ISSN:0975-9646

Deepali N. Gunjate et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3181 - 3185

Immense Review on Clustering Approaches

Deepali N. Gunjate, Prof. B. R. Kanawade

Dnyanganga College of Engineering and Research
Pune University, Pune, India

***Abstract* ---- Clustering is process for finding similarity groups in data. It is considered as unsupervised learning task. The motive of clustering is detecting groups of similar objects by separating dissimilar ones. Clustering plays wide role in data mining. Old versions of clustering methods compute a division of the data, grouping each object in at most one cluster or detecting it as noise. However, it is not always necessary that an object must belongs to only one cluster. It means that multiple meaningful groupings might exist for each object. In today's any applications, high dimensional databases are used. As multiple concepts described in the same data set by different attributes are mixed. And clusters are hidden in the subspace projections and do not exist in all directions. Furthermore, as a general contribution to the community, the benefits of evaluation study and evaluation framework are described. Both provide important basis for future research and ensure compatibility and repeatability of experiment results. To describe the objective more correctly it is necessary to define it by all possible and meaningful directions. To find out the similarity between objects become difficult because of increase in number of attributes. Hence. Its difficult to discover the distribution pattern. Thus, it is necessary to look for meaningful grouping in subspaces. This paper presents an overview of clustering techniques, its significance along with its various methods of implementation and challenges that need to be overcome.**

***Keywords*— Data mining, Subspace clustering, High Dimensional Data**

I. INTRODUCTION

The motive of this survey is to provide a inclusive review of different clustering techniques in data mining. Clustering is defined as division of data into groups of similar objects. Each group is named as cluster, which is made up of such objects that are similar to one another and dissimilar to objects of other groups.

From a machine learning perspective clusters correspond to hidden patterns, the search for cluster is unsupervised learning and the resulting system represents data concept. Hence, clustering is unsupervised learning of hidden data concepts. Data mining applications include scientific data exploration, information retrival, text mining, spatial databases, web analysis, computational biology, customer relation management, medical diagnostics and many others. The real challenges are presented to classic clustering algorithms by them.

For convenience of reader, we provide a classification of clustering algorithms followed by this review-

1. Hierarchical Methods
2. Partitioning relocation methods
    - k-medoids method
    - k-means method
3. Density based partitioning methods
    - Density based connectivity clustering
    - Density function clustering
4. Grid based methods
    - CLIQUE method
5. Algorithms for high dimensional data
    - Subspace clustering

An important tool for finding data analysis, which leads to summarize main features of data. Huge work has already done in past for this work. K- means algorithm is type of clustering algorithm, has history of fifty years[4]. Clustering techniques have been developed for categorical data. Clustering techniques are applied in pattern recognition [5], image processing and information retrival. Clustering has a huge history in other disciplines [6] such as psychiatry, biology, geology, geography, archaeology, psychology and marketing [7]. Cluster searching methods differs according to need of application. Clustering algorithms and techniques are based on number of samples to be considered, size of single sample, precision of result required. All these parameters cause for various techniques and algorithms. This paper presents an overview of clustering techniques, their differentiation, benefits and pitfalls of them and problems that needs attention. Part II explains methods of clustering, part III of this paper explore the working of clustering in case of high dimensional data. Part IV emphasis on subspace clustering. Comparison of different methods will be covered in part V and lastly discuss applications and conclusion in rest of the section.

II. CLUSTERING METHODS

Hierarchical clustering and partitioning clustering are very famous methods of clustering [6].

A) Hierarchical Clustering

Data set is subdivided into distinct partitions, in hierarchical clustering. There are two possible ways for partitioning methods, one is agglomerative and other is divisive. In agglomerative perspective, at beginning each object is considered as single cluster, later objects are combined together based on the criteria followed. In divisive perspective, originally complete data set is treated as cluster. Complete cluster is then divided into smaller sets based on their similarity. This method of hierarchical clustering produces a data structure, named as dendogram. A traditional dendogram is as shown in fig. 1 [6]. Dendogram explains the sequence in which the data sets are combined. Dendogram serves the hierarchy of clusters and therefore the name. When we crop the dendogram at a point then we obtain the cluster at that level.

At the beginning, x1 and x2 are objects and they are grouped together, as shown in the figure. This grouping of objects x1 and x2 is possible only if x1 and x2 manifest minimum level of similarity needed by the application. Objects x3 and x4 are also merged together at the same level. This merging on similarity two objects allocates. Any distance calculating measure can be similarity measure such as Euclidian distance, Manhatten distance. There are two types of hierarchical clustering one is agglomerative and other is divisive.

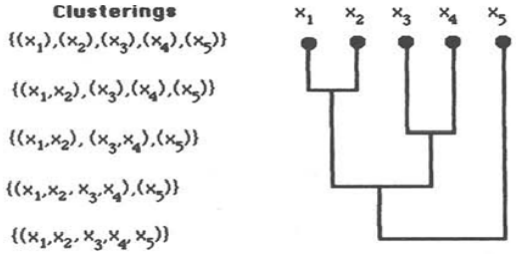

Fig. 1 A Dendogram demonstrating hierarchical clustering

B) Agglomerative Perspective:-

Bottom up approach is used to cluster the objects. Originally, each sample object is a cluster in its own. At start, it holds n-clusters, where n- represents the number of objects in the data set. Algorithms in this category iterate to merge objects which are similar [8] and stop when, only one cluster left having all n –data objects. The most similar data objects are merged by method, in each stage [4,9]. To find out the cluster border, standard is required. For this, correct perspective is to determine, what should be the distance between two clusters [10]. Single linkage clustering and complete linkage clustering came into picture, based on the distance measure used. In both methods, distance between two objects is calculated wisely, at the beginning only. One object belongs to first cluster and other object is from second cluster. In single linkage method, distance between two clusters is same as the minimum of all the pair wise distance calculated. In complete linkage method, distance between two clusters is the maximum of all the pair wise distance calculated [7]. Thus single linkage method uses minimum distance standard or nearest neighbor approach. Let A and B are two sets of objects considered as clusters. Let D(A,B) shows distance between cluster A cluster B; and d(a,b) is the distance between two elements a and b. in case of single linkage method, distance is as given below:

D(A,B) = min d(a,b) when a € A and b € B. In case of complete linkage method, distance is as given below:

D(A,B) = max d(a,b) when a € A and b € B.

1) Divisive Approach

This approach is completely opposite to the agglomerative perspective. It starts with only one cluster having complete objects as the member of that one cluster. After that, it keeps on dividing until there are n- clusters, each having only one object. This particular approach is able to partition the huge set of objects into small groups. Attributes of an objects plays an role of standard for division, which brings to new two views that are monoethic and polyethic techniques respectively [9]. A cluster with N objects gives 2^{N-1}-1 possible two subset partitions [1], which is exorbitant in computation. Hence this perspective in not mush usual in practice.

2) Partitioning Clustering

In this technique, each object is kept in perfectly one set. An object cannot share two different sets. In this, user needs to provide the number of clusters as input. This particular method provides another category of clustering algorithm, named as non – overlapping clustering algorithms. Result of partitioning techniques is based on the user input and hence same algorithm with same input can produce various output with huge difference between each of them based on number of clusters needed.

III. HIGH DIMENSIONAL DATA CLUSTER ANALYSIS

In data mining, the objects can have hundreds of attributes. Clustering in such a high dimensional spaces presents huge difficulty. In customer behavior analysis, customer purchases some amount of e of product is nothing but the attribute. We can consider the example of "Description data of glass". In this example of glass, all the components of glass are considered as various attributes like aluminum, magnesium, potassium, calcium, silicon, etc. and the percentage of component in one mg is considered as the value for an attribute. Like this there are huge numbers of real world situation, where many attributes are required to describe the data. It becomes very tedious task to analyze or to understand the data pattern, if the count of attributes increases. It is very difficult to find out the similarity between two objects, if they are not exactly same. There may be the possibility of case that two objects are not similar in all attributes, for this purpose we need to consider all the attributes. This is nothing but the subspace. The aim of modern techniques is to investigate the clusters in subspaces.

A) Hurdle in High Dimensional Data

"Curse of dimensionality" is the barrier for high dimensional data [12]. Two objects can be equidistance because of increase in dimensions. The dense regions which could be cluster in space are lost, if the count of objects decreases with increase in dimensions in a one container.

Let us take example of glass we have variety of glass as objects and component of glass will be considered as attributes of the objects. Let us assume any two components randomly such as silicon and magnesium and hence we take two dimensional data. We can represent this data in tabular form as below:

TABLE I
Sample Data Set

A	5	6	7	8	9	10	15	17	20
B	2	3	4	16	17	18	19	20	20

This data can be represented graphically, consider x-axis for magnesium and y- axis for silicon.

A = {5, 6, 7, 8, 9, 10, 15, 17, 20}
B = {2, 3, 4, 16, 17, 18, 19, 20, 20}
(Two dimensional) graph of sample data is in fig. 2

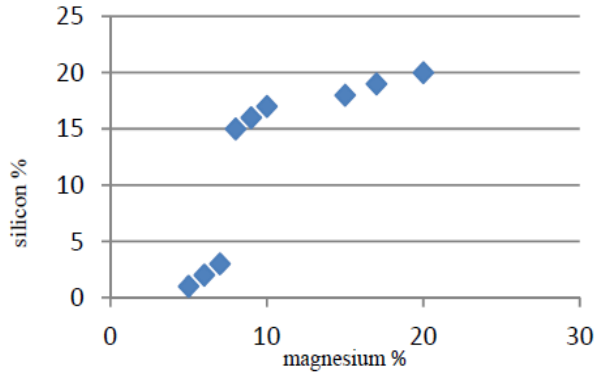

Fig. 2: Graph for objects in space in two dimensional views

Let us assume the range of 0-15 as one unit container. As it is 2-D graph of points, we get first four points in one container, considering both axes x and y. if we assume only one dimension axis x then we get graph as shown in fig. 3 that means we get all objects in one unit container and if we assume both x and y axes dimensions, then the count of objects per container reduces to four only. Hence, as the counter of objects in one unit container decreases, the count of dimension increases that means density decreases.

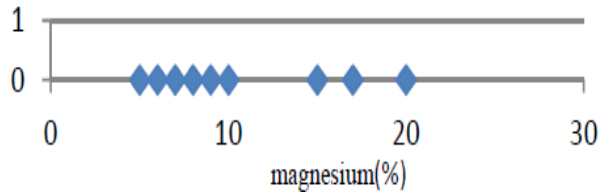

Fig. 3: graph of objects considering only one attribute.

We cannot provide the definition of the best cluster; definition varies from user to user, as per his demand. Originally, it is mandatory to check the algorithm by manual testing of result. Once the results are tested for quality then it is assured that for any set of data algorithm will generate some satisfactory output.

B) Dimensionality Reduction

Feature selection and feature transformation are included in dimensionality reduction [6]. Converting original sets of attributes into new is nothing but the feature transformation. There are varieties of transformation techniques available such as linear transformation, wavelet transformation, etc. feature selection incorporates selection of the most relevant set of attributes and excludes the rest. Let us consider the example, instead of having 3 attributes for particular student details namely name of student, surname of student and roll number of student. We can simply concentrate on roll number attribute and neglect the other two. However, when we cannot avoid the attributes as all attributes gives useful information then we can use this approach. The solution to this is given by researchers is to look clusters into subspaces.

IV. SUBSPACE CLUSTERING

Clustering faces the problem of curse of dimensionality [12]. If the dimension increases then data will get lost. Dimensionality reduction cannot allow excluding any of the attribute. Subspace clustering came into picture to solve this complication. The purpose of subspace clustering is to search such clusters which are not visible in subspaces and which cannot be recognized if we take all the attributes simultaneously. CLIQUE (CLUstering In QUEst) was the very first algorithm which manages the problem with large number of dimensions using subspace approach [13, 14]. ENCLUS algorithm (ENttropy based CLUstering) and MAFIA algorithm (Merging of Adaptive finite InervAls) [13, 16] was on the same path. Subspace clustering algorithm can be classified into two main types as grid based algorithms and density based algorithms.

A. Grid based algorithms

In grid based technique data set is divided into small grids with width w. objects belongs to specific grid structure are the part of cluster in that region. For each grid, density is searched and cells are arranged to their densities and center of cluster is search out. Hence, computational complexity of grid is lesser dependant on algorithm. Such algorithms are easy to design and fast in computing output. But the drawback of grid based method is shape of the cluster. Example of grid based algorithm are OptiGrid, STING(Statistical INformaton Grid based method), GRIDCLU [17].

1. *CLIQUE*: R. Agrawal et al. proposed CLustring In QUEst [14] algorithm in 1998. It produces cluster details in DNF expression form. This algorithm is indistinguishable to Apriori algorithm [18] which is used to mine frequent data items. Only grid based algorithm can divide data space into grid. The coverage term came into picture to select the subspace. Formal definition of coverage is fraction of dataset covered by dense units. Subspaces with high coverage value are kept remaining are cropped. CLIQUE gives result which is thoughtless to the given input data order. Algorithm is speedy and scale well with number of dimensions in output cluster.

B. *Density Based Algorithm*

The disadvantage of grid based algorithm is defeated by Density based algorithm. Algorithms of this types are DBSCAN [19] (Density Based algorithm for Discovering Clusters in Large Spatial Databases With Noise), DENCLUE [15] (Density Based CLUstering), SUBCLU is the method which is able to search all shaped and sized clusters.

1. SUBCLU

SUBCLU (SUBspace CLUstering) was introduced by Karin kailing, Hans-peter kriegel and peer Kroger [20] a density based algorithm. Clusters are named as density connected sets as per this algorithm. This algorithm works an bottom – up approach for searching cluster. It executes DBSCAN [19] algorithm as a first step on each dimension to form single dimension cluster. Similarly, DBSCAN [19] is applied on all (suitable) applicable subspaces. SUBCLU is a density based

algorithm hence it is capable for recognizing all shaped clusters. Idea of core object is registered in this algorithm. Minimum threshold and epsilon radius are two inputs of this algorithm. Epsilon radius is a area in which we get the similar objects. Minimum threshold value states that there are minimum number of objects which form cluster together. Core object is an object whose ε area has minimum threshold number of objects, such core objects along with neighboring objects obtain the cluster.

V. DIFFERENTIATION OF CLUSTERING ALGORITHMS

We have to select appropriate algorithm which gives best result for our requirement among all convenient algorithms. In case of hierarchical algorithm, time complexity various according to the object count that means increased number of objects leads to increase in time complexity. In case of partitioning algorithm, number of cluster is provided by the user as input, hence user must have idea about data distribution, its requirement for this particular algorithm. Grid based algorithm are not enough capable to recognize all shaped and sized clusters. Density based algorithms handle noise and can search clusters of all shapes. Differentiation of all these algorithms covered in table II.

TABLE II DIFFERENTIATION OF CLUSTRING ALGORITHMS

Type	Time Complexity	Input	Algorithm
Hierarchical	Typically $O(n^2)$	Branching factor, radius of cluster	BRICH[1] [22, 23] CURE[2] [21]
Partitioning	Typically O(n) [10]	Number of clusters needed	K- mean, K- mode, PAM
Grid based	Typically O(n)	Grid size, number of objects in cell	STING
Density based	Typically O(n log n)	Threshold, radius	DBSCAN, DENCLU

1 Balanced Iterative Reducing and Clustering using Hierarchies.
2 Clustering Using REpresentatives.

VI. SIGNIFICANCE

A) Web mining

In this area, clustering helps to find out the homogeneous and meaningful groups of documents on web.

B) Compression of data

Clustering plays wider role in data compression. Instead of mining whole data set we can mine clusters which are inductive for group of similar objects [10].

C) Spatial data analysis

Medical equipment, Geographical Information Systems (GIS), image database exploration, satellite images, etc all these produces tremendous amount of data. It is costly and harder for user to examine spatial data in detail. Clustering provides to automate the process of analyzing and understanding spatial data.

VII. CONCLUSION

The major provocation for clustering high dimensional data is to solve the "Curse of dimensionality" problem. There are n- numbers of latest techniques for clustering high dimensional data which are strongly applied in many areas. Comparison of all these methods is needed to grapes their advantages and disadvantages. The method which is most suitable for high dimensional data, it is not necessary that particular method will be suited to all other data distribution.

We need to concentrate on the technique which provides output in such a way that which is easy to interpret. Generated result should be able to produce some conclusion and data distribution details.

ACKNOWLEDGMENT

I would like to express gratitude towards my guide Prof. B. R. Kanawade for her punctual help. Except her guidance it would have not been possible to complete this task. I would also like to show gratitude towards my coordinator and head of department whose motivation enabled me to present this task.

REFERENCES

[1] Rui Xu and W. Donald, "Survey of Clustering Algorithms," *IEEE Transaction on Neural Network,* vol. 16, 2005.
[2] A Survey of Clustering Data Mining Techniques Pavel Berkhin Yahoo!, Inc. pberkhin@yahoo-inc.com
[3] S. Harkanth "A Survey on Clustering Methods and Algorithms",
[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of the Fifth Berkeley Symp. On Math. Stat. and Prob.,* vol. 1, pp. 281-296, 1967.
[5] Gan Guojan, Ma Chaoqun, and W. Jianhong, *Data Clustering: Theory, Algorithm and Applications*. Philadelphia.
[6] aA. Jain and R. Dubes, *Algorithms for Clustering Data*. New Jersey,1948.
[7] A. K. Jain, M. N. Murtyand, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys* vol. 31, pp. 264-324, 1999.
[8] P. Cimiano, A. Hotho, and S. Steffen, "Comparing Conceptual,Divisive and Agglomerative Clustering for Learning Taxonomiesfrom Text," in *European Conference On Artificial Intelligence*, 2004.
[9] B. Everitt, S., S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*.West Sussex, 2011.
[10] M. Halkidi, Y. Batistakis, and V. Michalis, "On Clustering Validation Techniques," *Journal of Intelligent Information Systems,* vol. 17, pp. 107-146, 2001.
[11] K. Bache and M. Lichman. (2013). *UCI Machine LearningRepository*.Available: http://archive.ics.uci.edu/ml/machinelearning-databases/
[12] M. Steinbach, L. Ertoz, and V. Kumar, " The Challenges of Clustering High Dimensional Data.," in *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition*, ed New Vistas: Springer, 2004.
[13] L. Parson, E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A Review " *Sigkdd Explorations,* vol. 14, pp. 90-106, 2004.
[14] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 1998, pp. 94-105.

[15] C. H. Cheng, A. W. Fu, and Y. Zhang, "Entropy-based subspace clustering for mining numerical data," in *International conference on Knowledge discovery and data mining*, 1999, pp. 84-93.
[16] S. Goil, H. Nagesh, and A. Choudhary, "Mafia: Efficient and scalable subspace clustering for very large data sets," Technical Report, Northwestern University, 1999.
[17] Y. Zhao and J. Song, "GDILC: a grid-based density-isoline clustering algorithm," in *International Conferences on Info-tech and Info-net Proceedings*, Beijing, 2001.
[18] R. Agrawal and S. Ramakrishnan, "Fast Algorithms for Mining Association Rules," in *International Conference on Very Large Data Bases*, 1994.
[19] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 169-194.
[20] K. Kailing, H. Kriegel, and P. Kroger, "Density Connected Subspace Clustering for High - Dimensional Data " in *International Conference on Data Mining*, Lake Buena Vista, FL, 2004.
[21] G. Sudipto, R. Rajeev, and S. Kyuswok, "CURE: An Efficient Clustering Algorithm for Large Databases," in *International Conference on Management of data*, 1998, pp. 73-84.
[22] Z. Tian, R. Raghu, and L. Miron, "BIRCH: A New Data Clustering Algorithm and Its Applications," *Data Mining and Knowledge Discovery,* vol. 1, pp. 141-182, 1997.