A sample hierarchical cluster victimization agglomative methodology and divisive methodology in data mining

[1]S.Shirisha, [2]Ch.Jaya Bharathi

[1]Assistant Professor, [2]Associate Professor
VITS, Deshmukhi, Hyderabad, India

***Abstract*: During this paper, we have a tendency to be discussing concerning hierarchical cluster victimization agglomative & divisive strategies in data processing for sample objects, scattered in several areas. A hierarchical cluster methodology works by grouping information objects into a tree of clusters. In the hierarchical clustering, for a sample of points (objects), by applying agglomative or divisive methods need to calculate min, max, mean, and average distance. Depends on values, afterward, we can form different clusters.**

***Keywords*: hierarchical clustering, agglomative method, divisive method**

1. INTRODUCTION:

Create a hierarchical clustering using set of knowledge (or objects) victimization. This methodology creates a hierarchical decomposition of the given set of knowledge objects. We can able to classify hierarchical strategies on the premise of however the objects are partitioned, and how cluster is created.

There are mainly 2 approaches in hierarchical clustering, those are:
1. **Agglomative Approach** (Bottom- up): This approach is additionally referred to as the bottom-up approach. In this, we have a tendency to begin with every object forming as a separate cluster. This merging of the objects is a continuous process, forming as a team. Finally forming as one team or cluster or continued until the condition terminates or condition reached. In This bottom-up method, **AGNES** – AGglomative NESting technique is employed here.

2. **Divisive Approach** (Top- down) this approach is additionally referred to as the top-down approach. In this, we have a tendency to begin with all of the objects within the same cluster. Within the continuous iteration, a cluster is getting a divorce into smaller clusters. It's down till every object in one cluster or the termination condition holds. This methodology is rigid, i.e., once a merging or ripping is finished, it will ne'er be undone. These top-down strategies will the reverse of agglomerated hierarchical cluster by beginning with all objects in one cluster. It subdivides the cluster into smaller and smaller items, till every object forms a cluster on its own or till it satisfies bound termination conditions. **DIANA** – Divisive ANalysis technique is employed here.

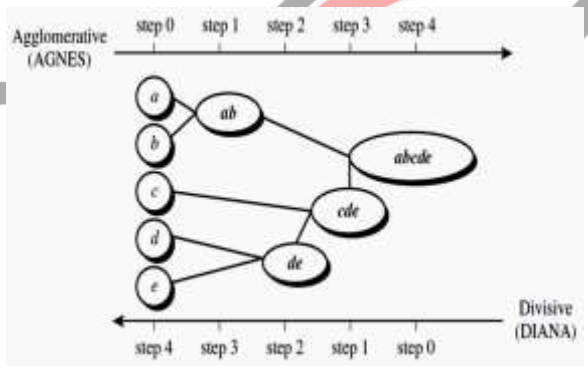

Fig 1: Agglomative &Divisive approach

In either agglomative or divisive hierarchical cluster, the user will specify the specified variety of clusters as a termination condition.

Four widely used measures for distance between clusters are as follows, where $|p-p'|$ is the distance between two objects or points, p and p'; m_i is the mean for cluster, C_i; and n_i is the number of objects in C_i.

$$\text{Minimum distance}: \quad d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'| \qquad (7.20)$$

$$\text{Maximum distance}: \quad d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'| \qquad (7.21)$$

$$\text{Mean distance}: \quad d_{mean}(C_i, C_j) = |m_i - m_j| \qquad (7.22)$$

$$\text{Average distance}: \quad d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'| \qquad (7.23)$$

Fig 2: formulae

A tree structure referred to as a dendrogram is often wont to represent the method of hierarchical cluster. It shows however objects are sorted along step by step Starts with individual points as clusters; in turn merge the two closest clusters till only one cluster remains.

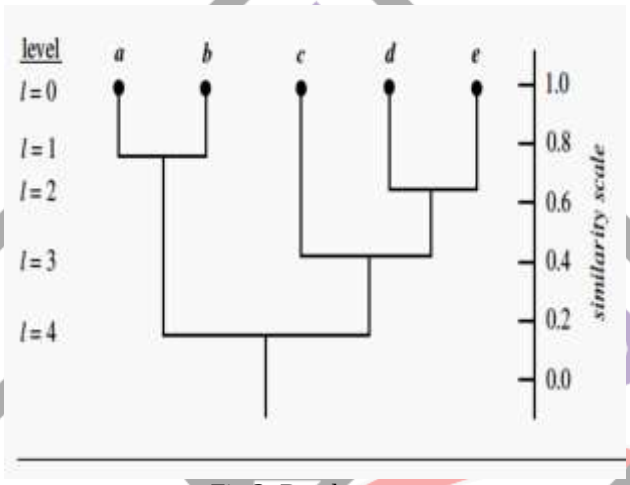

Fig 3: Dendrogram

2. RELATED WORK:

To solve the problem of high dimensionality, accuracy, scalability various researchers put their efforts. Hierarchical clustering method is an alternative technique to partitional clustering. In the proposed method we proceed by updating the hierarchical representation of the data instead of recomputing the whole tree, when new patterns have to be taken into account every time.

Euclidian distance matrix for 6 points:

	p1	p2	p3	p4	p5	p6
P1	**0**					
P2	**0.24**	**0**				
P3	**0.22**	**0.15**	**0**			
P4	**0.37**	**0.2**	**0.15**	**0**		
P5	**0.311**	**0.14**	**0.28**	**0.29**	**0**	
p6	**0.23**	**0.25**	**0.11**	**0.22**	**0.39**	**0**

Fig 4: Euclidian distance between 6 points

Defining proximity between clusters

Single Linkage:

In single linkage hierarchical cluster, the gap between 2 clusters is outlined as the shortest distance between 2 points in every cluster. i.e the left is adequate to the length of the arrow between their 2 nearest points.

Here the distance is calculated as the difference between **two most similar objects**

$$D_{sl}(C_i, C_j) = \min_{x,y} \{ d(x, y) | x \in C_i, y \in C_j \}$$

Single-link clustering: example

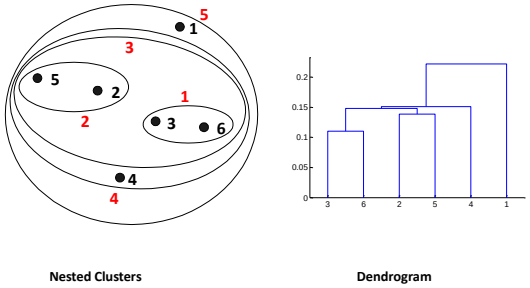

Fig 5: Single link clustering

Result:

- **Distance between {3,6} & {2,5 } is**

dis({3,6} , {2,5 })= min(d(3,2), d(6,2),d(3,5),d(6,5))
=min(0.15,0.25.0.28,0.39) =0.15

Advantages: Can handle non-elliptical shapes

Limitations: 1. Sensitive to noise and outliers 2. It produces long, elongated clusters

Complete Linkage: In complete linkage hierarchical cluster, the gap between 2 clusters is outlined as the longest distance between 2 points in every cluster. i.e the left is adequate to the length of the arrow between their 2 furthest points.

Here the distance is calculated as the difference between **two most dissimilar objects**

$$D_{cl}(C_i, C_j) = \max_{x,y}\{d(x,y) | x \in C_i, y \in C_j\}$$

Complete-link clustering: example

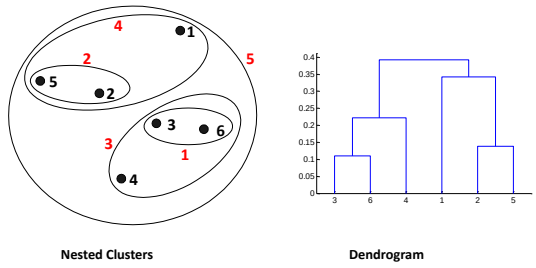

Fig 6: Complete link clustering

Result:

Points 3,6 are merged first, then {3,6} is merged with {4} instead of {2,5} or {1} because

Dist({3,6},{4})=max(d(3,4),d(6,4)) = max(.15,.22)=0.22

Dist({3,6},{2,5})=max(d(3,2),d(6,2),d(3,5),d(6,5))= max(.15,.25,.28,.39)=0.39

Dist({3,6},{1})=max(d(3,1),d(6,1))=max(.22,.23)=0.23

Advantages: 1. More balanced clusters (with equal diameter) 2.Less susceptible to noise

Limitations: 1. tends to break large clusters **2.** All clusters are having the same diameter; small clusters are merged with larger cluster.

Group Average linkage: In average linkage hierarchical cluster, the gap between 2 clusters is outlined because of the average distance between every purpose in one cluster to each purpose within the different cluster. i.e connecting the points of 1 cluster to the opposite.

Formula is:

$$D_{avg}(C_i, C_j) = \frac{1}{|C_i| \times |C_j|} \sum_{x \in C_i, y \in C_j} d(x,y)$$

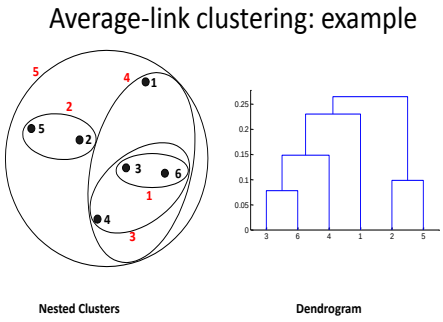

Fig 7: Average link clustering

Result:

Dist({3,6,4},{1})=(.22+.37+.23)/3*1=0.28

Advantages: Less susceptible to noise and outliers

Limitations: Biased towards globular clusters

Hierarchical Clustering: Time and area necessities For a dataset X consisting of n points, O(n2) area it needs storing the gap matrix , O(n3) time in most of the cases. There are n steps and at every step, the scale n2 distance matrix should be updated and searched. Complexness is reduced to O(n2 log(n)) time for a few approaches by victimization acceptable information structures To illustrate the behavior of assorted cluster algorithms, use sample information consists of five two-dimensional points.

Specific Techniques

Proximity of two clusters for

1. Single link or MIN
2. Complete Link or MAX or CLIQUE
3. Cluster average methodology (Group average method)
4. Ward methodology
5. Center of mass methodology or Centroid method

Key problems in hierarchical cluster

1. **Lack of global or worldwide Objective Function:** agglomerative hierarchical cluster techniques perform cluster on an area level and intrinsically there's no 0% Plagiarised 100% unique global objective operate like within the K-Means formula. This can be truly a bonus of this method as a result of the time and area complexness of world functions tends to be terribly overpriced.
2. **Ability to handle completely different cluster Sizes**: we got to decide the way to treat clusters of assorted sizes that are unified along a pair of approaches weighted-all clusters are equal and unweighted.
3. **Merging choices are Final**: one drawback of this method is that when 2 clusters are unified they cannot be getting a divorce at a later time for an additional favorable union.

Strengths of hierarchical cluster

1. No assumptions on the quantity of clusters-Any desired number of clusters are obtained by 'cutting' the dendogram at the right level.

2. Hierarchical clustering plays lead role in classification-Example in biological sciences (e.g., phylogenies reconstruction), web (e.g., product catalogs, marketing) etc
3. Easy to grasp and simple to do

Weaknesses of hierarchical cluster

1. Seldom provides the simplest answer
2. Capricious choices
3. Missing information
4. Data types
5. Mistaking of the dendrogram
6. High in time complexness Matched Source

3. CONCLUSION:

Hierarchical clustering is used in different areas such as search engines, text mining, web mining, information retrieval, machine learning and topological analysis. Here we studied number of document clustering algorithms. It can solve the problem of high dimensionality, accuracy and scalability.

REFERENCES:

1. Faculdade Campo Limpo Paulista (FACCAMP), Paulista" Case studies in divisive hierarchical clustering" IIJICA 2017 ISSN 1751-6498
2.Pranav Nerurkara , Archana Shirkeb, Madhav Chandanec , Sunil Bhirud "Empirical Analysis of Data Clustering Algorithms", Elsevier Volume 125, 2018, Pages 770-779.
3. Nisha and Puneet Jai Kaur, "A Survey of Clustering Techniques and Algorithms", IEEE (978-9-3805-4415-1), 2015

4. K. Kameshwaran and K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies (0975-9646), Vol. 5(2), 2014
5. R. Indhu, R. Porkodi "Comparison of Clustering Algorithm",IJSRCSEIT 2018,Volume 3 , Issue 1 , ISSN : 2456-3307
6. K. Chitra, Dr. D.Maheswari" A Comparative Study of Various Clustering Algorithms in Data Mining" IJCSMC, Vol. 6, Issue. 8, August 2017, pg.109 – 115
7. Oussama Rouane,Hacene Belhadef, Mustapha bouakkaz "Combine clustering and frequent itemsets mining to enhance biomedical text summarization" Volume 135, 2019, Pages 362-373
8. Omid Sadeghian,Arman Oshnoei, Morteza Kheradmandi, Rahmat Khezri, Behnam Mohammadi -Ivatloo "A robust data clustering method for probabilistic load flow in wind integrated radial distribution networks" International Journal of Electrical Power & Energy Systems, Volume 115, February 2020, Article 105392
9. Mahmood Shakir Hammoodi, Frederic Stahl, Atta Badii"Real-time feature selection technique with concept drift detection using adaptive micro-clusters for data stream mining" Knowledge-Based Systems, Volume 161, 2018, Pages 205-239
10. Alessia Amelio, Andrea Tagarelli "Data Mining:Clustering" Encyclopedia of Bioinformatics and Computational Biology, Volume 1, 2019, Pages 437-448
11. Dongzhi Zhang,Kyungmi Lee, Ickjai Lee "Hierarchical trajectory clustering for spatio-temporal periodic pattern mining" Expert Systems with Applications, Volume 92, February 2018, Pages 1-11
12. Kyle Woodward, Johan Wesseloo, Yves Potvin"A spatially focused clustering methodology for mining seismicity" Engineering Geology, Volume 232, 2018, Pages 104-113
13. Sean Carlisto de Alvarenga, Sylvio Barbon, Rodrigo Sanches Miani, Michel Cukier, Bruno Bogaz Zarpelão "Process minging and hierarchical clustering to help intrusion alert visualization" Computers & Security, Volume 73, March 2018, Pages 474-491
14. Kaveh Khalili-Damghani, Farshid Abdi, Shaghayegh Abolmakarem "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries" Applied Soft Computing, Volume 73, 2018, Pages 816-828.
15. An overlapping community detection algorithm based on rough clustering of links Elsevier online 4, DEC, 2019, 101777
16. Constrained distance based clustering for time-series: a comparative and experimental study 1663–1707(2018) springer link, Thomas lampert, thi-bich-hanhdao

Authors Profile:

1. Shirisha S ingireddy was awarded B.Tech from computer science & Engineering in the year 2011. She was awarded M.Tech in computer science & engineering in the year 2017. Her research interests include software Engineering, cloud computing. At present working as a Assistant professor in Computer science & engineering Department, VITS, JNTU, Hyderabad, India.

. 2. Jaya Bharathi Chintalapati, currently working as a Associate Professor in Department of computer science and Engineering, Vignan Institute of Technology & Science, Hyderabad, Telangana, India. Pursuing Ph.D from Vignan University,Guntur, AP, India. She received her Master's degree in computer science & Engineering from JNTU Kakinada in the year 2011. She also awarded Master of Computer Applications degree from JNTU in the year 2004. Her research areas include Data mining, Meachine learning Network security and Cloud Computing.