www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 5 May 2016, Page No. 16504-16508

An Analytic Survey On Current Clustering Technique of data Categorizing and Retrieving

Snehlata Bhadoria, U. Datta
MPCT
Gwalior, India
sneha.singh2k17@gmail.com

MPCT
Gwalior, India
deanacademics@mpct.org

***Abstract*— in this survey we have a furnished detail of Clustering. Clustering is not only bounded in boundary of grouping of same kind of objects in cluster, it would also be like to get or retrieve specific data by just analyzing clustering approach. This analytic survey focused on the current clustering technique of data categorizing and retrieving as faster as possible from huge amount of data because data is growing like square or cube of their current position. So saving of all information and easier retrieving will always face new challenges as proportion of data increasing in various aspects which would population of any country or data of any field related to them.**

Keywords—Density Based Clustring, OPTICS, DENCLUE, DBSCAN, CURE

INTRODUCTION

Clustering is the well-known technique to perform fast data accessing approaches from huge amount of data. Through this technique large data could divide, categorized and grouped by their similar attributes. Clustering is challenging than classification. High dimension dataset, arbitrary shapes of clusters, scalability, input parameter, domain knowledge and handling of noisy data are the basic requirement of cluster analysis. A various algorithms had been proposed yet, each addresses some specific requirements.

In this paper we have provided a detailed analytical comparison of some of the very well-known clustering algorithms.

Clustering could be classified into five major categories Partitioned technique, Hierarchical technique, Density-Based technique, Grid-Based technique and Model-Based technique as shown in Fig SRW 1.1.

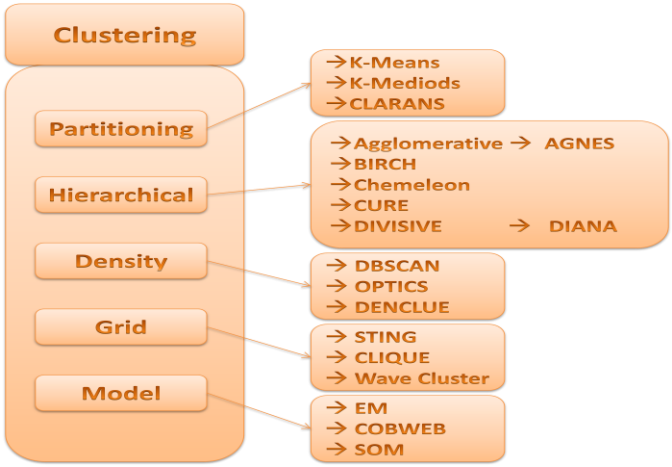

Fig SRW 1.1: Various Clustering Techniques

A. AIM:

The aim of that work is to survey the existing clustering algorithms which could process huge data and observe the feasibility of the algorithms to handle data with noise and outliers has been studied so as to confirm the requirements of specific application. Data set

size, Data dimensionality, Time complexity are main factors through which the comparative analysis of the algorithms could performed.

B. Survey overview

Section 2 is a summarized survey of basic and effective clustering approaches. Section 3 is representing the summarized table of comparison. Section 4 presents conclusions of the work.

BASIC AND EFFECTIVE CLUSTERING APPROACHES

There are five methodologies those are implemented for clustering process. The Partitioned technique, Hierarchical technique, Density-Based technique, Grid-Based technique and Model-Based technique are widely implemented for clustering. In this work all of them are briefly highlighted below.

A. Partitioning Techniques

In this technique number of k partitions of datasets is made with n objects, each partition represent a cluster, where k<= n. It tries to divide data into partition based on some evolutionary criteria. As checking of all possible partitions are computationally infeasible, certain greedy heuristics are used in the form of iterative optimization [5].

One such kind of approach to perform partition is based on the objective function, in which, instead of pair-wise computations of the proximity measures, unique cluster representatives are generated depending on how representatives are made iterative partitioning algorithms are classified into k-means and k-mediods [3] [8].

The partitioning algorithm in which each cluster is modeled by the gravity of centre is known as k-means algorithm. The one most efficient algorithm proposed under this scheme is known as k-means only.
The partitioning algorithm in which cluster is represented by one of objects located near of its centre is called as a k-mediods. PAM, CLARA and CLARANS are three main algorithms proposed under the k-mediod method [11].

B. Hierarchical Techniques

The hierarchical approach decomposes dataset of n objects into a hierarchy of groups. This hierarchical decomposition can be represented by a tree structure diagram known as *dendrogram*; whose root node represents the whole dataset and each leaf node is a single object of the dataset. The clustering results could be obtained by cutting the dendrogram at different level. There are two general approaches for hierarchical approach: agglomerative (bottom-up) and divisive (top down) [2] [11]. An hierarchical agglomerative clustering (HAC) or agglomerative method starts with n leaf nodes or n clusters. that is by considering each object in the dataset as a single cluster (node) and in successive steps applied merge operation to reach at root node, which is a cluster containing all data objects. The merge operation depends on the distance between two nodes. The Distance has three different notions of single, average and complete link.

A hierarchical divisive clustering (HDC) or divisive method, opposite to agglomerative, starts with root node which considers all data objects into single cluster and in successive steps divide dataset until reach to a leaf node containing a single object. For a dataset with n objects have 2n-1 – 1 possible two-subset divisions, which is very expensive in computation. The major problem with hierarchical methodology is its selection of merge or split points, as once done cannot be undone. This problem also causes impacted scalability of this method. Thus, in general hierarchical methods are used as one of the phase in the multi-phase clustering. Various algorithms proposed under these concepts are: BIRCH, ROCK and Chameleon [3] [8] [11].

C. Density Based Technique

The density based concept developed based on the density notation, which is the no of objects in the assigned cluster, in this context. The general idea is to continue growing the given cluster in proportion to the density in the neighborhood exceeds some threshold; that is for each data point within a given cluster; the neighborhood of given radius has to contain at least a minimum number of points.
The basic idea of density based clustering involves number of new below definitions

a. ε-neighborhood: the neighborhoods within radius ε of a given object is called as ε-neighborhood of the object.
b. Core object: if the ε-neighborhood of an object contains at least a minimum number, MinPts, of objects, then the object is called a core object.
c. Border point: A border point has fewer than MinPts within radius ε, but is in the neighbourhood of a core point.
d. directly density-reachable: given a set of objects D, an object p is directly density-reachable form object q if p is within the ε-neighbourhood of q, and q is a core object.

e. (Indirectly) density-reachable: an object p is density-reachable by object q w.r.t ε and MinPts in a set of objects, D, if there is a chain of objects p1,……..pn, where p1 = p and pn = q like pi+1 is directly density-reachable by pi w.r.t ε and MinPts, for 1≤i≤n.

f. Density-connected: an object is density-connected to object q w.r.t ε and MinPts in a set of objects, D, if there is an object o in D like both p and q are density-reachable from o w.r.t ε and MinPts.

The density based algorithms could again classify as: density based on connectivity of points and based on density function. The main algorithms in this former are DBSCAN and its extensions, OPTICS, whereas under the latter category are DENCLUE [3] [4] [6] [9].

D. Grid Based Techniques

As the name suggest, grid based clustering methodology uses a multi-dimensional grid data structure. It divides the object space into finite number of cells those forms grid structure on which all of the o algorithms in the former of DBSCAN and its extensions, partitions for clustering are performed. One of the distinct features of this method is the fast processing time, as it not dependant on the number of data objects but only on the number of cells. The algorithms based on this technique are STING, Wave Cluster, and CLIQUE [9].

E. Model Based Technique

Attempt to optimize the fit between given data and some mathematical model based on the assumption that Data are generated by a mixture of underlying probability distribution.
Typical methods are under this category are:
» Statistical approach• EM (Expectation maximization), Auto Class
» Machine learning approach• COBWEB, CLASSIT
» Neural network approach • SOM (Self-Organizing Feature Map)
These methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering used to identifies groups of objects, model-based clustering methods also find characteristic descriptions for each group, where each group represents class. The most frequently used induction methods are decision trees and neural networks.

a. Decision Trees.

The data is represented by a hierarchical tree in decision tree, where each leaf node refers to the concept and contains a probabilistic description of that concept. Several algorithms produce classification trees to represent unlabelled data. The most well-known algorithms are: COBWEB — This algorithm consider that all attributes are independent. Its aim is to achieve high predictability of nominal variable values with assigned cluster. This algorithm is not suitable for clustering large database data. CLASSIT, an extension of COBWEB for continuous-valued data, unfortunately has similar problems as the COBWEB algorithm.

b. Neural Networks.

This algorithm represents cluster by a neuron or prototype". Neurons represent input data at first layer, which are connected to the prototype neurons. Each such connection has a weight, which is learned adaptively during learning. A popular neural network algorithm for clustering is the self-organizing map (SOM). This algorithm based on a single-layered network. The learning process takes place in a "winner-takes-all" fashion, the prototype neurons compete for the current instance. The winner is neuron whose weight vector is closest to the instance currently presented.
The winner and its neighbors learn by having their weights adjusted. The SOM algorithm is successfully used to vector quantization and speech recognition work. It is useful for visualizing high-dimensional data in 2D or 3D space with sensitivity at initial selection of weight vector, as well as to its different parameters, such as the learning rate and neighborhood ra- dius.

COMPARATIVE ANALYSIS

The clustering is more challenging task than classification. various algorithms had been proposed till data, each to solve some specific issues. No clustering algorithm could adequately to handle all sorts of cluster structure and input data. A detailed comparative study of different clustering algorithms proposed under the different methods by considering the different aspects of clustering is given in table SRW 1.1. In table we had provided the remarks for each of the algorithm which gives the clear idea of the advantages and disadvantages of each of the algorithms.

DOI: 10.18535/ijecs/v5i5.28

Table SRW 1.1: Analytic comparison among various clustering technique

Name	DBSCAN	BIRCH	STING	DENCLUE	CLIQUE	Wave Cluster	OPTICS	ROCK	CHAMELEON
Proposed By	Martin Ester, Hans-Peter Kriegel & Xiaowei Xu	Zhang, Ramakrishnan & Linvy	Wang Wei, Jiong Yang & Richard Muntz	Hinneburg & Keim	Agrawal Rakesh, Johannes Gehrke, Dimitrios Gunopulos & Prahhakar Raghavan	Sheikholeslami, Gholamhosein, Surojit Chatterjee & Aidong Zhang	Ankerst	Guha Sudipto, Rajeev Rastogi & Kyuseok Shim	Karypis
Year	1996	1996	1997	1998	1998	1998	1999	1999	1999
Complexity	O(nlogn)	O(n)	O(k)-	O(n2)	Quadratic on # of dimensions	O(n) for low dimension	O(nlogn)	O(n2)	O(n2)
Types of Data	Numerical	Numerical	numerical	Numerical	Mixed	Numerical	numerical	Categorical	Discrete
Data Set	High Dimensional	Large	Any size	High Dimensional	High Dimensional	Large	High Dimensional	Small sized	Small
Cluster Shape	Arbitrary	Spherical	Rectangular	Arbitrary	Arbitrary	Arbitrary	Arbitrary	Graph	Arbitrary
Input Parameter	a) radius b) minimum points	branching factor B, threshold T (max. diameter of sub cluster)	Statistical	density parameter, noise threshold	density threshold	No	density threshold	similarity threshold	Min. Similarity
Remarks	can handle noise, more efficient than partitioning and hierarchical methods - Efficiency is dependent on the number of different input parameter –Can't handle clusters of different densities	Time complexity is linear works well only for spherical clusters	Support parallel processing and incremental updating, efficiency	Solid, mathematical foundation, good clustering properties with large amt of noisy data set, compact representation of clusters	insensitive to order of input, scales well -results are highly dependent on the input parameter	⁺ outperforms BIRCH, CLARANS & DBSCAN in terms of both efficiency and clustering quality, capable of handling data with up to 20 dimensions	⁺ No need for input parameter settings -Cannot handle clusters of different densities	⁺ based on HAC ⁺ more powerful than traditional hierarchical clustering	⁺ high quality clusters

I. CONCLUSION

Clustering is still a huge area of possibilities of efficient and fast data grouping and fast data accessing from the created group of data or data set. The clustering is applicable in every field of science and technology in which data manipulation is required. Many clustering algorithms had been proposed yet which satisfy certain criteria such as arbitrary shapes, high dimensional database, and domain knowledge and so on. So, it is difficult to select any algorithm for a specific application. In this paper we have comparison of the clustering algorithms by which we tried to satisfy our two aims first choosing most appropriate clustering algorithm for specific task and second is to make new clustering technique which could survives as proportional to data growth as longer as possible.

References

[01] Smiti, Abir, and Zied Eloudi, "Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory", In the IEEE 6th International Conference on Human System Interaction 2013, pp. 380-385, 2013.
[02] Neha Soni, Amit Ganatra, "Categorization of Several Clustering Algorithms from Different Perspective: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 8, pp. 63-68, Aug. 2012.
[03] Pragati Shrivastava and Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research, pp. 2249-7277, September 2012.
[04] Chaudhari Chaitali G., "Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Vol.2, Issue-2, pp. 212 – 215, December 2012.
[05] Chakraborty S., Prof. Nagwani N. K., Analysis and Study of Incremental DBSCAN Clustering Algorithm, International Journal of Enterprise Computing And Business Systems, Vol. 1, July 2011.
[06] Chandra. E, Anuradha. V. P, A Survey on Clustering Algorithms for Data in Spatial Database Management System, International Journal of Computer Applications, Col. 24, June 2011.
[07] Parimala M., Lopez D., Senthilkumar N. C., A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases, International Journal of Advanced Science and Technology, Vol. 31, June 2011.
[08] Santhisree K., Dr. Damodaram A., SSMDBSCAN and SSM-OPTICS : Incorporating new similarity measure for Density based clustering of Web usage data, in International Journal on Computer Sciences and Engineering, August 2011.
[09]Anant Ram, Sunita Jalal, Anand S. Jalal, Manoj kumar, "A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases", International Journal of Computer Application Vol. 3, No.6, June 2010.
[10] Dr. E. Chandra, V. P. Anuradha, " A Survey on Clustering Algorithms for Data in Spatial Database Management Systems", International Journal of Computer Application, vol. 24, pp. 19-26.
[11] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means, in Pattern Recognition Letters, vol. 31 (8), pp. 651-666, 2010.
[12] Peter J. H., Antonysamy A., An optimized Density based Clustering Algorithm, International Journal of Computer Applications, Vol. 6, September 2010.
[13] Ram A., Jalal S., Jalal A. S., Kumar M., A Density based Algorithm for Discovering Density varied clusters in Large Spatial Databases, International Journal of Computer Applications, Vol. 3, June 2010.