International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

A link concerning various clusters using hierarchical clustering Techniques

[1] D Veeraiah, [2] Dr. D Vasumathi
[1] Research Scholar, Dept of CSE, JNTUK, Kakinada, Andhra Pradesh, India
[2] Professor, Department of CSE, JNTUCEH, Hyderabad, Andhra Pradesh, India

***Abstract*— The process of grouping a set of physical or abstract object into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to another within the same cluster and are dissimilar to the objects in other clusters. The Hierarchical clustering analysis procedure for finding relationships in various clusters. It works by grouping data objects into a tree of clusters. These methods can be classified agglomerative or divisible, depending on whether the hierarchical decomposition is formed in a bottom-up or top down fashion The main aim of paper is to connection among different clusters using the single link and completed link techniques using Euclidian distance and Manhattan distance, words methods and centroid methods. The clustering that is produced is different from those produced by single link, complete link, group average .the ward's method ,the proximity between two clusters is defined as the increase in the squared error that results when two cluster are merged and the centriod method calculate the proximity between two clusters by calculating the distance between the centriod of clusters.**

Keywords*— Euclidian, single link, complete link, group average, centriod, Proximity

I. INTRODUCTION

Hierarchical clustering techniques are second important category of clustering method. There are two basic approaches for generating a Hierarchical clustering

1. Agglomerative
2. Divisive

The Agglomerative start with the points as individual clusters and, at each step, merge the closet pair of clusters and divisive start with one, all-inclusive clusters and, at each step, split a cluster until only singleton clusters of individual points remain.

A hierarchical clustering is often displayed graphically using a tree like diagram called dendrogram, which displays both the clusters-sub clusters relationship and order in which the clusters were merged. A hierarchical clustering can also be graphically represented using nested cluster diagram.

The following nested diagram example of these two types of figures for a set of four two dimensional points. These points were clustered using the single link technique.

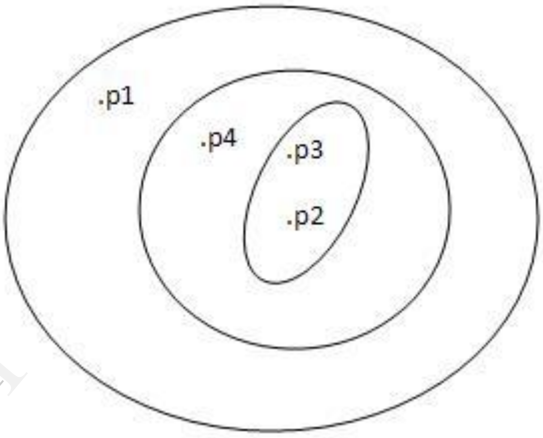


Fig : 1.1 Nested Diagram

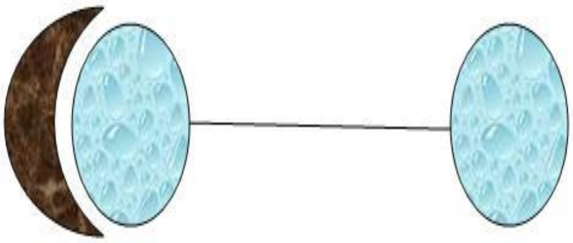
A. Basic Agglomerative Hierarchical clustering Algorithm

Algorithm:

1. Compute the proximity matrix, if necessary
2. Repeat
3. Merge the closest two clusters
4. Update the proximity matrix to reflect the proximity between the new clusters and the original clusters
5. Until only one cluster remains

B. Defining Proximity between clusters

The key operation of Agglomerative Hierarchical clustering Algorithm computing of the proximity between two clusters, and it is the definition of clusters proximity. Cluster proximity is typically defined with particular type of cluster in mind. For example Agglomerative Hierarchical clustering techniques, such as, MIN, MAX and group average, come from a graph based view of clusters. IN defines cluster proximity as the proximity between the closest two points that are in different clusters ,or using graph terms, the shortest edge between two nodes in different subset of nodes. This yields contiguity-based clusters. The contiguity-based cluster was each point is closer to at least one point in its cluster than to any point in another cluster.



Alternatively MAX takes the proximity between the farthest two points in different clusters to be the cluster proximity, are using graph terms, the longest edge between two nodes in different subjects of nodes. Another graph based approach, the group average technique, defines cluster proximity be the average pair wise proximities of all pairs of point s from different clusters.

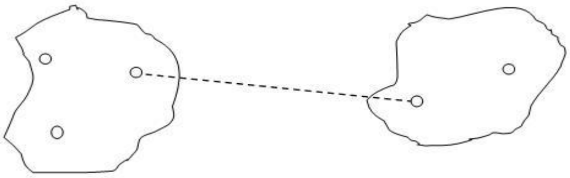


Fig: 1.2.1 MIN (single Link)

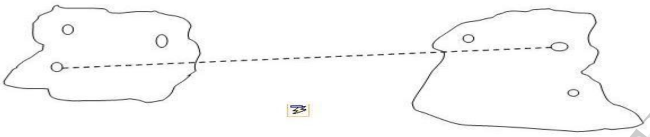


Fig: 1.2.2 MAX (complete Link)

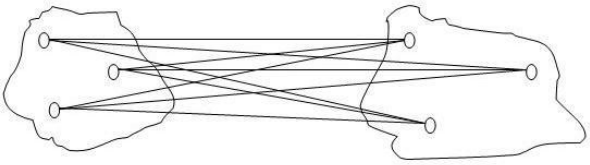


Fig: 1.2.3 Group average

II. Literature Survey

1.” **Agglomerative Mean-Shift Clustering** by Xiao-Tong Yuan, Bao-Gang Hu, “
In this paper, for the purpose of algorithmic speedup, we develop an agglomerative MS clustering method along with its performance analysis. Our method, namely Agglo-MS, is built upon an iterative query set compression mechanism which is motivated by the quadratic bounding optimization nature of MS algorithm. The whole framework can be efficiently implemented in linear running time complexity. We then extend Agglo-MS into an incremental version which performs comparably to its batch counterpart. The efficiency and accuracy of Agglo-MS are demonstrated by extensive comparing experiments on synthetic and real data sets.

2. “**A Variation Bayesian Framework for Clustering with Multiple Graphs** by Motoki Shiga and Hiroshi Mamitsuka”

Mining patterns in graphs has become an important issue in real applications, such as bioinformatics and web mining. We address a graph clustering problem where a cluster is a set of densely connected nodes, under a practical setting that 1) the input ismultiple graphs which share a set of nodes but have different edges and 2) a true cluster cannot be found in all given graphs. For this problem, we propose a probabilistic generative model and a robust learning scheme based on variation Bayesian estimation. A key feature of our probabilistic framework is that not only nodes but also given graphs can be clustered at the same time, allowing our model to capture clusters found in only part of all given graphs. We empirically evaluated the effectiveness of the proposed frame work on not only a variety of synthetic graphs but also real gene networks, demonstrating that our proposed approach can improve the clustering performance of competing methods in both synthetic and real data.

3. “**A Link-Based Cluster Ensemble Approach for Categorical Data Clustering** by Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price”

Although attempts have been made to solve the problem of clustering categorical data via cluster ensembles, with the results being competitive to conventional algorithms, it is observed that these techniques unfortunately generate a final data partition based on incomplete information. The underlying ensemble-information matrix presents only cluster-data point relations, with many entries being left unknown. The paper presents an analysis that suggests this problem degrades the quality of the clustering result, and it presents a new link-based approach, which improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble. In particular, an efficient link-based algorithm is proposed for the underlying similarity assessment. Afterward, to obtain the final clustering result, a graph partitioning technique is applied to a weighted bipartite graph that is formulated from the refined matrix. Experimental results on multiple real data sets suggest that the proposed link-based method almost always Out performs both conventional clustering algorithms for categorical data and well-known cluster ensemble techniques.

III. Methodology

We fallow the methodology for connection among the different clusters

1. Single Link(MIN)
2. Complete Link or CLIQUE(MAX)
3. Group Average Method
4. Ward’s and Centroids Methods

By using the sample data that consist of 6 two-dimensional points

Point	X-Coordinate	Y-Coordinate
p1	0.4	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.3

Fig: 3.1 x y coordinates of six points

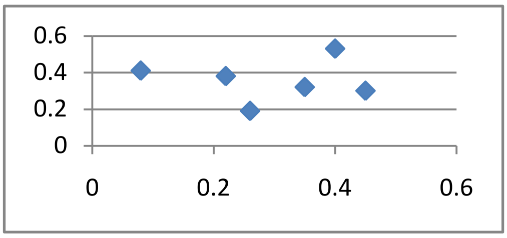


Fig: 3.2 set of 6 two dimensional points

	p1	p2	p3	p4	p5	p6
p1	0	0.24	0.22	0.37	0.34	0.23
p2	0.24	0	0.15	0.2	0.14	0.25
p3	0.22	0.15	0	0.15	0.28	0.11
p4	0.37	0.2	0.15	0	0.29	0.22
p5	0.34	0.14	0.28	0.29	0	0.39
p6	0.23	0.25	0.11	0.22	0.39	0

Fig: 3.3 Euclidian Distance Matrix for 6 points

A. *Single Link (MIN)*

The proximity of two clusters is defined as the Minimum of the distance between any two points in the two different clusters .using graph terminology if u start with all points as singleton clusters and add links between points one at time, shortest link first, then these single link combine the points into clusters .the single link techniques is good at handling non elliptical shapes, but is sensitive noise and outliers

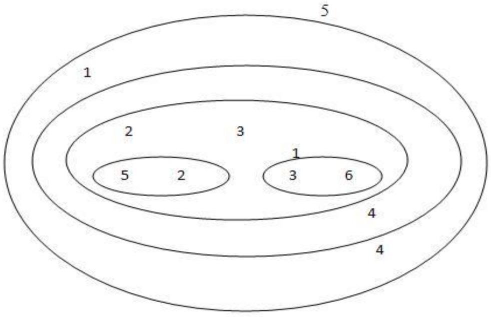


Fig: 3.1.1 Single Link Clustering

B. *Complete Link or CLIQUE (MAX)*

The proximity of two clusters is defined as the maximum of distance between any two points in the two different clusters. using graph terminology, if you start with all points as singleton clusters and add links between points one at a time, shortest link first, then a group of points is not a cluster until all the points in it or completely linked, i.e. such that from a CLIQUE .complete link is susceptible to noise e and outliers, but it can break large clusters and it favors globular shapes.

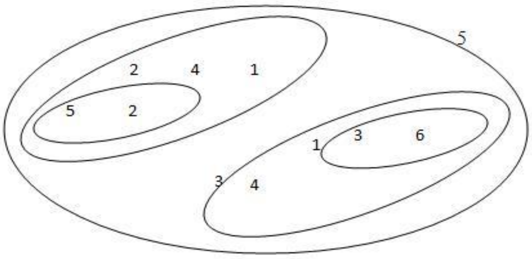


Fig: 3.2.1Complete Link Clustering

C. *Group Average Method*

The proximity of two clusters is defined as the average pair wise proximity among all pairs of points in the different clusters. This is an intermediate approach between the single and complete link approaches. Thus, for group average, the cluster proximity (C_i, C_j) of clusters Ci and Cj, which are of size m_i and m_j respectively, is, expressed by the following equation

$$\text{Proximity } (C_i, C_j)$$
$$=\left(\sum_{\substack{xe\ Ci\\ ye\ Cj}} proximity(x,y)\right)/m_i*m_j$$

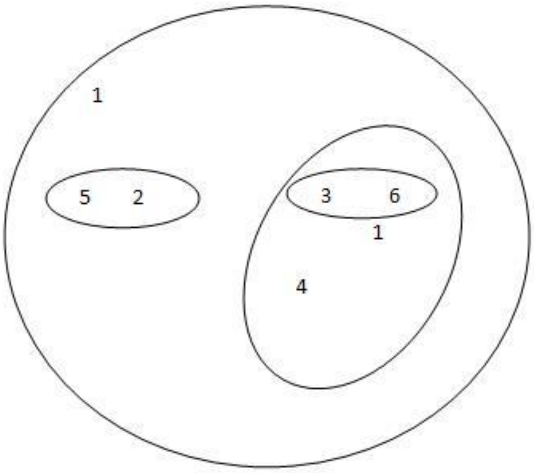


Fig: 3.3.1 Group Average

D. *Ward's and Centroids Methods*

The proximity between two clusters is defined as the increase in the squared error that results when two clusters are merged. Thus, this method uses the same objective function as K-means clustering. While it may see that this feature makes ward's method somewhat distinct from other hierarchical techniques, it can be shown mathematically that ward's method is very similar to the group average method when the proximity between two points is taken to be the square of the distance between them.

Centroid methods calculate the proximity between two clusters by calculating the distance between the centroids of clusters. This technique may seem similar to K-means, but as we have remarked, ward's method is the correct hierarchical analog. These methods also have a characteristics often consider bad that is not processed by the other hierarchical clustering techniques that we have discussed: the possibility of inversions. Specifically two clusters that are merged may be more similar then the pair of clusters that are merged in a previous step. For the other methods, the distance between merged clusters monotonically increases as we proceed from singleton clusters two one all-inclusive cluster

IV. RESULT ANALYSIS

We analysis the results based the on the sample data represented in graphical form

A. *Single Link:*

The distance between two clusters (3, 6) and (2, 5) by using Euclidian Distance

Dist({3,6},{2,5})=min(dist(3,2),dist(6,2),dist(3,5),dist(6,5))
=min (0.15, 0.25, 0.28, 0.39)=0.15

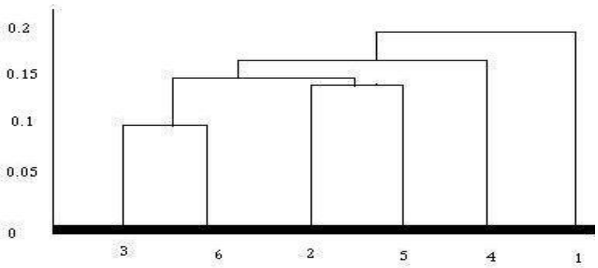


Fig: 4.4.1 Graphs for Single Link

B. *Complete Link*

As with a single link, points 3 and 6 are merged first. However, {3, 6} is merged with {4}, instead of {2, 5} or {1} because

dist({3,6},{}4) =max(dist(3,4),dist(6,4))

=max (0.15, 0.22)

= 0.22

dist ({3, 6}, {2, 5})
=max(dist(3,2),dist(6,2),dist(3,5),dist(6,5))

=max (0.15, 0.25, 0.28, 0.39)

=0.39

dist({3,6},{1}) = max(dist(3,1),dist(6,1))

=max (0.22, 0.23)
= 0.23

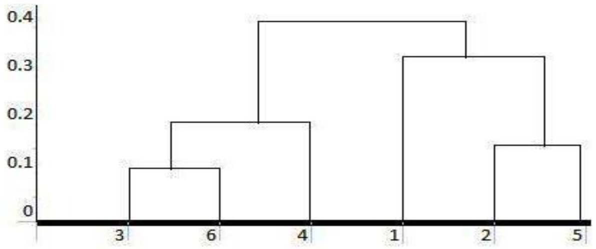


Fig: 4.2.1 Graphs for Complete Link

C. *Group Average*

We calculate the distance between some clusters
dist({3,6,4},{1}) = (0.22+0.37+0.23)/(3*1)
=0.28
dist({2,5},{1})=(0.2357+0.3421)/(2*1)=0.2889
dist({3,6,4},{2,5})
=(0.15+0.28+0.25+0.39+0.20+0.29)/(6*2)
=0.26

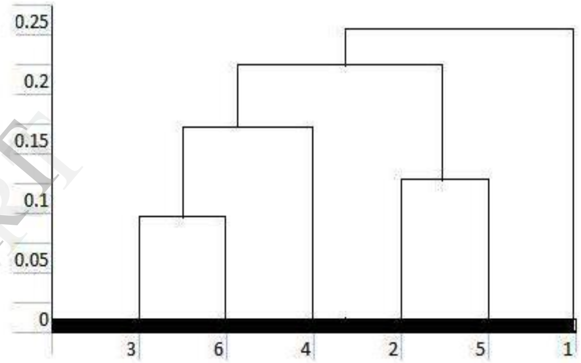


Fig: 4.3.1 Graphs for Group Average

D. *Wards Method*

The clustering that is produced in different from those produced by single link, complete link and group average.

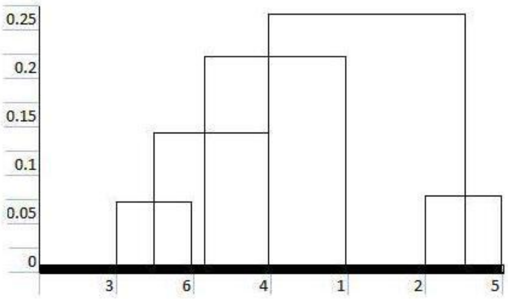


Fig: 4.4.1 Graph for Wards Method

V. CONCLUSION

The Hierarchical clustering analysis procedure for finding relationships in various clusters. It works by grouping data objects into a tree of clusters. These methods can be classified agglomerative or divisible, depending on whether the hierarchical decomposition is formed in a bottom-up or top down fashion.

REFERENCES

[1] S.E. Schaeffer, "Graph Clustering," Computer Science Rev., vol. 1, pp. 27-64, 2007.

[2] S. Arora, S. Rao, and U. Vazirani, "Geometry, Flows, and Graph-Partitio Algorithms,"Comm. ACM, vol. 51, no. 10, pp. 96-105,2008.

[3] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log,"Proc. ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining, pp. 407-416, 2000.

[4] M. Bilenko, S. Basu, and R. Mooney, "Integrating Constraints and Metric Learning in Semi-Supervised Clustering," Proc. Int'l Conf. Machine Learning, pp. 81-88, 2004.

[5] I. Davidson and S.S. Ravi, "Using Instance-Level Constraints in Agglomerative Hierarchical Clustering: Theoretical and Empirical Results," Data Mining and Knowledge Discovery, vol. 18, no. 2, pp. 257-282, Apr. 2009.

[6] L. Dragomirescu and T. Postelnicu, "A Natural Agglomerative Clusteringmethod forBiology," Biometrical J., vol. 33, no. 7, pp. 841-849, Jan. 2007.

[7] M. Fashing and C. Tomasi, "Mean Shift Is a Bound Optimization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 471-474, Mar. 2005.

[8] D. Freedman and P. Kisilev, "Fast Mean Shift by Compact Density Representation," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2009.

[9] K. Fukunaga and L. Hostetler, "The Estimation of the Gradient of a Density Function, with Application in Pattern Recognition," IEEE Trans. Information Theory, vol. 21, no. 1, pp. 32-40,Jan. 1975.

[10] M.R. Garey and D.S. Johnson, Computers and Intractability, A Guide to the Theory of NP-Completeness. Freeman, 1979.

[11] B. Georgescu, I. Shimshoni, and P. Meer, "Mean Shift Based Clustering in High Dimensions: A Texture Classification Example," Proc. IEEE Int'l Conf. Computer Vision, vol. 1, pp. 456-463, 2003.

[12] S. Guha, R. Rastogi, and K. Shim, "Cure: An Efficient Clustering Algorithm for Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 73-84, 1998.

[13] A.K. Jain, M.N. Murty, and P. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[14] K. Lang, "Newsweeder: Learning to Filter Netnews," Proc. Int'l Conf. Machine Learning, pp. 331-339, 1995.

[15] S. Paris and F. Durand, "A Topological Approach to Hierarchical Segmentation Using Mean Shift," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognitio, 2007.

[16] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Publishers, 1990.

[17] A.K. Jain and R.C. Dubes, Algorithms for Clustering. Prentice-Hall, 1998.

[18] P. Zhang, X. Wang, and P.X. Song, "Clustering Categorical Data Based on Distance Vectors," The J. Am. Statistical Assoc., vol. 101, no. 473, pp. 355-367, 2006.

[19] J. Grambeier and A. Rudolph, "Techniques of Cluster Algorithms in Data Mining," Data Mining and Knowledge Discovery, vol. 6, pp. 303-360, 2002.

[20] K.C. Gowda and E. Diday, "Symbolic Clustering Using a New Dissimilarity Measure," Pattern Recognition, vol. 24, no. 6, pp. 567-578, 1991.

[21] Z. Huang, "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, vol. 2, pp. 283-304, 1998.

[22] Z. He, X. Xu, and S. Deng, "Squeezer: An Efficient Algorithm for Clustering Categorical Data," J. Computer Science and Technology, vol. 17, no. 5, pp. 611-624, 2002.

[23] D.H. Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, vol. 2, pp. 139-172, 1987.

[24] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems," VLDB J., vol. 8, nos. 3-4, pp. 222-236, 2000.

[25] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS: Clustering Categorical Data Using Summaries," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.

[26] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering Aggregation," Proc. Int'l Conf. Data Eng. (ICDE), pp. 341-352, 2005.

[27] N. Nguyen and R. Caruana, "Consensus Clusterings," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 607-612, 2007.

[28] C. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles: Methods and Analysis," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 4, pp. 1-40, 2009.

[29] X.Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," Proc. Int'l Conf. Machine Learning (ICML), pp. 36-43, 2004.

[30] A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

BIOGRAPHIES

D.Veeraiah, Research Scholar in JNTUK, Kakinada.M.Tech in CSE from JNTUH, B.Tech in CSIT from JNTUH. His areas of interest Data Mining, Network Security

Dr D.Vasumathi Ph.D from JNTU Hyderabad. Currently Working as Professor in Department of CSE, JNTUCEH. Her areas of interest Data Mining, Network Security, Ad hoc Networks, Software Engineering