the 10 runs is 10K centre points. These 10K points were then themselves input to the K-means algorithm and the algorithm run 10 times, each of the 10 runs initialized using the K final centroid locations from one of the 10 subset runs. The resulting K centre locations from this run are used to initialize the K-means algorithm for the entire dataset.

Khan et al. [17] present an algorithm to compute initial cluster centers for K-means clustering. This algorithm is based on two observations that some of the patterns are very similar to each other and that is why they have same cluster membership irrespective to the choice of initial cluster centers. Also, an individual attribute may provide some information about initial cluster center.

III EXISTING CLUSTERING TECHNIQUES

The K-means algorithm

K-means algorithm follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centroids, one for each cluster. The simple K-means algorithm chooses the centroid randomly from the data set. The next step is to take each data belonging to a given data set and associate it to the nearest centroid. The K-means clustering partitions a data set by minimizing a sum of squares cost function[18].

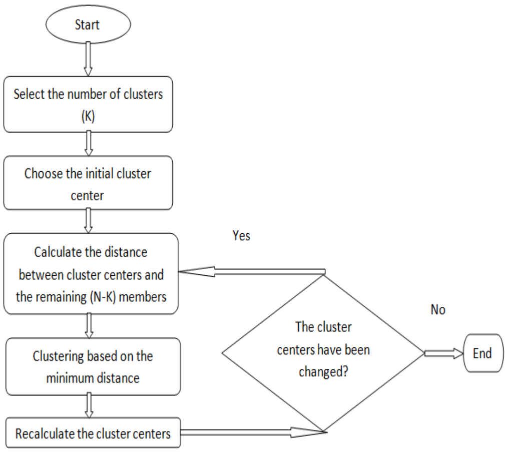
Start
Select the number of clusters (K)
Choose the initial cluster center
Calculate the distance between cluster centers and the remaining (N-K) members
Clustering based on the minimum distance
Recalculate the cluster centers
Yes
The cluster centers have been changed?
No
End

Figure 2: K-Means Algorithm Steps[6].

Hierarchical clustering

A hierarchical clustering method works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. That is, if a particular merge or split decision later turns out to have been a poor choice, the method cannot backtrack and correct it. Recent studies have emphasized the integration of hierarchical agglomeration with iterative relocation methods [17].

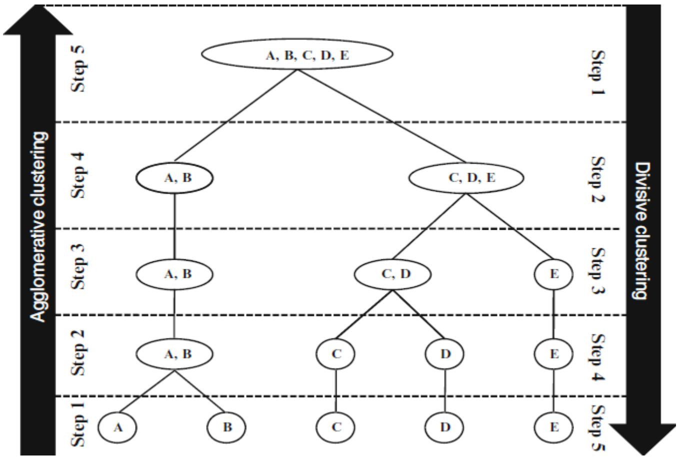

Figure 3: Agglomerative and divisive hierarchical clustering on data objects (a, b, c, d, e)

Hierarchical clustering takes as input a set of points. It creates a tree in which the points a internal nodes reveal the similarity structure of the points. The tree is often called a “dendogram “.The method is summarized below:

1. Place all points into their own cluster. While there is more than one cluster, do

2. Merge the closest pair of clusters algorithm depends on how “closest pair of clusters” is [17].

IV NEW TECHNIQUE: PROPOSED ENHANCED ALGORITHM

Proposed Enhanced Algorithm

Our Proposed Modified algorithm uses standard deviation that reduces the time to make the cluster in simple k-means. The main contribution is to divide the square root distance with standard deviation By this we optimized a distance measure method that gives good result as compare to simple k-means clustering algorithm..

The Euclidean distance between one vector x=(x1 ,x2,…xn) and another vector y=(y1 ,y2 ,…yn) is Euclidean distance d(xi, yi) .The Euclidean distance formula is given below

$$d = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

Our proposed algorithm also uses standard deviation that reduces the time to make the cluster in simple k-mean. The main contribution is to divide the square root distance with standard deviation is that by doing this deviation from mean can be reduced more.

Proposed Weighted Euclidean distance formula is given as:

$$d = \text{Math.ulp}(\ (\text{var})^{-1} \sum_{j=1}^{n} (x_j - y_j)^2\ ) \quad (1)$$

Variance is calculated using the given formula:

$$\text{var} = \frac{\sum_{j=1}^{n} (x_j - \overline{x}\ )^2}{\text{n - 1}}$$

Where, var is the variance of the attributes of instances and $\overline{x}$ is the mean of the attributes of instances.

The Math.ulp() method returns the distance from a number to its nearest neighbors. This distance is called an ULP for unit of least precision or unit in the last place [9]. Given one float or double, there is a next float; and there is a minimum finite distance between successive floats and doubles. These methods can be useful in modeling applications. Numerically, to sample a value at 10,000 positions between a and b, but if we're getting only enough precision to identify 1,000 unique points between a and b, then we're doing redundant work nine times out of ten. We can do a tenth of the work, and get results that are just as good.

The size of a function d is not constant. As numbers get larger, there are fewer floats between them. For instance, there are only 1,025 floats between 10,000 and 10,001; and they're 0.001 apart. Between 1,000,000 and 1,000,001 there are only 17 floats, and they're about 0.05 apart. Clustering algorithms based on global optimization are not applicable for large data sets. Algorithms which are applicable to such data sets can locate local minima of function d. These local minima can differ from global solutions significantly as number of clusters increases The value of function **E** given in equation(2) below must be minimized.

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{x \in C_i} \text{Math.ulp}(\ (\text{var})^{-1} \sum_{j=1}^{n} (x_j - y_j)^2\ ) \quad (2)$$

We have performed ensembles using Boosting approach. In this technique more weight is given to the observations which are misclassified in each step. By this we can decrease misclassification rate.

Proposed Enhanced Algorithm's steps are listed below:

```
Function distance()
1. For i=1 to n
```

```
2. For j=1 to k
3. Compute Euclidean distance using Eq(1)
4. d Proposed distance function
5. end for
6. Find the closest centroid mj to xi;
7. mj=mj+xi;
8. nj=nj+1;
9. Calculate E using Eq(2)and Minimize Value of E ;
10.   E=E+ d
11.   Clusterid[i]=number of the closest centroid;
12.   Pointdis[i]=Euclidean distance to the closest centroid;
13.   endfor
14.   For j=1 to k
15.   mj=mj/nj;
16.   endfor
17.   Ensemble above steps using Boosting approach;
17.1. Initially, weights are set equally.
17.2. Iterate:
17.2.1. Train weak learner on weighted data
17.2.2. Increase weights of incorrectly classified  examples (force weak learner to focus on difficult examples)
17.3. Final hypothesis: combination of weak hypotheses
```

This function is called distance().point number i and all k centroids. Line 5 searches for the closest centroid to point number i, say the closest centroid is number j. Line 6 adds point number i to cluster number j, and increase the count of points in cluster j by one. Lines 8 and 9 are used to enable us to execute the proposed idea; these two lines keep the number of the closest cluster and the distance to the closest cluster. Line 12 does centroids recalculation.

V EXPERIMENTAL SETUP & PERFORMANCE METRICS

Data sets Used:

Pima Indian Diabetes data set:

This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and has 9 attributes and 768 instances.

Echo Cardiogram data set:

This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 13 attributes and 131 instances.

Haberman data set:

This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 4 attributes and 306 instances.

Lympography data set:

This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 19 attributes and 148 instances.

Segment-test data set:

This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 20 attributes and 810 instances.

Cleveland data set :

This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 14attributes and 303instances.

Liver Disorder data set:

This data set is downloaded from archive.ics.uci.edu/ml/datasets.html and contains 7 attributes and 345 instances.

Comparative Analysis

The K-Means, Hierarchical and Proposed Enhanced K-Means clustering are applied on Biological data sets and their results are compared with respect to time complexity and accuracy. With help of analysis, it is shown that Enhanced Proposed K-Means has taken less time to make cluster on Biological datasets. It has more accuracy then K-Means and Hierarchical clustering algorithms. The table1 shows the accuracy in terms of correctly clustered instances by the above algorithms to make clusters. The given figure no.4 shows respective results in case of accuracy. The table2 shows the time taken by the above algorithms to make clusters. The given figure no.5 shows respective results in case of time complexity.

Finally, the generated results by Proposed Enhanced K-Means outperform K-Means and Hierarchical clustering in terms of, accuracy and time taken to build clusters.

TABLE 1. ACCURACY OF VARIOUS CLUSTERING ALGORITHMS

	Accuracy for Pima Indian diabetes data set %(values)	**Accuracy for Echo Cardiogram data set %(values)**	**Accuracy for Haberman data set %(values)**	**Accuracy for Lymphography data set %(values)**	**Accuracy for segment-test data set %(values)**	**Accuracy for Cleveland data set %(values)**	**Accuracy for Liver Disorder data set %(values)**
K-Means	64.8	67.8	54.11	53.26	59.26	69.33	50.3
Hierarchical	63.6	60.82	48.21	45.72	54.23	65.22	46.3
Proposed Enhanced K-Means	69.55	69.9	55.82	54.41	64.21	71.21	51.9

TABLE 2. TIME COMPLEXITY OF VARIOUS CLUSTERING ALGORITHMS FOR GIVEN DATA SETS .

	Time for Pima Indian diabetes data set	Time for EchoCardiogram set	Time for Haberman data sets	Time taken for Lymphography data set	Time taken for segment-test data set	Time taken for Cleveland data set	Time taken for Liver Disorder data set
K-Means	0.11	0.03	0.02	0.03	0.27	0.04	0.025
Hierarchical	0.16	0.08	0.09	0.16	0.38	0.12	0.14
Proposed Enhanced K-Means	0.07	0.02	0.012	0.01	0.19	0.02	0.020

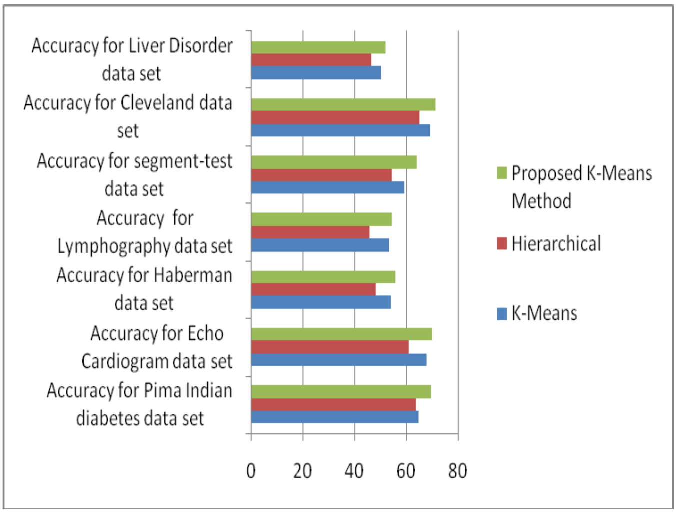

Figure 4: Shows Respective Results in Terms of Accuracy.

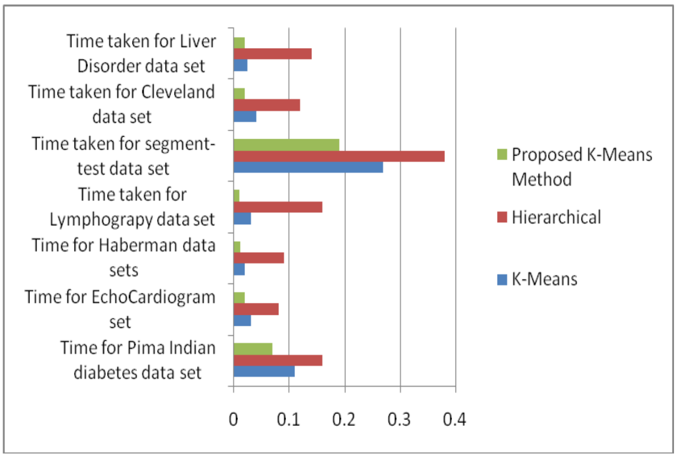

Figure 5: Shows Respective Results in Terms of Time Complexity.

VI. CONCLUSION

We have proposed an enhancement in the simple K-means algorithm and the experimental results have proved that with this enhanecment the clustering performance drastically increased in terms of time and complexity as compared to K-Means and Hierarchical clustering . We have used ensembles here to deal with imbalance data and local minima problem but still it takes less time by use of Math.ulp approach. The performance of proposed algorithm is tested across seven real world datasets and the results are quite encouraging and have established the effectiveness of the proposed algorithms.This properly classified data can be given to a predication model to enhance predictive accuracy. The proposed work can also be explored by use of various filtering algorithm for data preprocessing. Further we will apply certain constraints on Proposed algorithm which will not only improve its cluster quality but also its efficiency.

REFERENCES

[1] Johannes Grabmeier,Fayyad, Mannila, Ramakrishnan,“ Techniques of Cluster Algorithms inData Mining,”May 23 2001.

[2] Osama Abu Abbas, Jordan, “Comparisons Between Data Clustering Algorithms,”The International Arab Journal of Information Technology, vol. 5, no. 3, pp.320-326,Jul. 2008.

[3] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta,“ A Comparative Study of Various Clustering Algorithms in Data Mining,” International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com, vol. 2, Issue 3, pp.1379-1384,May-Jun. 2012.

[4] A.K. Jain, M.N. Murty, P.J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, vol. 31, no. 3, pp. 264-323,Sep. 1999.

[5] D.Napoleon,S.Pavalakodi,”A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set,” International Journal of Computer Applications (0975–8887),vol. 13, no.7, pp.41-46, Jan 2011.

[6] Adil M. Bagirov Karim Mardaneh,” Modified global k-means algorithm for clustering in gene expression data sets”, Workshop on Intelligent Systems for Bioinformatics (WISB2006), Hobart, Australia,vol 73,2006.

[7] Christopher M. Bishop and Michael E. Tipping, “A Hierarchical Latent Variable Model For Data Visualization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp.281-293, Mar. 1998.

[8] M.Awad, H. Pomares, I.Rojas, Member, IEEE,” Enhanced Clustering Technique in RBF Neural Networkfor Function Approximation”,Mathematical and Computer Modelling,Vol 55,Issues3-4, pp. 286-302, Feb. 2012.

[9] Fasahat Ullah Siddiqui and Nor Ashidi Mat Isa,”Enhanced Moving K-Means (EMKM) Algorithmfor Image Segmentation,”IEEE, pp.833-841.

[10] Mu-Chun Su and Chien-Hsing Chou, “A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry,”IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp.674-680, Jun. 2001.

[11] TapasKanungo,David M.Mount ,NathanS. Netanyahu, Christine D.Piatko, Ruth Silverman, and Angela Y.Wu,"An Efficientk-Mean Clustering Algorithm: Analysisan Implementation ,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-891,Jul 2002.

[12] Malay K. Pakhira, "A Modified k-means Algorithm to Avoid Empty Clusters," International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp.220-226, May 2009.

[13] Narendra Sharma ,Aman Bajpai,Mr.Ratnesh Litoriya," Comparison the various clustering algorithms of weka tools," International Journal of Emerging Technology and Advanced Engineering, vol. 2, pp.73-80, May 2012.

[14] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-meansclustering," Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1998, pp. 91–99.

[15] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for kmeans clustering," Pattern Recognition Letters, vol. 25, pp. 1293–1302,2004

[16] Ian Davidson S. S. Ravi" Using Instance-Level Constraints in Agglomerative Hierarchical Clustering: Theoretical and Empirical Results"International Journal of Data Mining and Knowledge Discovery,Springer,vol 18,257-282,June 2008.

[17] Prof .Mrs. J. R Prasad, R.S. Prasad, Dr. U.V Kulkarnai "Impact of Feature Selection Methods in Hierarchical Clustering Technique:A Review" Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008 Vol 2, IIMECS , Hong Kong, 19-21 March, 2008.

[18] Sunila Godara and Dr. Rishipal Singh, "An Efficient method to Improve Performance of K- Means Clustering Algorithms for Medical Domains", International Journal of Applied Engineering Research, Volume 10, Number 13 ,2015.

[19] http://www.openclinical.org/aisp_era.html