Volume 3, Issue 5, May 2018 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Comparative Study SLINK and CLINK Agglomerative Clustering

Amarkant Goud
M. Tech (C.S.E.) 4th Semester
Lord Krishna College of Technology Indore e
Indore M.P. India
kant.goud@gmail.com

Vijay Kumar Verma
Asst Prof. Department of C.S.E.
Lord Krishna College of Technology Indore
Indore M.P. India
vijayvermaonlione@gmail.com

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 A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Hierarchical clustering method works by grouping data objects into a tree of clusters. Agglomerative hierarchical is a "bottom up" approach each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. In this paper we proposed a comparative study between SLINK and CLINK to optimize clustering.

Keywords: Cluster, Divisive Hierarchical, Agglomerative, , SLINK , CLINK

I. INTRODUCTION

Some common definitions are collected from the clustering literature and given below
"A Cluster is a set of entities which are alike, and entities from different clusters are not alike."
"A cluster is an aggregation of points in the space such that the distance between two points in the cluster is less than the distance between any point in the cluster and any point not in it."
"Clusters may be described as connected regions of a multidimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points."
Although the cluster is an application dependent concept, all clusters are compared with respect to certain properties: density, variance, dimension, shape, and separation. The cluster should be a tight and compact high-density region of data points when compared to the other areas of space. From compactness and tightness, it follows that the degree of dispersion (variance) of the cluster is small. The shape of the cluster is not known a priori. It is determined by the used algorithm and clustering criteria. Separation defines the degree of possible cluster overlap and the distance to each other.

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

II .CLUSTER ANALYSIS

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

III. REQUIRMENT OF CLUATERING

The following are typical requirements of clustering in data mining

Scalability: Clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.
Ability to deal with different types of attributes: Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering
other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.
Ability to deal with noisy data: Most real-world databases contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
Incremental clustering: Some clustering algorithms cannot incorporate newly inserted data into existing clustering structures and, instead, must determine a new clustering from scratch. Some clustering algorithms are sensitive to the order of input data.

High dimensionality: A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling two to three dimensions. Finding clusters of data objects in high dimensional space is challenging, especially considering that such data can be sparse and highly skewed.

IV. LITERATURE REVIEW

In 2010 Revati Raman et al proposed "Fuzzy Clustering Technique for Numerical and Categorical dataset". They presented a modified description of cluster center to overcome the numeric data only limitation of Fuzzy c-mean algorithm and provide a better characterization of clusters. The fuzzy k-modes algorithm for clustering categorical data. They proposed a new cost function and distance measure based on co-occurrence of values. [5]
In 2011 K. Ranjini proposed "Performance Analysis of Hierarchical Clustering Algorithm" They explain the implementation of agglomerative and divisive clustering algorithms by using various types of data. They implements and analysis running time of the algorithms using different linkages (agglomerative) to different types of data are taken for analysis[6].
In 2012 Dan Wei, Qingshan Jiang et al. proposed "A novel hierarchical clustering algorithm for gene Sequences" .The proposed method is evaluated by clustering functionally related gene sequences and by phylogenetic analysis. They presented a novel approach for DNA sequence clustering based on a new sequence similarity measure, DMK, which is extracted from DNA sequences based on the position and composition of oligonucleotide pattern. [7].
In 2013 K. Sasirekha, P. Baby proposed "Agglomerative Hierarchical Clustering Algorithm- A Review". They showed that data mining hierarchical clustering method are used to build a hierarchy of clusters. They also show that hierarchical clustering generally fall into two types: Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy [8].
In 2014 Archana Singh and Avantika Yadav proposed "Hybrid Approach of Hierarchical Clustering". They proposed a hybrid approach of clustering based on AGNES and DIANA clustering algorithms, an extension to the standard hierarchical clustering algorithm. In the proposed algorithm, they used single linkage as a similarity measure. The proposed clustering algorithm provides more consistent clustered results from various sets of cluster centroids with tremendous efficiency [9].
In 2015 Olga Tanaseichuk "An Efficient Hierarchical Clustering Algorithm for Large Datasets". They show that Hierarchical clustering is a widely adopted unsupervised learning algorithm. Standard implementations of the exact algorithm for hierarchical clustering require $O(n)^2$ time and $O(n)^2$ memory and thus are unsuitable for processing datasets with large object. They present a hybrid hierarchical clustering algorithm requiring less time and memory [10].
In 2016 Amit Kumar Kar et al proposed "Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining". They analyzes the four major clustering algorithms namely: Partitioning methods, Hierarchical methods, Grid-based methods and Density-based methods and comparing the performance of these algorithms on the basis of correctly class wise cluster building ability of algorithm[11].
In 2017 Shubhangi Pandit et al "An Improved Hierarchical Clustering Using Fuzzy C-Means Clustering Technique for Document Content Analysis". They present work a clustering technique and proposed using fuzzy c-means clustering algorithm for recognizing the text pattern from the huge data base. The proposed work is also committed to advance the approach of clustering for computing the hierarchical relationship among different data objects [12].

V. SLINK AND CLINK

Consider a simple data set with six object and coordinate value. Each object is represented in two dimensional plan using x and y coordinates

Object	X	Y
A	4	4
B	8	4
C	15	8
D	24	4
E	24	12

Table 1 Simple dataset

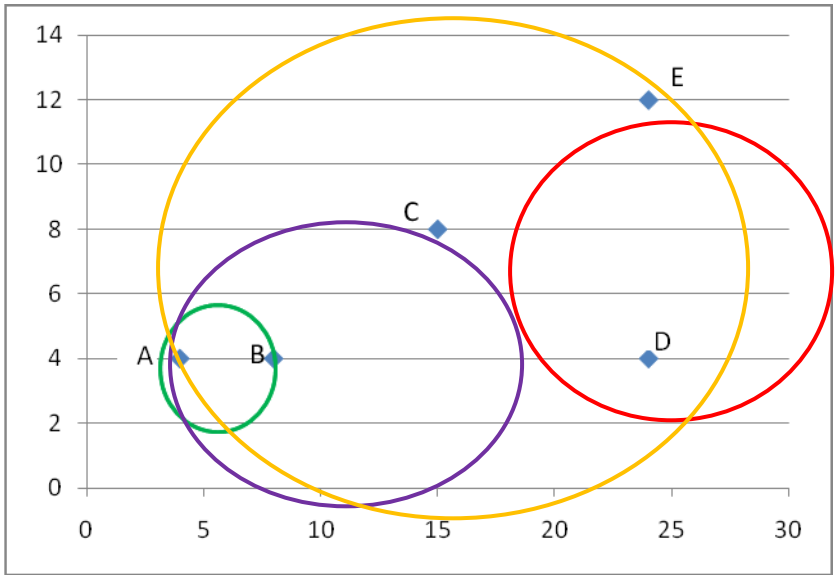

Figure.1: working of SLINK

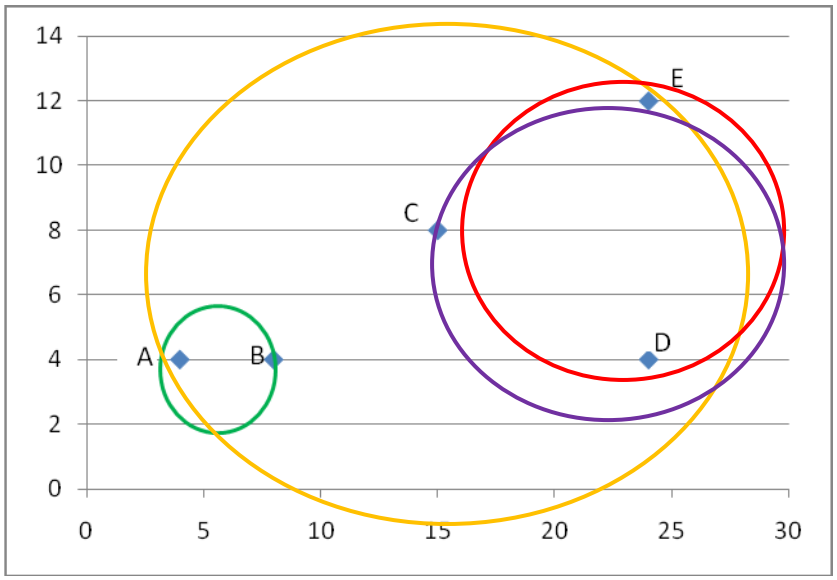

Figure 2: working of CLINK

VI. COMPLEXITY OF AGGLOMERATIVE CLUSTERING

1. Min linkage

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

2. Max linkage

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

3. Average linkage

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

Each of these linkage algorithms can yield totally different results when used on the same dataset, as each has its specific properties. So it is very difficult to decide which method is to best for select data set. 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

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 SLINK and CLINK . Each of these algorithms can yield totally different results when used on the same dataset, as each has its specific properties. The CLINK clustering methods usually produce more compact clusters and more useful hierarchies than the SLINK clustering methods, yet the SLINK 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.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.

REFERENCE

1. J. Han, M. Kamber, Data mining, Concepts and techniques, Academic Press, 2003.
2. Arun K. Pujari, Data mining Techniques, University Press (India) Private Limited, 2006.
3. D. Hand, H. Mannila, P. Smyth, "Principles of Data Mining, Prentice Hall of India, 2004
4. Nachiketa Sahoo "Incremental Hierarchical Clustering of Text Documents" May 5, 2006
5. 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.
6. K. Ranjini Performance Analysis of Hierarchical Clustering Algorithm Performance Analysis of Hierarchical Clustering Algorithm" Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011 (2011).
7. Dan Wei, Qingshan Jiang et al. proposed "A novel hierarchical clustering algorithm for gene Sequences" 2012 Wei et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License
8. K.Sasirekha, P.Baby Agglomerative Hierarchical Clustering Algorithm- A Review "International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013 1 ISSN 2250-3153.
9. Archana Singh and Avantika Yadav "Hybrid Approach of Hierarchical Clustering"World Applied Sciences Journal 32 (7): 1181-1191, 2014 ISSN 1818-4952 © IDOSI Publications, 2014
10. Olga Tanaseichuk, Alireza Hadj "An Efficient Hierarchical Clustering Algorithm for Large Datasets" Austin J Proteomics Bioinform & Genomics - Volume 2 Issue 1 - 2015 ISSN : 2471-0423
11. Amit Kumar Kar "A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining". ACEIT Conference Proceeding 2016
12. Shubhangi Pandit et al proposed " An Improved Hierarchical Clustering Using Fuzzy C-Means Clustering Technique for Document Content Analysis" Volume 7, Issue 4, April 2017 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com.`