REVIEW ON DATA COLLECTION, PREPARATION AND CLUSTERING FOR QUALITY PURPOSE

[1]MEGERSA DESTA JELDU, [2]GIZAW TADELE
[1]Parul University, College of Engg, Vadodara, Gujarat, India, [2]Jimma University, JiT, Jimma, Oromia, Ethiopia

Abstract: Clustering is a process of grouping data points into various groups which implies that; data points in the same groups are more similar to each other than those in other groups of data points. The purpose of this task is to segregate groups with similar traits and assign them into clusters. This data clustering implies to statistical data classification technique to discover whether the individuals of a population fall into different classes by making quantitative comparisons of various characteristics. The main two fields of clustering that data focuses on are Partitioning and Hierarchical clustering methods. The paper covers about data clustering methods, data clustering tools, clustering algorithms, benefits and features of clustering tools. Lastly, this Paper concludes by summary and conclusion with core concepts raised points in it.

Keywords: Clustering, Clustering Algorithm, Hierarchical Clustering, K-Means Clustering, Data Clustering Tools

I.INTRODUCTION

Clustering is a process, which is, organising data into meaningful similar groups, and these groups are called clusters. Clustering categorizes objects based on their perceived similarity in their characteristics. Even though clustering is generally a field of unsupervised learning, knowledge about the type and source of the data has been found to be useful in both selecting the clustering algorithm, and for better clustering methods.

This clustering can be seen as a generalisation of classification. It is about wisely knowing and identifying "Where to put the new object in".

The most general definition is that given N items can be divided into k groups based on the measure of similarity between the items in group.

Clusters are not only data points; they also could have different forms which are curves, lines, complex shapes or spirals.

The aim of data clustering is to discover the natural grouping(s) of a set of patterns, points, or objects. Cluster analysis is prevalent in any discipline that involves analysis of multivariate data. It is difficult to exhaustively list the numerous scientific fields and applications that have utilized clustering techniques as well as the thousands of published algorithms [2]

II. METHODS OF DATA CLUSTERING

The goal of clustering is to minimize the amount of data by grouping similar data items together. Such grouping is pervasive in the way human's process information, and one of the motivations for using clustering algorithms is to provide automated tools helping in constructing taxonomies or categories [Jardine and Sibson, 1971, Sneath and Sokal, 1973].

The methods may also be used to reduce the effects of human factors in the process.

Clustering methods can be divided into two basic types: hierarchical and partitional clustering.

Hierarchical clustering is performed either merging smaller clusters into larger ones, or by splitting larger clusters. The end result of its algorithm is a tree of clusters called a dendrogram, which shows how the clusters are related to each other. By removing the dendrogram at a desired level a grouping of the data items into disjoint clusters is obtained [2]. There are two main types of hierarchical clustering.

a. **Agglomerative:**
   - ✓ Start with the points as individual clusters
   - ✓ At each step, merge the closest pair of clusters until only one cluster left.

b. **Divisive:**
   - ✓ Start with one, all-inclusive cluster
   - ✓ At each step, split a cluster until each cluster contains a point (there are k clusters)

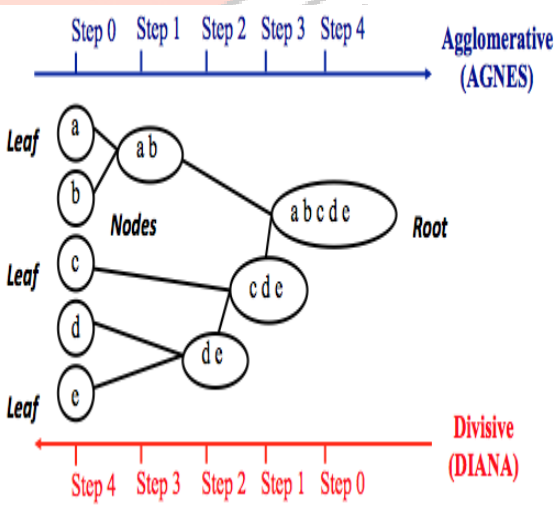

Figure1. Hierarchical clustering in terms of agglomerative and divisive

A partitional clustering a simply a division of the set of data objects into subsets in which each data object is in exactly one subset. The criterion function that the clustering algorithm tries to minimize may emphasize the local structure of the data, as by assigning clusters to peaks in the probability density function, or the global structure. Typically the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters. K-means clustering is commonly used partitional clustering method.

III. K-MEANS CLUSTERING ALGORITHM

K-means clustering algorithm is one of the simplest unsupervised learning algorithms that solve any clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed priory.

The core idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. Then taking each point belonging to a given data set and associating it to the nearest centres. Without pending any data point, the first step is completed and an early group age is done. In this manner, a loop has been generated. As a result of this loop we may understand that the k centres change their location step by step until no more changes are done. In other words, centres do not move any more [7]. Finally, k-means algorithm aims at reducing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in i^{th} cluster.

'c' is the number of cluster centres

A problem with the clustering methods is interpretation of the clusters may be difficult. Majority of clustering algorithms prefer certain cluster shapes, and the algorithms will always assign the data to groups of such shapes even though there were no clusters in the data.

Hence, if the goal is not just to compress the data set, it is essential to analyze whether the data set exhibits a clustering tendency. Another potential problem is selecting or choosing number of clusters is critical; quite different kinds of clusters may emerge when K is altered

Clustering can be used to minimize amount of data points and to induce a categorization. The clusters shall be illustrated somehow for more understanding what they looks like. Example, in the case of the K-means algorithm centroids that represent clusters are still high-dimensional and more additional illustration methods are needed for visualize them.

Use of clustering algorithm

Even on the same data, different clustering algorithms often result in entirely different partitions. An interesting question may be to identify algorithms that generate similar partitions irrespective of the data. The similarity among clustering algorithms is measured as the averaged similarity between the partitions resulted on the datasets.

Clustering algorithms also can be compared at the theoretical level based on their objective functions. To perform such a comparison, a distinction should be made among a clustering method and clustering algorithm. A clustering algorithm is simply an instance of a method. A clustering method is on the other hand, a general strategy employed to solve a clustering problem. Minimizing the squared error is a clustering method, and there are many different clustering algorithms, including K-means, that implement the minimum squared error method. Some equivalence relationships even between different clustering methods have been shown. We can finally say that there is no best clustering algorithm. Each clustering algorithm imposes a structure on the data either explicitly or implicitly. When there is a good match between the model and the data, good partitions are obtained. Since the structure of the data is not known before, one needs to try competing and diverse approaches to determine an appropriate algorithm for the clustering task at hand. This idea of no best clustering algorithm is partially captured by the impossibility theorem [Kleinberg, 2002], which says that "no single clustering algorithm simultaneously satisfies the three basic axioms of data clustering", i.e., scale invariance, consistency, and richness.

IV. DATA CLUSTERING TOOL

Figure 2 below shows, implemented interface for the data preparation tool to generate an object view from a relational database (RDB). Using the information provided by the user via the interface, the algorithm to generate an object-view works as follows: as the DB name and the data set of interest are given, the attributes from the data set of interest in the database are first extracted; then the related attributes in related tables are selected through joining with related tables; lastly, the object attribute(s) is selected from the attributes and the object-view is created by grouping the rows with the same values for the object attribute(s) into one object with the bags of values for the related attributes [5].

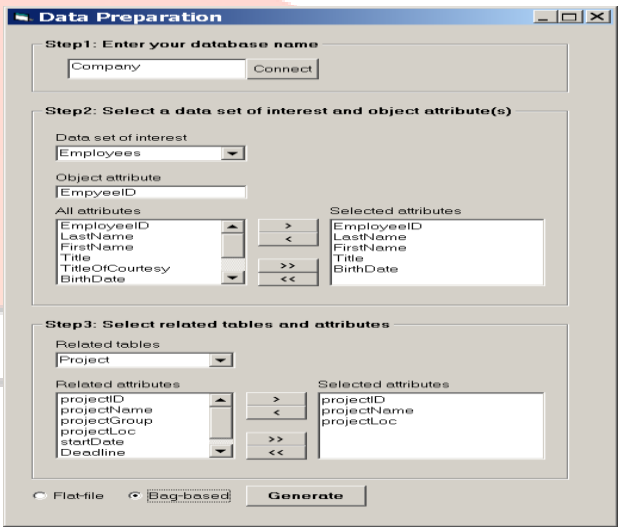

Figure 2. Interfaces for data preparation tool

V. FEATURES OF THE CLUSTERING TOOL

Shown in figure 3, the class diagram for clustering tool in Unified Modelling Language, which is a notational language for software architecture and design. The class diagram describes the developed operations, classes, attributes and the relationships among classes.

Get Analysis Info: class obtains basic information from the user like name of the selected data set, data types for attributes, the interested attributes, and the chosen similarity measure that might be applied to the selected data set.

Read Data Set Objects: class receives or reads the selected data set.

Similarity Measure class defines our similarity measure. For similarity measure in this implementation, the average dissimilarity measure for quantitative attributes and the Tversky's ratio model for qualitative attributes considering the contextual assessment of similarity is chosen.

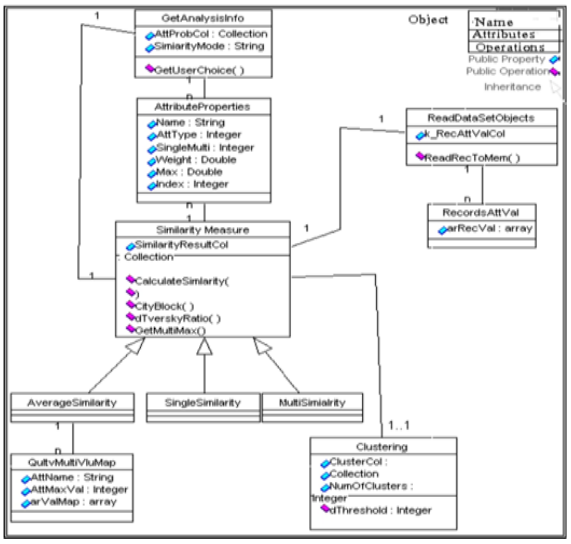

Figure 3. Class diagram for the clustering

VI. SUMMARY & CONCLUSIONS

Organizing data into sensible groupings arises naturally in many scientific fields. It is, therefore, not surprising to see the continued popularity of data clustering. It is important to remember that cluster analysis is an exploratory tool; the output of clustering algorithms only suggest hypotheses. While numerous clustering algorithms have been published and new ones continue to appear, there is no single clustering algorithm that has been shown to dominate other algorithms across all application domains. Most algorithms, including the simple K-means, are admissible algorithms. With the emergence of new applications, it has become increasingly clear that the task of seeking the best clustering principle might indeed be futile. As an example, consider the application domain of enterprise knowledge management. Given the same set of document corpus, different user groups (e.g. legal, marketing, management, etc.) may be interested in generating partitions of documents based on their respective needs. A clustering method that satisfies the requirements for one group of users may not satisfy the requirements of another. As mentioned earlier, "clustering is in the eye of the beholder" – so indeed data clustering must involve the user or application needs.

In most databases, data are stored in several tables or classes and related information are represented as relationships among related tables or classes, while most traditional clustering algorithms assume that input data are stored in a single flat file format. Based on this observation, we showed that the traditional flat file format is not appropriate for storing related information since it restricts each attribute in a data set to have a single value while once related objects in related tables or classes are combined, objects are frequently characterized by bags of values.

Acknowledgement

First of all, we would like to thank Mighty God for His help through all our works. Also special thanks to our families for their moral and financial support to be here.

REFERENCES

1. Tae-Wan Ryu, "A Database Clustering Methodology and Tool" Department of Computer Science, California state university, Fullerton, Fullerton, California 92834, 2005.
2. T. Soni Madhulatha, "An Overview On Clustering Methods", Apr. 2012
3. Christoph F. Eick, "A Database Clustering Methodology and Tool" Department of Computer Science, University of Houston, Houston, Texas 77204-3010, 2005.
4. Anil K. Jain, "Data Clustering 50 Years Beyond K-means" Department of Computer Science Michigan State University.
5. E. Backer, "Computer-Assisted Reasoning in Cluster Analysis", Prentice Hall, London, 1995.
6. Workineh Tesema, "Afan Oromo Sense Clustering in Hierarchical and Partitional Techniques", Department of Information Science, Jimma University, Jimma, 378, Ethiopia, 2016.
7. Jain AK, Dubes RC, "Algorithms for Clustering Data", 1948.
8. M. Amadasun and R. A. King, "Low-level segmentation of Multispectral images via Aglomerative Clustering of Uniform Neighborhoods", Pattern Recognition 21(3): 261-268, 1988.
9. M.R Anderberg, "Cluster Analysis for Applications", Academic Press, Inc., New York, 1973.
10. E. Backer, "Computer-Assisted reasoning in Cluster Analysis", Prentice Hall, London, 1995.
11. G. Biswas, J. Weinberg, and C. Li, " A conceptual Clustering Method for Knowledge Discovery in Databases", Editions Technip, 1995.
12. V. Brailovski, "A probabilistic approach to clustering", pattern recognition letters 12(4): 193-198, 1991.
13. Yizong Cheng, "Mean Shift, Mode Seeking, and Clustering", IEEE Transactions on pattern analysis and machine intelligence, vol. 17, No. 8, 1995.
14. G.B. Coleman and H.C. Andrews, "Image Segmentation by Clustering", Proc. IEEE 67(5): 773-785, 1979