International Journal of Scientific Engineering and Science
Volume 1, Issue 10, pp. 11-14, 2017. ISSN (Online): 2456-7361

Identify Best Similarity Matrix to Find Accurate Cluster Using Dendrogram Distance

Deepika Patidar[1], Vijay Kumar Verma[2]
[1]M. Tech. (CSE) IV Semester, Lord Krishna College of Technology, Indore M.P. India
[2]Asst. Prof. CSE Dept. Lord Krishna College of Technology, Indore M.P. India

***Abstract*—** *Data Mining has several techniques clustering is one of them. Clustering techniques are used in several real life applications like artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing. There are several algorithms and methods have been developed to improve clustering process. There are several issue like cluster size, depth, number of cluster, Breadth, and relation between object are consider for a clustering method. There are several new methods and techniques have been proposed by various researchers to improve the clustering process in term of accuracy and scalability. In this paper we proposed new concepts based on dendrogram distance to identify the correct distance matrix to find more accurate cluster. The Proposed approach provides which clustering techniques are suitable for a particular data.*

***Keywords*—** *Cluster, partition, hierarchical accuracy, efficiency, agglomerative.*

I. INTRODUCTION

Clustering is one of the most important unsupervised learning techniques it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". 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.

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.

II. PROBLEMS CLUSTERING ALGORITHM

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering. There are some problems which effect the clustering algorithms are

1. Scalability.
2. Dealing with different types of attributes.
3. Discovering clusters with arbitrary shape.
4. Minimal requirements for domain knowledge.
5. Ability to deal with noise and outliers.
6. Insensitivity to order of input records.
7. High dimensionality;
8. Interpretability and usability.

III. CLUSTER PROCESS

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.

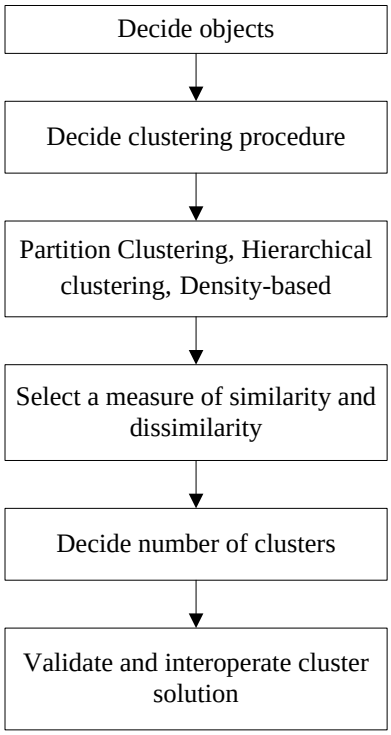

Fig. 1. Clustering process.

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

IV. Literature Review

In 2010 Revati Raman Dewangan, Lokesh Kumar Sharma, Ajaya Kumar Akasapu 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 Explains agglomerative and divisive clustering algorithms and applied on various types of data. The details of the victims of Tsunami in Thailand during the year 2004, was taken as the test data. Visual programming is used for implementation and running time of the algorithms using different linkages (agglomerative) to different types of data are taken for analysis [6].

In 2011 Hussain Abu-Dalbouh et al. proposed "Bidirectional Agglomerative Hierarchical Clustering using AVL Tree Algorithm". Proposed Bidirectional agglomerative hierarchical clustering to create a hierarchy bottom-up, by iteratively merging the closest pair of data-items into one cluster. The result is a rooted AVL tree. The *n* leafs correspond to input data-items (singleton clusters) needs to *n/2* or *n/2+1* steps to merge into one cluster, correspond to groupings of items in coarser granularities climbing towards the root. As observed from the time complexity and number of steps need to cluster all data points into one cluster perspective, the performance of the bidirectional agglomerative algorithm [7].

In 2012 Dan Wei, Qingshan Jiang et al. proposed "A novel hierarchical clustering algorithm for gene Sequences". They proposed method is evaluated by clustering functionally related gene sequences and by phylogenetic analysis. They presented a novel approach for DNA sequence clustering, mBKM, based on a new sequence similarity measure, DMk, which is extracted from DNA sequences based on the position and composition of oligonucleotide pattern. Proposed method may be extended for protein sequence analysis and meta genomics of identifying source organisms of meta genomic data [8].

In 2013 K. Sasirekha, P. Baby proposed "Agglomerative Hierarchical Clustering Algorithm- A Review". They proposed comparison between two algorithms. Comparing between the results of algorithms using normalized data or non-normalizes data that give different results. Normalizationl affect the performance of the algorithm and quality of the results [9].

In 2013 Elio Masciari et al. proposed "A New, Fast and Accurate Algorithm for Hierarchical Clustering on Euclidean Distances" A simple hierarchical clustering algorithm called CLUBS (for Clustering Using Binary Splitting) is proposed in this paper. CLUBS is faster and more accurate than existing algorithms, including k-means and its recently proposed refinements. The algorithm consists of a divisive phase and an agglomerative phase; during these two phases, the samples are repartitioned using a least quadratic distance criterion possessing unique analytical properties [10].

In 2014 J Anuradha, B K Tripathy proposed "Attribute Dependency for Attention Deficit Hyperactive Disorder". They proposed a hierarchical clustering algorithm to partition the dataset based on attribute dependency (HCAD). HCAD forms clusters of data based on the high dependent attributes and their equivalence relation. Proposed approach is capable of handling large volumes of data with reasonably faster clustering than most of the existing algorithms. It can work on both labeled and unlabelled data sets. Experimental results reveal that this algorithm has higher accuracy in comparison to other algorithms [11].

In 2015 Z. Abdullah, A. R. Hamdan "Hierarchical Clustering Algorithms in Data Mining" The proposed method builds the solution by initially assigning each points to its own cluster and then repeatedly selecting and merging pairs of clusters, to obtain a single all inclusive cluster. The key parameter in agglomerative algorithms is the method used to determine the pair of clusters to be merged at each step. Experimental results obtained on synthetic and real datasets demonstrate the effectiveness of the proposed various width cluster method [12].

V. Problem Statement

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

Distance measure: For numerical attributes, distance measures can be used. But identification of measure for categorical attributes in strength association is difficult.

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.

Types of attributes: 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.

VI. Proposed Approach

The proposed approach is used on generation of ensembles based cluster on the basis of few operations like mapping & combination. These operations can be performed with the help of two operators' similarity association & probability for correct classification or classifier analysis of cluster. In this proposed approach our main aim is to identify the cluster partitioned data for hierarchical clustering. It may be represented via parametric representation of nested clustering & dendogram.

International Journal of Scientific Engineering and Science
Volume 1, Issue 10, pp. 11-14, 2017. ISSN (Online): 2456-7361

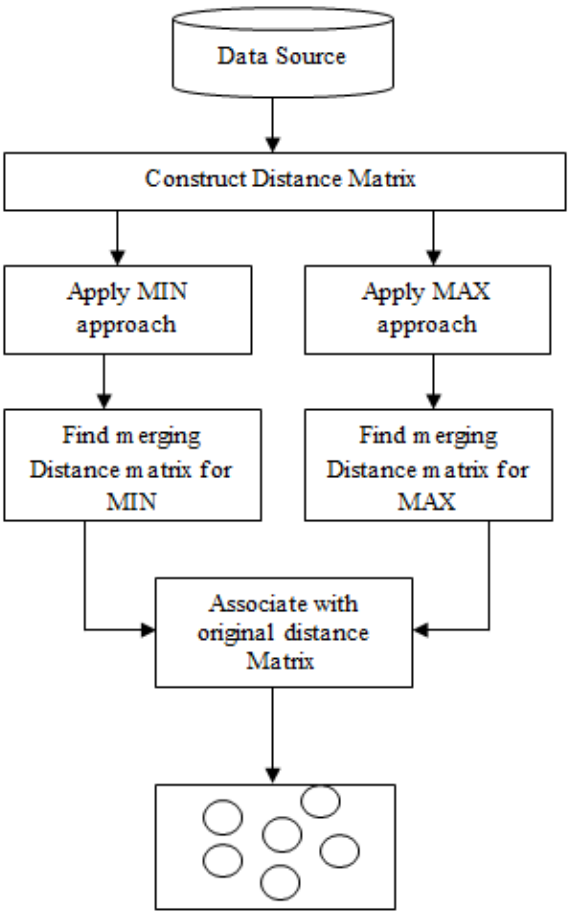

Fig. 2. Working process of proposed approach.

VII. Proposed Algorithms

The proposed approach is used on generation of ensembles based cluster on the basis of few operations like mapping & combination. These operations can be performed with the help of two operators' similarity association & probability for correct classification or classifier analysis of cluster. In this proposed approach our main aim is to identify the cluster partitioned data for hierarchical clustering. It may be represented via parametric representation of nested clustering.

1. Consider each object as separate cluster.
2. Find the distance matrix D, by using any distance formula.
3. Find the nearest pair of clusters in the current clustering. Merge clusters to form into a single cluster. Store merged objects with its corresponding distance in tress distance Matrix.
4. Update distance matrix, D, by deleting the rows and columns corresponding to clusters. Adding a new row and column corresponding to the merged cluster
5. If all objects are in one cluster, stop. Otherwise, go to step 3.
6. Find association relation coefficient value with single, complete and average linkage methods

VIII. Experimental Analysis

We evaluate the performance of proposed algorithm and compare it with single linkage, complete linkage and average linkage methods. The experiments were performed on Intel Core i5-4200U processor 2GB main memory and RAM: 4GB In built 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 plans. Maximum value for X coordinated, 100 and Maximum value for Y coordinated is also 100. User can give the coordinated value for any object between 0 to 100 for pair of X and Y. SQL Server R2 (2008) to store our database. Database contain three attribute first is name or number of the object, second X coordinated value and third is Y coordinated value.

Table I show number of objects and accuracy for single linkage and complete linkage

TABLE I. Accuracy with different objects.

Number of Objects	MIN Linkage	MAX Linkage
50	0.425	0.733
100	0.495	0.638
150	0.382	0.614

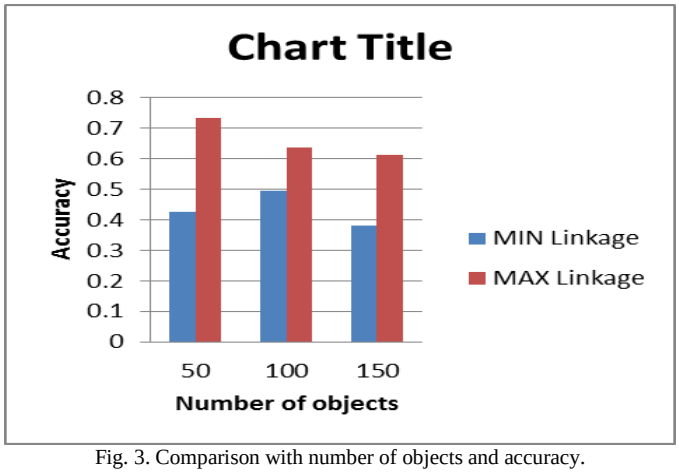

Fig. 3. Comparison with number of objects and accuracy.

IX. Conclusion and Future Works

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 Single linkage ,Complete linkage , Average linkage and Centroid.

Each of these linkage algorithms can yield totally different results when used on the same dataset, as each has its specific properties. 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. 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 assembling clusters like neural network, fuzzy logic, genetic algorithms etc. to enhance the clustering

REFERENCE

[1] J. Han and M. Kamber, *Data mining, Concepts and Techniques*, Academic Press, 2003.
[2] A. K. Pujari, *Data Mining Techniques*, University Press (India) Private Limited, 2006.
[3] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, Prentice Hall of India, 2004.
[4] N. Sahoo, "Incremental hierarchical clustering of text documents," May 5, 2006
[5] R. R. Dewangan, L. K. Sharma, and A. K. Akasapu, "Fuzzy Clustering Technique for Numerical and Categorical dataset," *International Journal on Computer Science and Engineering (IJCSE), NCICT Special Issue*, pp. 75-80, 2010.
[6] K. Ranjini, "Performance analysis of hierarchical clustering algorithm," *Int. J. Advanced Networking and Applications*, vol. 03, issue 01, pp. 1006-1011, 2011.
[7] H. Abu-Dalbouh and N. Md. Norwawi, "Bidirectional agglomerative hierarchical clustering using AVL tree algorithm," *IJCSI International Journal of Computer Science Issues*, vol. 8, issue 5, no 1, pp. 95-102, 2011.
[8] K. Sasirekha and P. Baby, "Agglomerative hierarchical clustering algorithm- A review" *International Journal of Scientific and Research Publications*, vol. 3, issue 3, pp. 1-3, 2013.
[9] D. Wei, Q. Jiang, Y. Wei, and S. Wang, "A novel hierarchical clustering algorithm for gene Sequences," *BMC Bioinformatics*, 13:174, 2012.
[10] E. Masciari, G. Massimiliano Mazzeo, and C. Zaniolo, "A new fast and accurate algorithm for hierarchical clustering on euclidean distances," J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 111–122, 2013. Springer-Verlag Berlin Heidelberg 2013.
[11] J. Anuradha and B. K. Tripathy, "Hierarchical clustering algorithm based on attribute dependency for attention deficit hyperactive disorder," *I.J. Intelligent Systems and Applications*, vol. 06, issue 6, pp. 37-45, 2014.
[12] Z. Abdullah and A. R. Hamdan, "Hierarchical Clustering Algorithms in Data Mining," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 10, pp. 2201-2206, 2015.