A Performance Evaluation of SMCA Using Similarity Association & Proximity Coefficient Relation For Hierarchical Clustering

Mr. Mayank Gupta, Mr. Ritesh Jain

PG Scholar, Asst. Professor
Computer Science & Engineering Department
Oriental University, Indore, India

***Abstract*— Clustering techniques have a wide use and importance nowadays. This importance tends to increase as the amount of data grows and the processing power of the computers increases. Clustering applications are used extensively in various fields such as artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing.There are several algorithms and methods have been developed for clustering problem. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. This type of dilemma motivated us to develop new algorithm and process for clustering problems. There are several another issue are also exits like cluster analysis can contribute in compression of the information included in data. In several cases, the amount of available data is very large and its processing becomes very demanding. Clustering can be used to partition data set into a number of "interesting" clusters. Then, instead of processing the data set as an entity, we adopt the representatives of the defined clusters in our process. Thus, data compression is achieved. Cluster analysis is applied to the data set and the resulting clusters are characterized by the features of the patterns that belong to these clusters. Then, unknown patterns can be classified into specified clusters based on their similarity to the clusters' features. Useful knowledge related to our data can be extracted [1].**

***Keywords*— Cluster, Hierarchical Clustering, Feature Matrices**

I. INTRODUCTION

Clustering is unsupervised learning because it doesn't use predefined category labels associated with data items. Clustering algorithms are engineered to find structure in the current data, not to categories future data. A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster [4].

Fig. 1 clustering of raw data

Cluster analysis is a convenient method for identifying homogenous groups of objects called clusters; objects in a specific cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster. After having decided on the clustering variables we need to decide on the clustering procedure to form our groups of objects. This step is crucial for the analysis, as different procedures require different decisions prior to analysis. These approaches are: hierarchical methods, partitioning methods and two-step clustering. Each of these procedures follows a different approach to grouping the most similar objects into a cluster and to determining each object's cluster membership. In other words, whereas an object in a certain cluster should be as similar as possible to all the other objects in the same cluster, it should likewise be as distinct as possible from objects in different clusters.An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data [5].

1.1 CLUSTERING METHODS

There are many clustering methods have been developed, each of which uses a different induction principle. Farley and Raftery suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid based methods. An alternative categorization based on the induction principle of the various clustering methods is presented in (Estivill-Castro, 2000). We discuss some of them here [14, 16].

1.2 PARTITIONING METHODS

Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. To achieve global optimality in partitioned-based clustering, an exhaustive enumeration process of all possible partitions is required. Because this is not feasible, certain greedy heuristics are used in the form of iterative optimization. Namely, a relocation method iteratively relocates points between the *k* clusters. The following subsections present various types of partitioning methods.

1.3 HIERARCHICAL METHODS

These methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. These methods can be subdivided into Agglomerative hierarchical clustering and Divisive hierarchical clustering. In agglomerative hierarchical clustering each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained. In divisive hierarchical clustering all objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained. The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.

1.4 DENSITY-BASED METHODS

Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution (Banfield and Raftery, 1993).The overall distribution of the data is assumed to be a mixture of several distributions. The aim of these methods is to identify the clusters and their distribution parameters. These methods are designed for discovering clusters of arbitrary shape which are not necessarily convex, namely: The idea is to continue growing the given cluster as long as the density (number of objects or data points) in the neighborhood exceeds some threshold. Namely, the neighborhood of a given radius has to contain at least a minimum number of objects. When each cluster is characterized by local mode or maxima of the density function, these methods are called mode-seeking Much work in this field has been based on the underlying assumption that the component densities are multivariate Gaussian (in case of numeric data) or multinomial (in case of nominal data). An acceptable solution in this case is to use the maximum likelihood principle. According to this principle, one should choose the clustering structure Density-based clustering may also employ nonparametric methods, such as searching for bins with large counts in a multidimensional histogram of the input instance space.

1.5 MODEL-BASED CLUSTERING METHODS

These methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering, which identifies groups of objects; model-based clustering methods also find characteristic descriptions for each group, where each group represents a concept or class. The most frequently used induction methods are decision trees and neural networks. Decision Trees. In decision trees, the data is represented by a hierarchical tree, where each leaf refers to a concept and contains a probabilistic description of that concept. Several algorithms produce classification trees for representing the unlabelled data. Neural Networks is used to represents each cluster by a neuron or "prototype". The input data is also represented by neurons, which are connected to the prototype neurons. Each such connection has a weight, which is learned adaptively during learning [6].

II. RELATED STUDY

We get idea from different research material that uniform approach for "Bidirectional Agglomerative Hierarchical Clustering using AVL Tree Algorithm". Proposed Bidirectional agglomerative hierarchical clustering to create a hierarchy bottom-up, by iteratively merging the closest pair of data-items into one cluster. The result is a rooted AVL tree. The *n* leafs correspond to input data-items (singleton clusters) needs to *n/2* or *n/2+1* steps to merge into one cluster, correspond to groupings of items in coarser granularities climbing towards the root. As observed from the time complexity and number of steps need to cluster all data points into one cluster perspective, the performance of the bidirectional agglomerative algorithm using AVL tree is better than the current agglomerative algorithms. One of the advantages of the proposed bidirectional agglomerative hierarchical clustering algorithm using AVL tree and that of other similar agglomerative algorithm is that, it has relatively low computational requirements. The overall complexity of the proposed algorithm is O(*logn*) and need (*n/2* or *n/2+1*) to cluster all data points in one cluster whereas the previous algorithm is O(n^2) and need (*n-1*) steps to cluster all data points into one cluster [7].

This revolution began with "A novel hierarchical clustering algorithm for gene Sequences" .The proposed method is evaluated by clustering functionally related gene sequences and by phylogenetic analysis. In this paper, they presented a novel approach for DNA sequence clustering, mBKM, based on a new sequence similarity measure, DMk, which is extracted from DNA sequences based on the position and composition of oligonucleotide pattern. Proposed method can be applied to study gene families and it can also help with the prediction of novel genes. mBKM with DMk can generate cluster trees that are useful to understand the processes governing the gene evolution. Proposed method may be extended for protein sequence analysis and Meta genomics of identifying source organisms of Meta genomic data [8].

The proposed method based on"A New, Fast and Accurate Algorithm for Hierarchical Clustering on Euclidean Distances"A simple hierarchical clustering algorithm called CLUBS (for Clustering Using Binary Splitting) is proposed in this paper. CLUBS is faster and more accurate than existing algorithms, including k-means and its recently proposed refinements. The algorithm consists of a divisive phase and an agglomerative phase; during these two phases, the samples are repartitioned using a least quadratic distance criterion possessing unique analytical properties that. CLUBS derives good clusters without requiring input from users, and it is robust and impervious to noise, while providing better speed and accuracy than methods, such as BIRCH, that are endowed with the same critical properties. The naturalness of the hierarchical approach for clustering objects is widely

recognized, and also supported by psychological studies of children's cognitive behaviors1. CLUBS is providing the analytical and algorithmic advances that have turned this intuitive approach into a data mining method of superior accuracy, robustness and speed [9].

This idea proposes "Algorithm Portfolios Based on Cost-Sensitive Hierarchical Clustering". Different solution approaches for combinatorial problems often exhibit incomparable performance that depends on the concrete problem instance to be solved. Algorithm portfolios aim to combine the strengths of multiple algorithmic approaches by training a classifier that selects or schedules solvers dependent on the given instance. Proposed algorithm devises a new classifier that selects solvers based on a cost-sensitive hierarchical clustering model. They devised a cost-sensitive hierarchical clustering approach for building algorithm portfolios. The empirical analysis showed that adding feature combinations can improve performance slightly, at the cost of increased training time, while merging cluster splits based on cross-validation lowers prediction accuracy [10].

III. Problem Domain

In a clustering application involving multiple datasets may demand partitioned clusters as the output instead of a dendrogram. Any hierarchical ensemble method can be used to combine the dendrogram generated from multiple datasets and then the resulting single dendrogram can be cut [7] to extract the desired number of partitioned clusters. However, partitioned clusters generated this way may not be of satisfactory quality. This necessitates the need for developing methods for producing partitioned clusters as dendrogram are combined.

In the work [2], the author proposed a graph-based ensemble method EHC (Ensemble for Hierarchical Clustering) that can be used to combine hierarchical clustering generated from multiple contextually related datasets where the datasets are represented by heterogeneous feature sets and may not necessarily contain the same set of objects. This method uses the cluster hierarchies generated from individual dataset and combines them to yield a set of partitioned clusters. Instead of extracting partitioned clusters by cutting a combined dendrogram, EHC directly generates partitioned clusters from two or more dendrogram by capturing the cluster membership of data objects from multiple dendrogram. EHC works by generating an undirected weighted graph using the combined strength of association of each pair of objects in the dendrogram. Each object is represented by a vertex in this graph and the strength of association between each pair of objects is represented as the weight of the edge connecting the corresponding vertices. The purpose is to bring together the objects that are strongly associated with each other in the form of a sub-graph [11].

EHC based approach is focusing on a specific area of documents arrangement & retrieval problem for proposed approach as application area. The document clustering domain, when heterogeneous feature sets are available to represent a set of documents, EHC yields higher quality clusters than hierarchical clustering based on individual feature sets and hierarchical clustering based on unified feature sets. The algorithm also outperforms the super-tree and consensus tree methods and *k*-means based on unified feature sets [2]. Even though EHC does not perform as well as graph-based clustering on unified feature sets when two congruent datasets are used, EHC significantly outperforms all other baseline methods including graph-based partitioning when two semi-congruent datasets are used. EHC is particularly helpful in situations where hierarchical clustering is performed using heterogeneous feature sets at multiple sites, but the datasets are not accessible after the clustering is complete. Also, the EHC algorithm is easily parallelizable since the association strengths for individual dendrogram can be computed independently.

Some of the approaches also lead towards non parametric data arrangements like in [16]. This paper presents a comparison of strategies for non-parametric document ensemble clustering [12].

Concerning Issues

The important problems with ensemble based cluster analysis that this work have identified are as follows:

Problem 1: The identification of distance measure, for numerical attributes, distance measures can be used. But identification of measure for categorical attributes in strength association is difficult.

Problem 2: The number of clusters, identifying the number of clusters & its proximity value is a difficult task if the number of class labels is not known in advance. A careful analysis of inter & intra cluster proximity through number of clusters is necessary to produce correct results.

Problem 3: Structure of database, Real life data may not always contain clearly identifiable clusters. Also the order in which the tuples are arranged may affect the results when an algorithm is executed if the distance measure used is not perfect. With a structure less data (Having lots of missing values), identification of appropriate number of clusters will not yield good results. So some sort of global objective function needs to be defined.

Problem 4: Types of attributes in a database, the databases may not necessarily contain distinctively numerical or categorical attributes. They may also contain other types like nominal, ordinal, binary etc. So these attributes have to be converted to categorical type to make calculations simple.

Problem 5: Classification of Ensemble Clustering Algorithm, Clustering algorithms can be classified according to the

method adopted to define the individual clusters. So which algorithm is used for specific purpose is not mentioned.

Problem 6: Merging decision in not given, Hierarchical clustering tends to make good local decisions about combining two clusters since it has the entire proximity matrix available. However, once a decision is made to merge two clusters, the hierarchical scheme does not allow for that decision to be changed. This prevents a local optimization criterion from becoming a global optimization criterion.

From the above mentioned problems of generalize approach of ensemble based methods the few basic question. We believe that two questions remain unanswered in the state of the art with respect to the use of ensemble methods for document clustering: PROPOSED

IV. PROPOSED ENSEMBLE BASED HIERARCHICAL CLUSTERING ALGORITHMS

The proposed approach is used on generation of ensembles based cluster on the basis of few operations like mapping & combination. These operations can be performed with the help of two operators' similarity association & probability for correct classification or classifier analysis of cluster. In this proposed approach our main aim is to identify the cluster partitional data for hierarchical clustering. It may be represented via parametric representation of nested clustering & Dendograms.

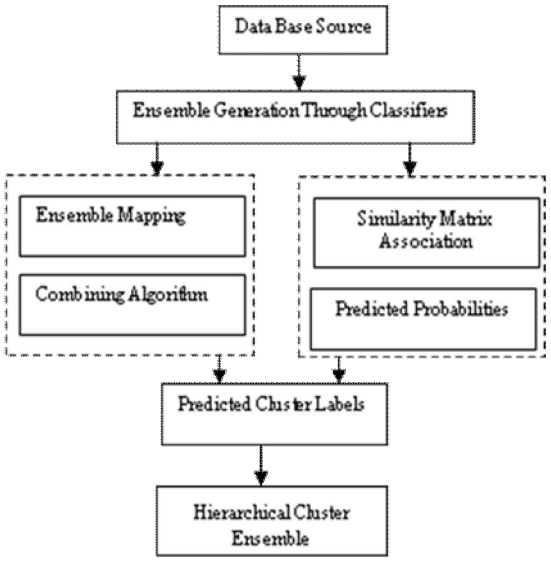


Fig. 2 working process of proposed method

According to the proposed approach initially the dataset of various features having n element in inserted which is arranged in dendrogram. Based on this features similar or dissimilar values may be separated. It gives various representations of same data sets but the views are same & the partition logic is separately mentioned. We need to categorize the boundaries properly & place the element in correct cluster. So to identify the correct element of each cluster strength of bond or dependencies is calculated which later shown in weak or strong manner. Assign those associations a value term as weights of vertex or dendograms. Now the ensemble mapping is achieved by clustering feature parameter of the most strong nearest entity. It gives the labelled value of mapped data and can be shown as similarity association metrics. We view the similarity association of two objects as a measure of how closely they are associated in the different hierarchical clustering. We assume that it is a measure of proximity of the two objects.

4.1 PROPOSED ARCHITECTURE

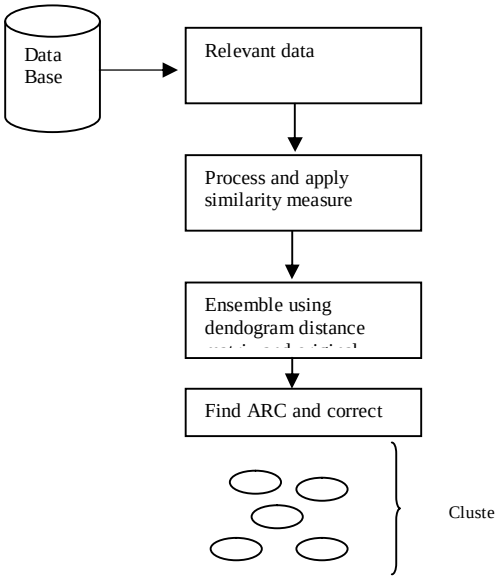


Fig. 3 working process of proposed method

PROPOSED ALGORITHMS

1. Assign each object as individual cluster like c_1, c_2, c_3, .. c_n where n is the no. of objects
2. Find the distance matrix D, using any similarity measure
3. Find the closest pair of clusters in the current clustering, say pair (r), (s), according to d(r, s) = mind (i, j) { i, is an object in cluster r and j in cluster s}
4. Merge clusters (r) and (s) into a single cluster to form a merged cluster. Store merged objects with its corresponding distance in Dendrogram distance Matrix.
5. Update distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s). Adding a new row and column corresponding to the merged cluster(r, s) and old luster (k) is defined in this way: d[(k), (r, s)] = min d[(k),(r)], d[(k),(s)]. For other rows and columns copy the corresponding data from existing distance matrix.
6. If all objects are in one cluster, stop. Otherwise, go to step 3.

7. Find association relation coefficient value with single, complete and average linkage methods.

V. RESULT EVALUATION

In the results will show the effectiveness of proposed scheme

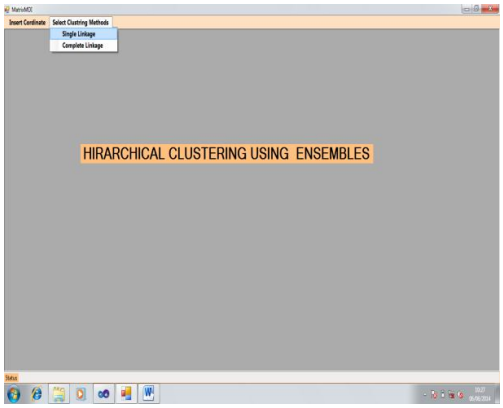


Fig.4 Methods select form

This snapshot show the working of single linkage method . In this snapshot distance matrix and object list is shown. There is open graph option for displaying the plotted object in two dimensinal plan. This snapshot also display the required execution time and memory used making final clusters. The execution time is cacluated in milliseconds and required memory is calculated in kilobytes .

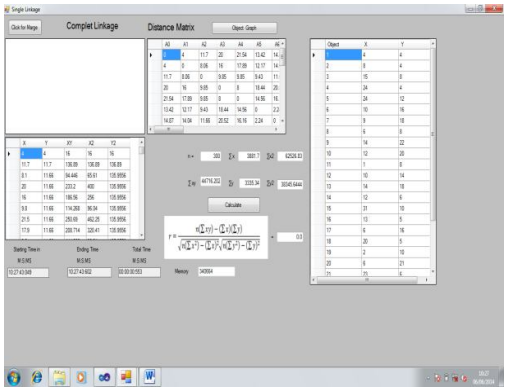


Fig. 5 single linkage method working process

Cluster forming (merging) using Single linkage Method

This snapshot displays the merging process for clustering. When user click on calculate button the accuracy value is shown in the text box. From the click for merge button user can see the step by step merging of clusters. [13]

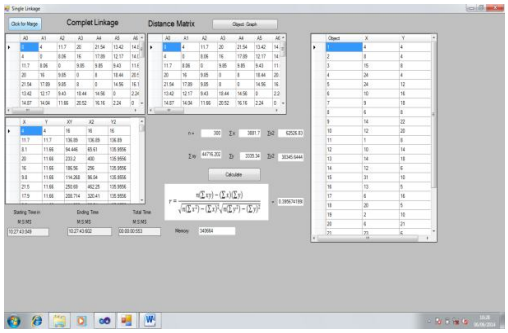


Fig. 6 Display the merging process for single linkage method

Objects Position on two dimensional plan

The following snapshots show the position of the object in two dimensional plans. This is a graph dynamic which display the inserted object position and automatically increase the coordinated value in both the directions. This Snapshot show the 25 object in the database

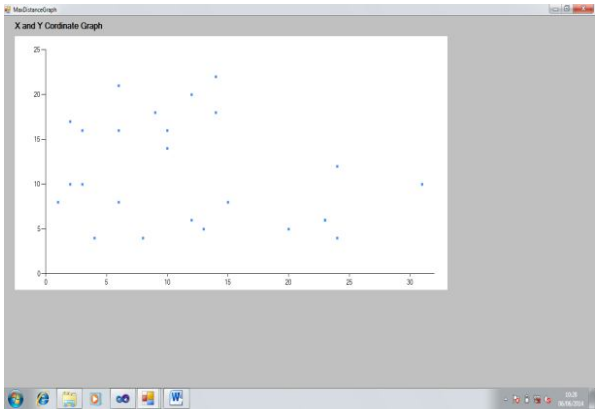
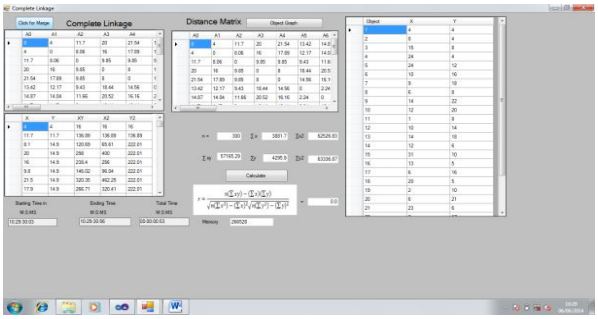


Fig. 7 25 objects position of two dimensional plan

Cluster forming (merging) using complete linkage Method

This snapshot displays the merging process for clustering. When user click on calculate button the accuracy value is shown in the text box. From the click for merge button user can see the step by step merging of clusters. [14]



.

Fig. 8 Display the merging process for complete linkage objects.

RESULT

Number of Objects and accuracy

TABLE I
Number objects and accuracy for single linkage and complete linkage

Number of Objects	Single Linkage	Complete Linkage
50	0.395674	0.33396
100	0.155668	0.158981
150	0.282241	0.224759

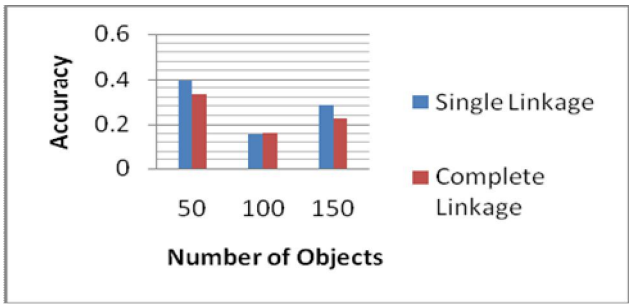


Fig. 9 Comparison with Number of objects and accuracy

IV. CONCLUSION

There are several algorithms and methods have been developed for clustering problem. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency The most popular agglomerative clustering procedures are Single linkage ,Complete linkage , Average linkage and Centroid. [15]

Each of these linkage algorithms can yield totally different results when used on the same dataset, as each has its specific properties. The complete-link clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, yet the single-link methods are more versatile. Final conclusion is that the all methods are fine but to select a method for a given Situations it depends the nature of the objects.[16]

In future enhancement we can also apply various other techniques for ensembling clusters like neural network, fuzzy logic, genetic algorithms etc to enhance the clustering.

VI. FUTURE WORK

Our proposed methods are based on clustering ensemble and association which is a probability measure. Clustering Association coefficient has he value between 0 to 1. The methods which has clustering Association coefficient near to 1 is more accurate method for a given data set. So this is a probability measure which is not 100% true.

VII. ACKNOWLEDGMENT

This research work is self financed but recommended from the institute so as to improve the Ensemble Based Hierarchical Clustering Method for Data Separation Using Various Feature Matrices Thus, the authors thank the anonymous reviewers for their valuable comments, which strengthened the paper. The authors also wish to acknowledge institute administration for their support & motivation during this research. They also like to give thanks to Mr. Ritesh Jain for discussion regarding the situational awareness system & for producing the approach adapted for this paper.

REFERENCES

[1] Mayank Gupta, Dhanraj Verma "A Novel Ensemble Based Cluster Analysis Using Similarity Matrices & Clustering Algorithm (SMCA)" in International Journal of Computer Application Vol. 100, No.10, ISBN 973-93-80883-40-8, 20 Augest 2014. pp. 1-6

[2] J. Han, M. Kamber, Data mining, Concepts and techniques, Academic Press, 2003.

[3] Arun K. Pujari, Data mining Techniques, University Press (India) Private Limited, 2006.

[4] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, Prentice Hall of India, 2004

[5] Nachiketa Sahoo Incremental Hierarchical Clustering of Text Documents May 5, 2006

[6] Sanjoy Dasgupta Philip M. Long Performance guarantees for hierarchical Clustering Preprint submitted to Elsevier Science 24 July 2010

[7] Tapas Kanungo, Nathan S. Netanyahu "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 7, July 2002.

[8] R. M. Castro, M. J. Coates, R. D. Nowak, Member, IEEE Department of Electrical and Computer Engineering, Rice University, MS366, Houston, TX 77251-1892 USA

[9] Matej Franceti, Mateja Nagode, and Bojan Nastav Hierarchical Clustering with Concave Data Sets Metodoloski zvezki, Vol. 2, No. 2, 2005, 173-193

[10] Ming-Chuan Hung, Jungpin Wu, Jin-Hua Chang and Don-Lin Yang "An Efficient k-Means Clustering Algorithm Using Simple Partitioning "Journal of Information Science And Engineering 21, 1157-1177 (2005).

[11] Yi Lu Lily R. Liang Hierarchical Clustering of Features on Categorical Data of Biomedical Applications Computer Science Department Prairie View A&M University Prairie View, Texas, 77446, USA.

[12] Dar-Jen Chang, Mehmed Kantardzic, Ming Ouyang Hierarchical clustering with CUDA/GPU Computer Engineering & Computer Science Department University of Louisville Louisville, Kentucky 40292

[13] Mahmood Hossain, Susan M. Bridges, Yong Wang, and Julia E. Hodges "An Effective Ensemble Method for Hierarchical Clustering " June 27-29, Montreal, QC, CANADA Editors: B. C. Desai, S. Mudur, E. Vassev Copyright c_2012 ACM 978-1-4503-1084-0/12/06.

[14] Xiaoke Su, Yang Lan, Renxia Wan, and Yuming Qin " A Fast Incremental Clustering Algorithm" ISBN 978-952-5726-02-2 (Print), 978-952-5726-03-9 (CD-ROM) Proceedings of the 2009 International Symposium on Information Processing (ISIP'09)

[15] Revati Raman Dewangan, Lokesh Kumar Sharma, Ajaya Kumar Akasapu Fuzzy Clustering Technique for Numerical and Categorical dataset Revati Raman Dewangan et al. / International Journal on Computer Science and Engineering (IJCSE) NCICT 2010 Special Issue.

[16] Parul Agarwal, M. Afshar Alam, Ranjit Biswas Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes International Journal of Innovation, Management and Technology, Vol. 1, No. 2, June 2010 ISSN: 2010-0248