Volume 2, Issue 12, December 2017 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

A More Accurate Approach to Construct Numeric Clusters

Zaid Rehman Qureshi
M. Tech. (CSE) IV Sem
Lord Krishna College of Technology
Indore M.P. India

Vijay Kumar Verma
Asst. Prof. C.S.E. Dept.
Lord Krishna College of Technology
Indore M.P. India

Abstract: ***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. 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. In this paper we proposed a new and efficient approach to construct more accurate clusters.***

Keywords: Cluster, Distance, Similarity, average distance, Splitter

I. INTRODUCTION

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 [1,2]. 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 [1,3].

II. BACKGROUND

Defining the characteristics of a cluster is a difficult task, although different authors emphasize on different characteristics

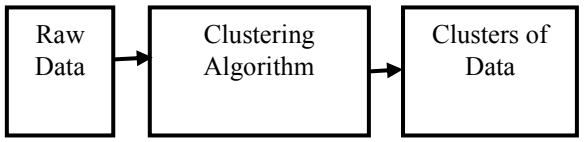

Figure 1. clustering process

Boundaries of a cluster are not exact. Clusters vary in size, depth and breadth. Some clusters consist of small and some of medium and some of large in size. The depth refers to the range related by vertically relationships. Furthermore, a cluster is characterized by its breadth as well. The breath is defined by the range related by horizontally relationships [3,4].

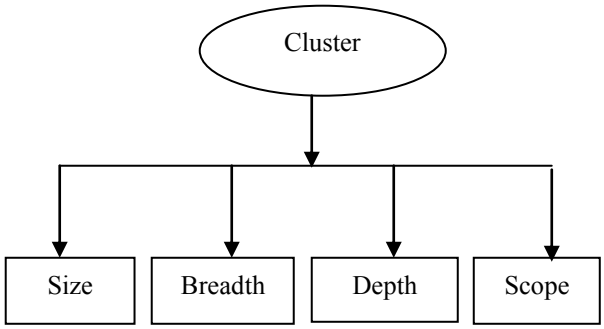

Figure 2. Attribute of a cluster

III. 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 [1,3,4]

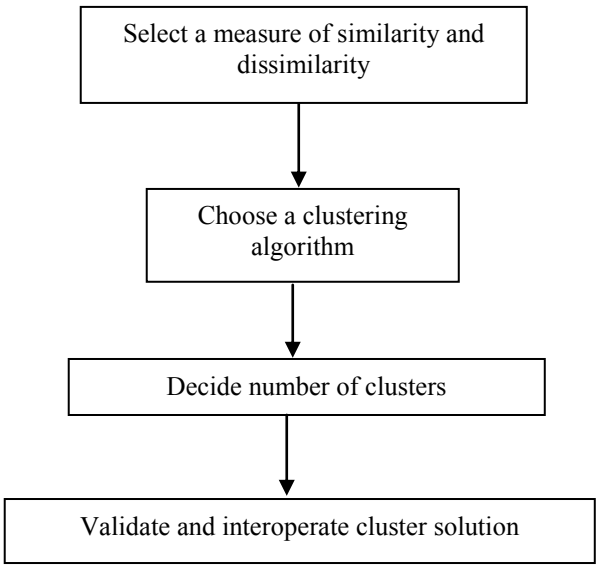

Figure 3. Analysis of clustering

IV. LITERATURE REVIEW

In 2012 Neepa Shah et al Document Clustering: A Detailed Review" .They gave an overview of various document clustering methods, starting from basic traditional methods to fuzzy based, genetic, co-clustering, heuristic oriented etc. They also include the document clustering procedure with feature selection process, applications, challenges in document clustering, similarity measures and evaluation of document clustering algorithm is explained[5].

In 2013 M. Emre Celebi et al. proposed "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm" .They proposed an overview of K-Means method with computational efficiency. They compared eight commonly used linear time initialization method for a large and diverse collection of real and synthetic data sets[6].

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 [7].

In 2015 Y. S. Thakare et al. proposed "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics". They check the performance of basic k means algorithm using various distance metrics for real life dataset. By the experimental analysis and result. They showed that the performance of k means algorithm is varying on the distance metrics for selected database. The proposed work helps to select suitable distance metric for particular application[8].

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[9].

In 2016 Jitendra Pal et al "A Survey on K-Means Clustering Algorithms for Large Datasets" They used two methods global k-means (GKM) and the fast global k-means (FGKM) algorithms for analysis. They iteratively append one cluster center at a time. By using numerical experiments they show the performance of these methods. They implement both algorithms and compare their time and memory for clustering creation[10].

In 2017 Shaoning Li et al proposed "A Novel Divisive Hierarchical Clustering Algorithm for Geospatial Analysis". They proposed a new method, cell-dividing hierarchical clustering (CDHC), based on convex hull retraction. They used following steps a convex hull structure is constructed to describe the global spatial context of geospatial objects. Then, the retracting structure of each borderline is established in sequence by setting the initial parameter. The objects are split into two clusters with the borderlines. Finally, clusters are repeatedly split and the initial parameter is updated until the terminate condition is satisfied [11].

V. PROBLEM STATEMENTS

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

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.
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.
3. 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.
4. 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 what specific purpose is not properly mentioned?

VI. PROPOSED APPROACH

Input: 2-dimensional dataset.
Output: Clusters in 2-dimensional space.
Assumptions: Single linkage is used as a similarity measure and initially all objects are in single cluster.

Step 1: Find the object, which has the highest average dissimilarity to all other objects. This object initiates a new cluster- a sort of a splinter group.
Step 2: For each object i outside the splinter group compute
Step 3: Distance i = [average distance (i,j) j Ssplinter group] - [average distance(i,j) j Ssplinter group]
Step 4: Find an object x for which the difference Distance x is the largest. If Distance x is positive, then x is, on the average close to the splinter group.
Step 5: Repeat Steps 2 and 3 until all differences Distance x are negative. The data set is then split into two clusters.
Step 6: Select the cluster with the largest diameter. The diameter of a cluster is the largest dissimilarity between any two of its objects. Then divide this cluster, following steps 1-4.\ Consider simple data set with six objects

Table 1 simple object with coordinates

Objects	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Table 2 average distance between objects

Distance	Average dissimilarity with objects
A	(0.71+5.66+3.61+4.24+3.20)/5=3.484
B	(0.71+4.95+2.92+3.54+2.50)/5=2.924
C	(5.66+4.95+2.24+1.41+2.50)/5=3.352
D	(3.61+2.92+2.24+1.00+0.50)/5=2.054
E	(4.24+3.54+1.41+1.00+1.12)/5=2.262
F	(3.20+2.50+2.50+0.50+1.12)/5=1.964

Object A has maximum dissimilarity. Object A is chosen to initiate the so-called splinter-group.
Remaining group {A}, {B, C, D, E, F}.We repeat this process for all objects finally we got Closest groups are {{D, F},E}, {A,B}

VII. EXPERIMENTAL ANALYSIS

We evaluate the performance of proposed algorithm and compare it with MIN linkage, MAX linkage methods

Number of objects and accuracy

Number of Objects	Min	Proposed approach
50	0.335674	0.39396
100	0.155668	0.178981
150	0.232241	0.294759

Table 3 comparison based on accuracy

The experiments were performed on Intel Core i5-4200U processor 2GB main memory and RAM: 4GB Inbuilt HDD: 500GB OS: Windows 8. The algorithms are implemented in using C# Dot Framework Net language version 4.0.1. Synthetic datasets are used to evaluate the performance of the algorithms. We have taken 50 objects in two dimensional plan.

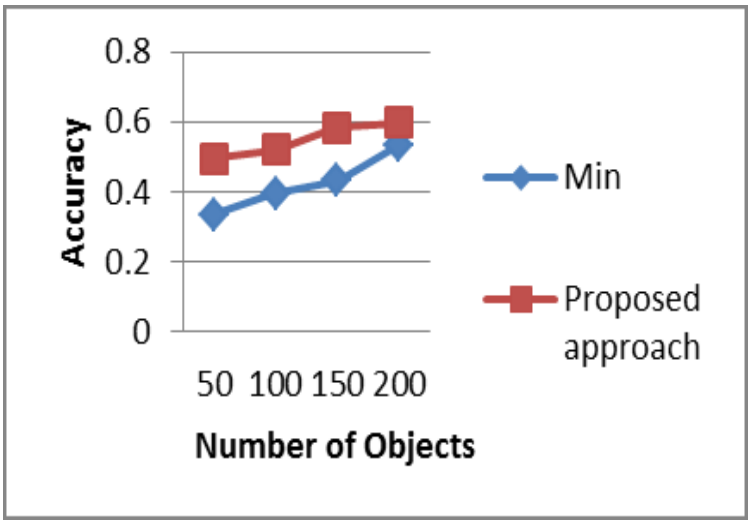

Figure 4. Comparison graph

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. Neepa Shah et al proposed Document Clustering: A Detailed Review" International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5, October 2012.
6. M. Emre Celebi et al. "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm" Expert Systems with Applications, 40(1): 200–210, 2013
7. 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.
8. Y. S. Thakare et al. "Performance Evaluation of K-means Clustering Algorithm with Various Distance MetricsY" International Journal of Computer Applications (0975 – 8887) Volume 110 – No. 11, January 2015.
9. Amit Kumar Kar "A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining". ACEIT Conference Proceeding 2016.
10. Jitendra Pal Singh Parmar " A Survey on K-Means Clustering Algorithms for Large Datasets" IJARCCE ISSN (Online) 2278-1021 ISSN (Print) 2319 5940 International .
11. Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 9, September 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