Vol 2 Issue 9 March 2013

Impact Factor : 0.1870 ISSN No :2231-5063

Monthly Multidisciplinary Research Journal

Golden Research Thoughts

IMPACT FACTOR : 0.2105

Welcome to ISRJ

RNI MAHMUL/2011/38595 **ISSN No.2230-7850**

Indian Streams Research Journal is a multidisciplinary research journal, published monthly in English, Hindi & Marathi Language. All research papers submitted to the journal will be double - blind peer reviewed referred by members of the editorial Board readers will include investigator in universities, research institutes government and industry with research interest in the general subjects.

Address:-Ashok Yakkaldevi 258/34, Raviwar Peth, Solapur - 413 005 Maharashtra, India
Cell : 9595 359 435, Ph No: 02172372010 Email: ayisrj@yahoo.in Website: www.isrj.net

Golden Research Thoughts
Volume 2, Issue. 9, March. 2013
ISSN:-2231-5063

Available online at www.aygrt.isrj.net

ORIGINAL ARTICLE

DATA CLUSTERING ALGORITHMS – A SURVEY

P.PRABHU AND N.ANBAZHAGAN

Assistant Professor in Information Technology
DDE, Alagappa University, Karaikudi, Tamilnadu, India.
Associate Professor, Department of Mathematics
Alagappa University, Karaikudi, Tamilnadu, India

Abstract:

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Clustering is used in many areas, including artificial intelligence, biology, customer relationship management, data compression, data mining information retrieval, image processing, machine learning marketing, medicine, pattern recognition, psychology and statistics. This paper gives survey of various types of clustering algorithms. It describes its functionality, parameters needed and the time and space complexity required for clustering.

KEYWORDS:

Clustering, large database, Partitioning Clustering, Hierarchical clustering.

1.INTRODUCTION:

Data mining deals with large databases that impose on clustering analysis with additional severe computational requirements. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [10]. There are various types of clustering algorithms that has been proposed in the literature. The objective of this paper is to provide a survey of different clustering techniques used in data mining.

2 CLASSIFICATION OF CLUSTERING ALGORITHMS

Clustering algorithms can be divided into the following categories: They are,

Partitioning Methods
Grid Based Methods
Model Based Methods
Hierarchical Methods
Other Clustering methods

2.1 Partitioning Methods

Source:Golden Research Thoughts [2231-5063] P.PRABHU AND N.ANBAZHAGAN. yr:2013 vol:2 iss:9

Partitioning methods attempt to directly partition the data into disjoint clusters (first creates initial k partitions, where parameter k is the number of partitions to construct; then it uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another). These methods work well for finding spherical-shaped clusters in small or medium-sized datasets.

K-means clustering

K-means (James MacQueen, (1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem [16]. It is partitioning clustering method. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters k fixed apriori. Finally this algorithm aims at minimizing the objective function, in this case a squared error function. The objective function is the sum of squared error (SSE),

$$SSE= \sum_{j=1} \sum_{i=1} \| x_i - C_j \|^2 \quad \text{------ (1)}$$

where $\| x_i - C_j \|^2$ is a chosen distance measure between a data point x I and the cluster centre Cj is an indicator of the distance of the n data points from their respective cluster centres.

This algorithm does not necessarily find the most optimal configuration corresponding to the global objective function minimum. The k means algorithm can be run multiple times to reduce this effect. The storage required is O ((m+K) n), where m is the number of points and n is the number of attributes. The time required is O (I*K*m*n), where I is the number of iterations required for convergence.

It is very simple and easy to implement and terminates at local optimum. There are some disadvantages of this method. It is applicable only when mean is defined. Significantly sensitive to the initial randomly selected cluster centres. Need to specify k as the number of clusters. This method is unable to handle noisy data and outliers. It is suitable for only spherical shape clusters.

Partitioning Around Medoids (PAM) Algorithm

In PAM method each cluster is represented by one of the objects located near the centre of the cluster. The algorithm begins by selecting an object as medoid for each of n clusters, and then each of the non selected objects is grouped with the medoid to which it is the most similar. The algorithm swaps the medoids with other non selected objects until all objects qualify as medoid.

Clustering LARge Application (CLARA)

CLARA method was designed by Kaufman, L. and Rousseeuw, P.J. based on samples. It is the implementation of the PAM algorithm in a subset of the dataset. Instead of finding representative objects for the entire data set, it takes different samples of the dataset, applies PAM on the samples, and finds the medoids of the sample. This algorithm can handle large datasets than PAM algorithm.

Clustering Large Application based on RANdomised Search (CLARANS)

Raymond t.Ng and Jiawei Han proposed a new clustering method CLARANS whose aim is to identify spatial structures that may present in data. Searching a graph where every node is a potential solution, that is, a set of k medoids. It selects the node and compares it to a user-defined number of their neighbours searching for a local minimum and moves to the neighbour's node. If the local optimum is found, it starts a new randomly selected note in search for a new local optimum.

Density-Based Methods

In density-based clustering, [9] clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu (1996) proposed a density-based algorithm for discovering clusters in large spatial databases with noise. These methods group neighbouring objects into clusters based on density conditions. It either grows closer according to the density of neighbourhood or according to some density function.

DENsity-based CLUstEring (DENCLUE)

DENCLUE is a kind of generalized density-based clustering method, which has many outstanding properties, such as capability of discovering clusters of arbitrary shapes, good scalability for the size of databases, high dimensions and large amounts of noisy data, etc.,. But the quality of clustering results strongly depends on the careful choice of two parameters: density and noise.

It has capability to clustering datasets with large amount of noise. It allows arbitrary shape clusters in high dimensional dataset. It depends on density parameter and noise threshold. These selections of parameters decide the quality of clusters.

Density-Based Spatial Clustering of Application with Noise (DBSCAN)

It is a partitional type of clustering algorithm. For each point in the cluster, the neighbourhood of a given radius has to contain at least a minimum number of data points. Data points are classified as core, border and Noise. Core points lie in the interior of density based clusters and should lie within specified radius or threshold value, minimum no of points which are user specified parameters. Border point lies within the neighbourhood of core point and many core points may share same border point. The point which is neither a core point nor a border point is a noise

The space requirement of DBSCAN, even for high-dimensional data, is O (m) because it is only necessary to keep a small amount of data for each point, i.e., the cluster label and the identification of each point as a core, border, or noise point. The time complexity of this algorithm is O (n.log n) [20].

This algorithm can handle noise and discover the clusters of arbitrary shape. This method has a problem of detecting meaningful clusters in data of varying density.

Ordering Points To Identify the Clustering Structure (OPTICS).

This algorithm was designed by Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jorg Sander in [1999]. It is similar to density based clustering. This algorithm computes an augmented clustering ordering for automatic and interactive cluster analysis. The time complexity of this algorithm is O (n.log n).

2.2 Grid-based Methods

The methods quantize the space into a finite number of the cells that form a grid structure, and then perform clustering on the grid structure. The time and space complexity of defining the grid, assigning each object to a cell, and computing the density of each cell is only O (m), where m is the number of points. STING, CLIQUE (CLustering In QUEst), Wave cluster, O-CLUSTER (Orthogonal partitioning CLUSTERing), MAFIA (Merging of Adaptive Intervals Approach to Spatial Data Mining), Axis Shifted Grid Clustering Algorithm (ASGC), Adaptive Mesh Refinement (AMR) and BANG are examples for this algorithm. Grid-based methods are fast and handle outliers efficiently.

Statistical Information Grid-based method (STING)

STING is a grid based multi resolution clustering technique in which the spatial area is divided into rectangular cells (using latitude and longitude) and employs a hierarchical structure [24]. The time complexity of this algorithm is O (k) where k is number of cells in the bottom of the layer.

Wave cluster

Gholamhosein Sheikholeslami Surojit Chatterjee, Aidong Zhang [1998], propose WaveCluster, a novel clustering approach based on wavelet transforms, which satisfies all the above requirements. Using multi- resolution property of wavelet transforms, we can effectively identify arbitrary shape clusters at different degrees of accuracy. The author also demonstrate that WaveCluster is highly efficient in terms of time complexity, O (n) where n is number data points.

2.3 Model-based Methods

A model-based approach consists in using certain models for clusters and attempting to optimize the fit between the data and the model.

Expectation Maximization (EM) Method

The Expectation Maximization (EM) method is based on the assumption that the objects in the dataset have attributes whose values are distributed according to some linear combination (or mixture) of simple probability distributions. This method assigns objects to different clusters with certain probabilities in an attempt to maximize expectation (or likelihood) of assignment. The EM method consists of two step iterative algorithm. The first step is estimation step or E-step. The second step is Maximization step or M-step, which maximize the likelihood function.

The Self-Organizing Maps (SOM)

The Kohonen Self-Organizing Feature Map (SOFM or SOM) is a clustering and data visualization technique based on neural network viewpoint [20]. The goal of SOM is to find a set of centroids and to assign each data point in the data set to the centroid that provides the best approximation of that data point. It imposes a topographic (spatial) organization on the centroids (neurons).

Juha Vesanto and Esa Alhoniemi proposed Clustering of the Self-Organizing Map [2000]. Different approaches to clustering of the SOM are considered. The two-stage procedure—first using SOM to produce the prototypes that are then clustered in the second stage—is found to perform well when compared with direct clustering of the data and to reduce the computation time. The experiments indicated that clustering the SOM instead of directly clustering the data is computationally effective approach. The clustering results using SOM as an intermediate step were also comparable with the results obtained directly from the data.

The user must choose the settings of parameters, neighbourhood function, grid type, and the number of centroids. SOM is not guaranteed to converge and lacks a specific objective function.

2.4 Hierarchical Methods

These methods proceed iteratively by either merging the smaller clusters into larger ones, or by splitting the larger clusters. The result of the algorithm is a tree of clusters, called dendrogram. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [3]. The agglomerative (AGNES) method is bottom-up and starts with each object forming the separate group. The divisive analysis method (DIANA) is top-down and starts with all the objects in the same cluster, based on how the hierarchical decomposition is formed.

This method requires distance measure between clusters to be computed. Some of the distance measure called linkage metrics are single-link, complete-link, centroid and average-link. AGNES, DIANA, CHAMELEON, BIRCH, CURE and ROCK are examples for hierarchical clustering Algorithms

These methods are able to handle any distance measure or similarity. It Provides clusters at different levels of granularity. Time complexity of hierarchical method is O (m3).

AGglomerative NESting method (AGNES)

Agglomerative nesting method (AGNES) is an iterative method starts with n clusters for m data points, that is, each cluster containing of a single data point. In each step the method merges two nearest clusters, thus reducing the number of clusters and building larger clusters using distance measure. The process continues until the number of clusters has been reached or all the data points are in one cluster.

Divisive ANAlysis Method (DIANA)

Divisive analysis method is an iterative method. Initially there is one large cluster consisting of all n objects. At each subsequent step, the largest available cluster is split into two clusters until each cluster has only one data point or any other threshold has been reached. Thus, the hierarchy is built in n-1 steps.

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

Zhang et al. [1996], work had a significant impact on overall direction of scalability research in clustering. It uses a hierarchical data structure called CF-tree for partitioning the incoming data points in an incremental and dynamic way. CF-tree is a height-balanced tree, which stores the clustering features and it is based on two parameters: branching factor B and threshold /t, which referred to diameter of the cluster of each cluster must be less than T.

Clustering Using REpresentative (CURE)

Guha et al. [1998] introduced the hierarchical agglomerative clustering algorithm CURE (Clustering Using REpresentatives).It represents each cluster by a certain number of points that are generated by selecting well scattered points and then shrinking them towards the cluster centroid by specified fraction. It uses combination of random sampling and partition clustering to handle large databases. Time complexity of CURE is O (m2).

Robust Clustering Algorithms (ROCK)

S. Guha, R. Rastogi, and K. Shim. introduces ROCK: A Robust Clustering Algorithm for Categorical Attributes and propose a novel concept of links to measure the similarity/proximity between a pair of data points.. ROCK that employs links and not distances when merging clusters[22]. This method naturally extend to non-metric similarity measures that are relevant in situations where a domain expert/similarity table is the only source of knowledge.

2.5 Other Clustering Techniques

There are other clustering algorithms are also proposed in the literature. Here constraint based clustering and Genetic Based algorithms are discussed.

Constraint Based Clustering

Finding clusters that satisfy user specified constraints is highly desirable in many applications[1].A scalable constraint starts by finding initial solution that satisfies user specified constraints then refines the solution by performing confined object movements under constraints.

Evolutionary Algorithms

There are many evolutionary algorithms like Simulated Annealing (SA) and Genetic algorithms (GA)) for clustering has been proposed in the literature. These algorithms are based on the optimization of some objective function that guides the evolutionary search.

Sheikh, R.H., Raghuwanshi, M.M.; Jaiswal, A.N. discussed survey of GA based clustering. The capability of GAs is applied to evolve the proper number of clusters and to provide appropriate clustering. The author presents some existing GA based clustering algorithms and their application to different problems and domains.

Initializing K-means using Genetic algorithm [7] overcomes the initializing problem of the K-means algorithm by using the genetic algorithm. This approach solves the blind search problem of the K-means algorithm. The author tested the algorithm using four different dataset chosen from MATLAB. The algorithms used for comparison were K-means, Genetic algorithm and Genetic algorithm initializing K-means (GAIK). The trade-off between average error rate and average time between these algorithms were listed by the author. Genetic algorithms are difficult to understand. Determining the best fitness function is also difficult.

CONCLUSION

In this paper survey of various clustering algorithms are discussed with their merits and demerits. The time and space complexity of these algorithms are also discussed. These clustering algorithms are used based on the size, shapes, data types and their application. There are also some variants in these algorithms with better efficiency.

REFERENCES

[1]Anthony K.H. Tung, Jiawei Han, Laks V.S.Lakshmanan and Raymond T. Ng, Constraint- Based clustering in Large Databases, In: ICDT (2001), p. 405-419.

[2]A. K. Jain and R. C. Dubes (1988), Algorithms for clustering data. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

[3]JAIN, A. and DUBES, R. (1988), Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

[4]Bezdek, James C. (1981), Pattern Recognition with Fuzzy Objective Function Algorithms, ISBN 0-306-40671-3.

[5]B Al-Shboul, SH Myaeng, (2009), Initializing K-Means using Genetic Algorithms. World Academy of Science, Engineering and Technology.

[6]Gholamhosein Sheikholeslami Surojit Chatterjee, Aidong Zhang, WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases, VLDB '98 Proceedings of the 24rd International Conference on Very Large Data Bases Pages 428-439 ,Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©1998.

[7]GOLDBERG, D. (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.

[8]GUHA, S., RASTOGI, R., and SHIM, K. (1998). CURE: An efficient clustering algorithm for large databases. In Proceedings of the ACM SIGMOD Conference, 73-84, Seattle, WA.

[9]Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30.

[10]Jiawei Han, Micheline Kamber (2001), “Data Mining concepts and Techniques”, Morgan Kaufmann Publishers, San Francisco, CA, USA.

[11]Juha Vesanto and Esa Alhoniemi, Clustering of the Self-Organizing Map, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 11, NO. 3, MAY 2000, pp. 586-600.

[12]KAUFMAN, L. and ROUSSEEUW, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York, NY.

[13]Kaufman, L. and Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the –Norm and Related Methods, edited by Y. Dodge, North-Holland, pp.405–416.

[14]M. Livny, R.Ramakrishnan, T. Zhang, (1996). BIRCH: An Efficient Clustering Method for Very Large Databases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery: pp.103-114.

[15]Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu (1996) "A density-based algorithm for discovering clusters in large spatial databases with noise". Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.

[16]MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

[17]MCLACHLAN, G. and BASFORD, K. (1988). Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York, NY.

[18]Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jorg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[19]P.Prabhu and N.Anbazhagan, (2011), Improving the performance of k-means clustering for high dimensional dataset, International Journal of Computer Science and Engineering”, Vol 3. No.6. pp. 2317-2322.

[20]Pang-Ning Tan, Vipin Kumar, Michael Steinbach (2006), Introduction to Data Mining, Pearson Education Inc,.

[21]Raymond GT.Ng and Jiawei Han (2002) CLARANS: A method for clustering objects for spatial mining. IEEE Trans. Knowl. Data Eng. 14(5): pp.1003-1016.

[22]S. Guha, R. Rastogi, and K. Shim, (2000). ROCK: A Robust Clustering Algorithm for Categorical Attributes. Information Systems, vol. 25, no. 5:pp. 345-366.

[23]Sheikh, R.H., Raghuwanshi, M.M.; Jaiswal, A.N., Genetic Algorithm Based Clustering: A Survey, Emerging Trends in Engineering and Technology, 2008. ICETET '08.pp.314- 319.

[24]Wei Wang, Jiong Yang, and Richard Muntz: STING: A Statistical Grid Approach to Spatial Data Mining: Department of Computer Science, University of California, Los Angels.

[25]Zhang, T., Ramakrishnan, R. and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. In Proceedings of the ACM SIGMOD Conference, pp.103-114, Montreal, Canada.

[26]Zhang, T., Ramakrishnan R. and Livny M. (1997).BIRCH: A new data clustering algorithm and its applications. Journal of Data Mining and Knowledge Discovery, 1, 2,141-182.