International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 9 2590 – 2595

Analysis of Clustering Approaches for Data Mining In Large Data Sources

Srinivas Sivarathri
Research Scholar,
Department of Computer Science
Acharya Nagarjuna University
Guntur , Andhra Pradesh, India
Email:- vasulibya@gmail.com

A.Govardhan
Director,
School of Information Technology
Jawaharlal Nehru Technological University
Hyderabad, Telangana , India
Email ;- govardhan_cse@yahoo.co.in

Abstract-A plethora of algorithms exist for clustering to discover actionable knowledge from large data sources. Given un-labeled data objects, clustering is an unsupervised learning to find natural groups of objects which are similar. Each cluster is a subset of objects that exhibit high similarity. Quality of clusters is high when they feature highest intra-cluster similarity and lowest inter-cluster similarity. The quality of clusters is influenced by the similarity measure being employed for grouping objects. The clustering quality is measured the ability of clustering technique to unearth latent trends distributed in data. The usage of data mining technique clustering is ubiquitous in real time applications such as market research, discovering web access patterns, document classification, image processing, pattern recognition, earth observation, banking, insurance to name few. Clustering algorithms differ in type of data, measure of similarity, computational efficiency, and linkage methods, soft or hard clustering and so on. Employing a clustering technique correct depends on the technical knowhow one has on various kinds clustering algorithms and suitable scenarios to apply them. Towards this end, in this paper, we explore clustering algorithms in terms of computational efficiency, measure of similarity, speed and performance.

Index Terms - *Data mining, clustering techniques, similarity measure, unsupervised learning*

*****

I. INTRODUCTION

Clustering is generally carried out with an unsupervised learning approach using certain error functions in order to minimize the distances. The supervised clustering is different from this traditional clustering as it is used with examples which have been classified. The objective of this clustering approach is to identify the probability density from the standpoint of a single class. In case of supervised learning it is also required to ensure the number of clusters [1]. Cluster analysis has various elements such as representing data, choice of objects, choice in choosing variables, determining what to cluster, normalization, determining similarity measure, choosing objective function, determining missing data strategy, algorithm, number of clusters required and results interpretation [2]. K-means is widely used algorithm which has been around for many years. It still occupies its presence in the top 10 clustering algorithms. It is simple and effective. However it causes more computational overhead while calculating centroids in the process of convergence. Still it has plenty of utility as it can be customized to meet the requirements [3]. Spatial databases with noise can be used for clustering purposes. However, input parameters are important for the success of such clustering algorithms. The requirements of both domain knowledge and discovery of clusters are to be fulfilled. DBSCAN is a clustering algorithm that is density based. It needs one parameter as input from user for making clusters. For discovering clusters of arbitrary shape this algorithm is every effective [4]. In the same fashion, density based clustering can be applied to uncertain data too. Uncertain data is produced by distance computation between objects, face recognition systems, location based services and sensor databases [5].

There are some scenarios in the real world where there is not sharp boundary between clusters. For such scenarios fuzzy clustering is best used. Fuzzy clustering facilitates mentioning of membership degree so as to make it very flexible. This kind of clustering is also known as soft clustering [6]. The top ten data mining algorithms are C4.5, K-Means, Support Vector Machines (SVMs), Apriori, Expectation Maximization, PageRank, AdaBoost, kNN (k-nearest neighbor classification), Naïve Bayes, and Classification and Regression Trees (CART) [7].Open source tools for data mining are R [8], Tanagra [9], Weka [10], YALE [11], KNIME [12], Orange [13], and GGobi [14]. ALCALÁ-FDEZ [15] described the importance of evolutionary algorithms (EAs). EAs are meant for optimization purposes. They are very powerful search techniques in the domain of artificial intelligence. Knowledge extraction is the main motivation for applying EAs which became promising techniques in the field of data mining [16], [17], [18], and [19]. Clustering of uncertain data has been studied for clustering. In [20] and [21] experiments are done on variants of K-means. In case of uncertain data we hot got K-means variants such as K-median exhibited no uncertainties. Based on DBSCAN and optics are explored in FDBSCAN and FOPTICS respectively. Regions in the data with high density are identified these algorithms based probability density functions.

II. CLUSTERING METHODS

In the context of data mining, the term cluster represents a group of objects which are similar. According to Estivill-Castro there is no precise definition as such. Hierarchical clustering approach is deterministic and it does not need to specify number of clusters priori [22] and [23]. Agglomerative hierarchical clustering is the most widely used algorithm for embedded classification schemes [24].

 ISSN: 2321-8169
2590 – 2595

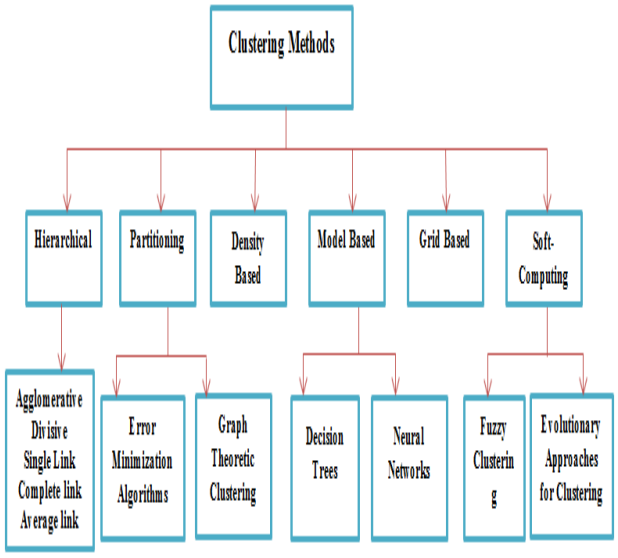

Fig. 1 – Classification of Clustering Algorithms

Hierarchical clustering is a process of constructing clusters recursively in bottom-up or top-down fashion. It generates a hierarchy of clusters. It is of two types namely agglomerative and divisive. In the former approach initially every object forms a cluster and later on merged into different clusters. In the latter initially all objects belong to a single cluster. Then the objects are gradually divided into sub clusters. The nested groups of objects are the end result of either approach. Based on the manner in which similarity measure is used, the hierarchical clustering is of three types. They are single-link clustering, complete-link clustering, and average link clustering [33]. The strength of hierarchical clustering lies in versatility and multiple partitions. The drawbacks of hierarchical clustering include no backtracking capability and less scalability [25]. Partitioning clustering is a technique in which achieve group of clusters by relocating objects from one cluster to another cluster after initial partitioning. Generally they need to know the number of cluster priori. Partitioning algorithms are of two types namely error minimization algorithms and graph theoretic clustering. The error minimization algorithms are used to reduce error rate in the clustering process. They make use of Sum of Squared Error (SSE) to achieve this. Graph theoretic methods are clustering methods that produce clusters through graphs. Examples for this approach are Limited Neighborhood Sets [33] and Minimal Spanning Tree [35]. Density Based Clustering is a process making clusters based on density. The objects that are with low density clusters are known as noise or outliers. This is best used to get rid of outliers. DBSCAN is best example for this kind of clustering [36].

Model Based Clustering Methods are the techniques that make use of mathematical models in making clusters. The most widely used methods of this kind are decision trees and neural networks. Decision trees represent data in the form of hierarchical trees while neural networks represent each cluster as a neuron [25]. Grid Based methods divide the space into a grid on which clustering operations are performed [36]. Soft computing techniques are also used for clustering. For instance fuzzy clustering [37] and evolutionary approaches are soft computing techniques used for clustering. Fuzzy C Means [29] is the best example for fuzzy clustering where Genetic Algorithms are is used for evolutionary approaches.

III. MEASURES USED IN CLUSTERING

A measure is essential to determine similarity between objects. There are two types of measures namely distance measures and similarity measures [25]. Distance between two objects x_i and $x_j$is denoted as $d(x_i, x_j)$. Distance measures can be used for numeric attributes, binary attributes, nominal attributes, ordinal attributes, and mixed-type attributes. According to Han and Kamber [26], for numeric attributes Minkowski metric is used to compute distance between two instances. Given xi, xj instances the distance is measured as follows.

$$d(x_i, x_j) = \left(|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \ldots + |x_{ip} - x_{jp}|^g\right)^{1/g} \quad (1)$$

For binary attributes distance is measured based on the contingency table. If the attributes are symmetric simple matching coefficient is used as (2) and Jaccard coefficient is used to for asymmetric attributes (3).

$$d(x_i, x_j) = \frac{r + s}{q + r + s + t} \quad (2)$$

$$d(x_i, x_j) = \frac{r + s}{q + r + s} \quad (3)$$

For nominal attributes simple matching (4) and binary attribute approach are followed. When the attributes are original they are mapped numeric ones. Mapping is done as (5).

$$d(x_i, x_j) = \frac{p - m}{p} \quad (4)$$

$$z_{i,n} = \frac{r_{i,n} - 1}{M_n - 1} \quad (5)$$

When the attributes are of mixed type the dissimilarity between objects is computed as (6). It combines two or more distance measures described above.

 ISSN: 2321-8169
2590 – 2595

$$d(x_i, x_j) = \frac{\sum_{n=1}^{p} \delta_{ij}^{(n)} d_{ij}^{(n)}}{\sum_{n=1}^{p} \delta_{ij}^{(n)}} \quad (6)$$

The similarity measures used to find the similarity between two objects include cosine measure, Pearson correlation measure, extended Jaccard measure and dice coefficient measure. A similarity measure results in a value between 0.0 and 1.0. Zero indicates dissimilar, 1 indicates completely similar while other values indicate the degree of similarity or dissimilarity. Their details are shown in table 1.

Similarity Measure	Equation
Cosine	$s(x_i, x_j) = \frac{x_i^T \cdot x_j}{\|x_i\| \cdot \|x_j\|}$ (7)
Pearson Correlation	$s(x_i, x_j) = \frac{(x_i - \bar{x}_i)^T \cdot (x_j - \bar{x}_j)}{\|x_i - \bar{x}_i\| \cdot \|x_j - \bar{x}_j\|}$ (8)
Extended Jaccard	$s(x_i, x_j) = \frac{x_i^T \cdot x_j}{\|x_i\|^2 + \|x_j\|^2 - x_i^T \cdot x_j}$ (9)
Dice Coefficient	$s(x_i, x_j) = \frac{2x_i^T \cdot x_j}{\|x_i\|^2 + \|x_j\|^2}$ (10)

Table 1 – Similarity measures

The measures are used to find similarity between objects to make clustering decisions. There are many metrics for evaluating clusters or finding quality of clusters thereby the utility of given clustering method. However, as Bonner [27] said there is no universal definition for determining what is good clustering. The criteria available for finding quality of clusters are of two types namely internal (compactness) and external (structure). Internal criteria include Sum of Squared Error (SSE), minimum variance criteria, scatter criteria, Condorcet's criterion, the C-Criterion, category utility metric, and edge cut metrics [25]. The external criteria include mutual information based measure, precision-recall measure and rand index [25].

IV. K-MEANS AND FUZZY K-MEANS

K-means is one of the clustering algorithms in data mining. It has been around for many years. It is in the top ten clustering algorithms. K-means is very flexible and good for grouping similar objects. It is widely used in many real world applications [28]. For instance it can be used in credit card fraud detection systems employed in banking domain. K-means algorithm takes two inputs such as dataset and also the number of clusters. It generates specified number of clusters. However, it is very expensive for large datasets. K-means generates k clusters always. Each cluster should have one item at least. The clusters are not overlapped. There is highest similarity between objects within cluster and lowest similarity between objects of different clusters. When compared to hierarchical clustering, K-means is computationally faster if number of clusters is less while it produce tighter clusters when the clusters are globular. K-means has several limitations. They include finding quality of clusters is difficult; prediction of k is difficult; With non globular clusters it does not work properly; final clusters depend on initial partitions [28].

Fuzzy K-Means

It is an extension to K-means algorithm. It is also known as Fuzzy C-Means. K-means makes hard clusters while Fuzzy K-Means makes soft clusters. When a data object belongs to only one cluster we call it hard cluster. In case of soft cluster an object can belong to multiple clusters with probability. It also uses distance measure that acts on objects kept in n-dimensional vector [29]. Fuzzy C Means is a widely used algorithm for clustering objects. Bezdek [31], [32] developed this algorithm originally. It is used to solve real world problems with fuzzy intelligent control. Classification of patterns, classification of network faults are some of the applications of Fuzzy C Means. This algorithm uses a distance measure such as Euclidean Distance which helps in making clustering decisions. The similarity between two objects is between 0.0 and 1.0. The value 0.0 represents no similarity while 1.0 indicates highest similarity. Feature vectors are generated by FCM and they are kept into clusters with associated values. Such value is between 0 and 1 known s fuzzy truth value.

The main purpose of FCM is mapping set of given representatives to improved ones by partitioning data points. Fuzzy C Means algorithm has the following steps. In step 1 initial cluster centers SC_0= {Cj(0)} and set p=1. In the second step computer cluster centers and update memberships as follows.

$$u_{i,j} = \left((d_{ij})^{1/m-1} \sum_{l=1}^{k} \left(\frac{1}{d_{il}} \right)^{1/m-1} \right)^{-1}$$

In step 3, compute the cluster center for each cluster in order to obtain new cluster representatives as follows.

$$\mathbf{C}_j(p) = \frac{\sum_{i=1}^{N} u_{ij}^m \mathbf{X}_i}{\sum_{i=1}^{N} u_{ij}^m}$$

In step 4, $\text{If } \|\mathbf{C}_j(p) - \mathbf{C}_j(p-1)\| < \varepsilon \text{ for } j = 1$ to k then step. Otherwise set p+1 → p and move to step 2.

ISSN: 2321-8169
2590 – 2595

As can be seen, major computational complexity of FCM is in steps 2 and 3. The computational complexity of 3 is comparatively less that that of 2. Therefore the computational complexity of the algorithm can be reduced in terms of number of distance calculations [38]. Pre-processing algorithms such as Canopy Clustering can be used to reduce the computational complexity of the FCM and speed up its process while developing clusters. In our future work we are going to combine the FCM and Canopy clustering algorithms in order to improve the performance of FCM.

Chang et al. [38] made such experiments to improve FCM by building two more algorithms such as CDFKM and MCDFKM. For a given dataset the average computing time in seconds is found and the results are as presented in table 1.

Method	*k*			
	16	32	64	128
FKM	509.50	1328.92	2925.56	15355.94
CDFKM	468.86	1212.73	2329.64	7920.42
MCDFKM (*M* = 1)	338.67	894.00	3192.59	7623.59
MCDFKM (*M* = 2)	349.08	911.80	3434.36	7791.73
MCDFKM (*M* = 3)	394.97	983.69	3678.59	8404.13
MCDFKM (*M* = 4)	412.97	1072.31	4069.98	10147.78
MCDFKM (*M* = 5)	475.51	1138.83	4390.61	10205.95
MCDFKM (*M* = 6)	530.64	1294.08	5385.81	14342.30

Table 2 – Experimental Results (excerpt from [38])

As can be seen in table 2, it is evident that the computational complexity of CDFKM and MCDFKM are much lower when compared to FKM. The results are graphically presented in figure 2.

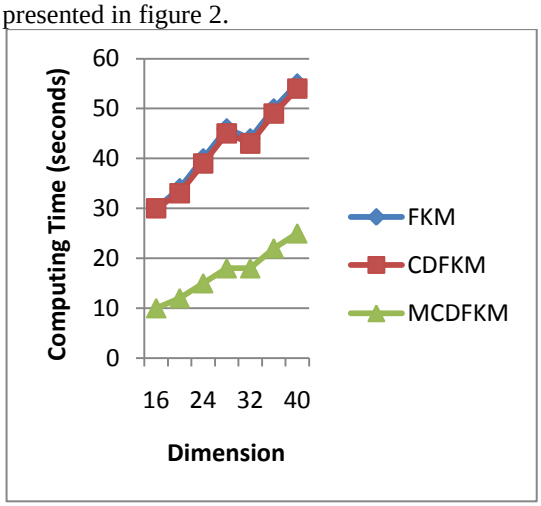

Figure 2 – Average computing time for FKM, CDFKM, and MCDFKM (excerpt from [38])

As seen in Figure 2, the results of the three algorithms are presented graphically. The CDFKM and MCDFKM outperform the traditional FKM.

Canopy Clustering

It is very simple and fast and can produce clusters accurately. In this case objects are represented as points in the feature space which is essentially multidimensional. Fast approximate distance measure is used in canopy clustering. For processing the clusters it makes use of two distance thresholds with T1>T2. Algorithm starts with a set of initial points [30]. At each point distance is measured and grouping decision is applied. When a point is <T1 add it to the cluster and when it is >T2 remove it from the cluster. By the end of the process, the algorithm produces a set of canopies. Each canopy is a group of objects which are similar. An object may belong to more than one canopy. This algorithm is best used as pre-processing approach to clustering techniques like K-means. This can reduce computational expenses by starting with initial clustering as it can ignore points which are not in the purview of canopies [30].

More Methods, Issues and Challenges of Clustering

Single Link Clustering Algorithm

With respect to document clustering, the single link clustering algorithm follows an approach in which the similarity between two clusters represents the actual similarity of their most similar members. It does mean that attention is paid on only the members who are closest to each other while other members who are more dissimilar are not considered. Here the criterion is local [24].

Complete Link Clustering Algorithm

In complete-link clustering, the similarity of two clusters is represented by their members that are highly dissimilar. The criterion for this kind of clustering is non-local. The merge decisions are influenced by the entire structure of clustering. This kind of clustering is sensitive to outliers. When any document is far from the center, it has the capacity to increase diameter and thus can have influence on final clustering while making besides making merge decisions [24].

Group-Average Agglomerative Clustering (GAAC)

GAAC makes use of all similarities between documents to be clustered by evaluating cluster quality. Thus it is capable of overcoming the drawbacks of both single link and complete link clustering algorithms [24].

Decisive Clustering

Top-down clustering is also known as decision clustering. It is complex when compared with bottom-up clustering. However, it is more efficient when complete hierarchy need not be generated. It is linear with respect to number of documents and said to be much faster than other algorithms [24].

Centroid Clustering

In this kind of clustering, the similarity of documents is determined by the centroids of respective clusters. Thus it is different from that of GAAC which considers all pairs of documents for making clustering decisions [24].

DBSCAN Algorithm

It is an algorithm that can mine clusters of various shapes and useful for spatial data mining. It takes each object and searches for neighbor of object to find whether for more than the minimum number of objects [24].

Self Organizing Map (SOM)

It is one of the neural networks algorithms which make use of a single – layered network. "Winner-takes-all" is the fashion in which it follows for learning. The neurons try to find the current instance and neural whose weight vector appears closest is considered winner. The winner and its neighbor learn by adjusting their weights. It is best used in speech recognition and vector quantization [24].

Issues and Challenges in Clustering

There are many algorithms that automate the process of clustering. The following are the issues or challenges pertaining to clustering algorithms.

- It is not easy to choose a suitable algorithm based on dataset to be used for clustering.
- Different algorithms might produce quite different results on same dataset.
- The result of clustering algorithm is based on the kind of dataset given to it.
- Size of dataset also has its influence on the results of clustering algorithms.
- Variety of data present in dataset also makes the clustering algorithms to behave differently.
- Clustering algorithms might produce results with uncertainty as they do not focus on all requirements simultaneously.
- Clustering algorithms heavily depend on distance function and at the mercy of efficiency of such function.

V. DISCUSSION

Two important things such as clustering methods and similarity measures are focused in this paper. The clustering methods are used for grouping objects. As there are many real world applications that need these methods, they assume significance. Every clustering technique needs some underlying similarity measure. For this reason various similarity and distance measures are presented in this paper. Moreover the quality of clusters depends on the efficiency of similarity measure. Various kinds of clustering algorithms that meet varied requirements of real world applications are presented in this paper in some detail. The main clustering approaches discussed include hierarchical, partitioning, density based, model based, and soft computing methods. We also provided some of the open source tools used for data mining including Weka. It also covers similarity measures and distance measures. The similarity measures presented in this paper include cosine similarity, Pearson correlation, extended Jaccard and dice coefficient. We also discussed K-Means its limitations, fuzzy K-Means and Canopy clustering. Fuzzy K-Means is one of the soft computing techniques which can be used along with other techniques to solve complex problems in the real world. On the other hand, the Canopy clustering is best used as pre-processing step to clustering method. From the insights from review of literature we could envisage an important research are that is to combine the methods such as Canopy clustering and Fuzzy K-Means for high quality clusters.

VI. CONCLUSION AND FUTURE WORK

In this paper we studied data mining algorithms especially clustering methods and various similarity measures that are used by them. Clustering has important utility in real world applications. The quality of clusters depends on the quality of the similarity measure used for finding similarity between objects. Knowledge about various clustering methods, the type of data, computational efficiency, similarity measure, type of clustering (soft or hard), linkage methods will help in making important development decisions while solving real world problems. In this paper we explored many clustering methods in terms of computational efficiency, measure of similarity, speed and performance. We have many directions for future work. First, evaluating the performance of Fuzzy K-Means; second, evaluating the processing efficiency of Canopy Clustering and third, exploring the combination of Fuzzy K-Means and Canopy Clustering for clustering efficiency.

REFERENCES

[1] NidalZeidat and Christoph F. Eick. (n.d). K-medoid-style Clustering Algorithms for Supervised Summary Generation. *IEEE*.0 (0), p1-7.

[2] Sami AyramoTommiKarkkainen. (2006). Introduction to partitioningbased clustering methods with a robust example. *University of Jyvaskyl*.0 (0), p1-36.

[3] MING-CHUAN HUNG, JUNGPIN WU+, JIN-HUA CHANG AND DON-LIN YANG. (2005). An Efficient k-Means Clustering Algorithm Using Simple Partitioning. *JOURNAL OF INFORMATION SCIENCE AND ENGINEERING*.0 (0), p1157-1177.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, XiaoweiXu. (n.d). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD-96*.0 (0), p1-6.

[5] Hans-Peter Kriegel and Martin Pfeifle. (2005). Density-Based Clustering of Uncertain Data. *KDD'05*.0 (0), p672-677.

[6] Hoppner F., Klawonn F., Kruse R., Runkler T: Fuzzy Cluster Analysis. Wiley (1999)

[7] XindongWu. (2008). Top 10 algorithms in data mining. *KnowlInf Syst*. 0 (0), p1-13.

[8] r-project. (n.d). The R Project for Statistical Computing. *R-Project*.0 (0), p1-4.

[9] Eric.univ. (2013). Le laboratoireERIC . *Eric*. 0 (0), p1-4.

[10] Weka. (2013). Weka 3: Data Mining Software in Java. *waikato*. 0 (0), p1.

[11] Rapid. (2013). Open source software for big data analytics.. *Rapid*.0 (0), p1.

[12] knime. (2012). *KNIME - Professional Open-Source Software.* Available: http://www.knime.org/. Last accessed 15th Oct 2013.

[13] Orange. (2012). *Orange Software.* Available: http://orange.biolab.si/. Last accessed 15th Oct 2013.

[14] Ggobi. (2010). *GGobi Making plots of data is just smart*. Available: http://www.ggobi.org/. Last accessed 10th Oct 2013.

[15] J. ALCALÁ-FDEZ, A. FERNÁNDEZ, J. LUENGO, J. DERRAC, S. GARCÍA, L. SÁNCHEZ AND F. HERRERA. (2011). KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *J. of Mult.-Valued Logic & Soft Computing*.17 (0), p255-287.

[16] M.L. Wong and K.S. Leung. (2000). *Data mining using grammar based genetic programming and applications*. Kluwer Academic Publishers.

[17] S.K. Pal and P.P.Wang.(1996). *Genetic Algorithms for Pattern Recognition*. CRC Press

[18] J.J. Grefenstette. (1993). *Genetic Algorithms for Machine Learning*.Kluwer Academic Publishers.

[19] A. Ghosh and L.C. Jain. (2005). *Evolutionary Computation in Data Mining*. Springer- Verlag.

[20] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, "Efficient clustering of uncertain data," in *IEEE International Conference on Data Mining (ICDM) 2006*, pp. 436–445.

[21] M. Chau, R. Cheng, B. Kao, and J. Ng, "Data with uncertainty mining: An example in clustering location data," in *Proc. of the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD 2006)*, 2006.

[22] Estivill-Castro, V. and Yang, J. A Fast and robust general purpose clustering algorithm.Pacific Rim International Conference on Artificial Intelligence, pp. 208-218, 2000.

[23] FionnMurtagh and Pedro Contreras (2011). *Methods of Hierarchical Clustering*. London: IEEE. 21

[24] Cambridge University (2009). *Hierarchical clustering*. London: IEEE. 26.

[25] LiorRokach and OdedMaimon, "Clustering Methods" . Available online at: <http://www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf> [Accessed: 10 October 2013]

[26] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[27] Bonner, R., On Some Clustering Techniques. IBM journal of research and development, 8:22-32, 1964.

[28] KhaledAlsabti, Sanjay Ranka and Vineet Singh.(n.d). An Efficient K-Means Clustering Algorithm. *Information Technology Lab (ITL)*.0 (0), p1-6.

[29] CHIH-TANG CHANG, JIM Z. C. LAI AND MU-DER JENG. (2011). A Fuzzy K-means Clustering Algorithm Using Cluster Center Displacement.*JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 27, 995-1009*. 0 (0), p995-1009.

[30] Jeff Eastman. (2012). Canopy Clustering. Available: https://cwiki.apache.org/confluence/display/MAHOUT/Canopy+Clustering. Last accessed 10th Oct 2013.

[31] J. Bedeck, R. Ehrlich and W. Full, "FCM: the fuzzy c-means clustering algorithm," Computers & Geosciences, vol. 10, pp.191-198, 1984.

[32] J. Bedeck, R. Hathaway, M. Sabin and W. Tucker, "Convergence theory for fuzzy c-means: counterexamples and repairs," IEEE Transactions on Systems, Man and Cybernetics, vol. 17, pp. 873-877, 1987.

[33] Jain, A.K. Murty, M.N. and Flynn, P.J. Data Clustering: A Survey. ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[34] Zahn, C. T., Graph-theoretical methods for detecting and describing gestalt clusters. IEEE trans. Comput. C-20 (Apr.), 68-86, 1971.

[35] Urquhart, R. Graph-theoretical clustering, based on limited neighborhood sets. Pattern recognition, vol. 15, pp. 173-187, 1982.

[36] Jiawei Han, Micheline Kamber. (2001). *Data Mining: Concepts and Techniques*. London, United Kingdom: Academic Press.

[37] Hoppner F. , Klawonn F., Kruse R., Runkler T., Fuzzy Cluster Analysis,Wiley, 2000.

[38] CHIH-TANG CHANG, JIM Z. C. LAI AND MU-DER JENG, "A Fuzzy K-means Clustering Algorithm Using Cluster Center Displacement", JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 27, 995-1009 (2011) , p1-15