International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Survey on Various Clustering Techniques in Data Mining

Lavanya .N[1], N. Deepika[2]

[1]New Horizon College of Engineering, Bangalore, India

Abstract: ***Clustering is the run through of grouping the data into classes, so that objects within a cluster are similar to one another but these objects are different to the objects that are in other clusters. Differences and likeness are refereed on the attribute values telling the objects and often involves in measuring distance. This paper provides an review about few clustering methods: Partitioning method, hierarchical method, Density based method, Grid based method, and Model based method in data mining.***

Keywords: Clustering, Partitioning, Hierarchical, Density Based, Grid based, Model based.

1. Introduction

Clustering is a technique of combining objects or data into clusters in which objects within the cluster have high communication, but are very different to objects in the other clusters. Differences and likeness are stately on the attribute values which describes the objects. Clustering methods are used to convey and categorize the data, for data looseness and model creation, for identifying of outliers etc. Common style of all clustering methods is to find clusters center which represent each cluster. Based on the likeness metric and input direction cluster center helps in shaping which cluster is nearest or most similar one.

Clustering can be rash the most important unproven ignorance problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unnamed data.

2. Few Requirements of Clustering

The following are typical requirements of clustering in data mining:
1) Scalability
2) Ability to deal with different types of attributes
3) Discovery of clusters with arbitrary shape
4) Requirements for domain knowledge to determine input parameters
5) Ability to deal with noisy data
6) Incremental clustering and insensitivity to input order

3. Clustering Methods

3.1 Partitioning Method

Trendy partitional clustering method clustering generates the clusters in a single step as an alternative of generating quite a few steps. Only one usual of clusters is made at the end of clustering , although numerous groups of clusters may be created internally. The best generally used partitioning methods are k-means, k- medoids. [6]

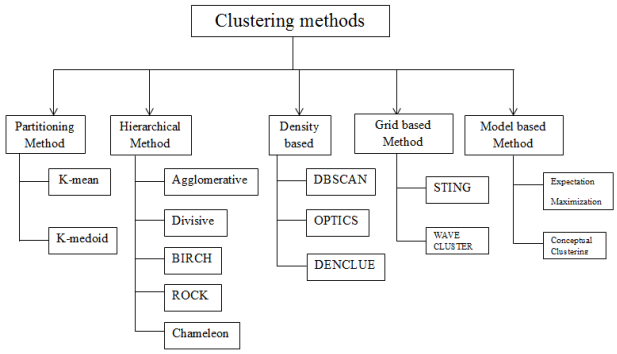

Figure 1: Classification of clustering methods [1]

3.1.1. k-means method: centroid based method

The K-means algorithm allocates every single point to the cluster whose center also called centroid is nearest. The center is the typical of all the points in the cluster that is, it arranges are the arithmetic mean for each measurement separately over all the points in the cluster.

Algorithm k-mean:
Input:
C: the number of cluster.
D: a data set containing m objects.

Output:
A set of C cluster.

Method:
1) Choose m objects unsystematically from dataset as the initial cluster centers;
2) Grounded on the mean value of the object which is similar to cluster re assign object to that cluster.
3) Educated guess the mean value of the objects for each cluster and make percentage increase until no updation made.[1]

The k-means algorithm has the following important material goods:
1. It is efficient in handling large data sets.
2. It often ends at a local optimum.
3. It works only on numbers.
4. The clusters have coiled shapes. [9]

3.1.2 k-medoid method

The simple line of attack of k-medoids algorithm is each cluster is denoted by single of the objects to be found near the center of the cluster. Partitioning around Medoids was one-of-a-kind of the leading k-medoids set of rules is said.

Algorithm k-medoid:

Inputs:
C: the total of clusters,
D: a data set covering m objects.

Output:

A stable of C clusters.

Method:

1) Choose m objects with composure in D as the original common objects;
2) Then each leftover object is assigned to the cluster which have nearest classic object,
3) Then informally select a non representative object.
4) Work out total cost for varying the illustrative object with non-representative object.
5) If Total cost is a smaller amount of than zero then change representative object with non-representative object to make a new set of m representative objects.

Pros and Cons of Partitioning Clustering:

It is calm to appliance and by k-mean set of rules. Removal monotonically falls G since each vector is given to the next centroid and disadvantage of this algorithm is whenever a point is close to the center of another cluster, then it gives lowly result due to casing of data points. [6]

3.2 Hierarchical Method

Is a method of cluster examination which treasure trove to build a pecking order of clusters. The value of a clean hierarchical clustering technique undergoes from its letdown to implement modification, once a merge or split decision has been completed. In overall there are two types of hierarchical method:

- Agglomerative approach
- Divisive approach [6]

3.2.1 Agglomerative Method

This method uses a bottom-up approach. It in general jumps by agree to each object form its own cluster and iteratively merges clusters into larger and larger clusters, until all the objects are in a on its own cluster or certain conclusion state of affairs are satisfied.

Algorithm:

Input:

J: Fixed number of objects
M: Matrix showing distance between objects

Output:

DG // Dendrogram

Method:

Step 1: Set J=0;
Step 2: Set p=n;
Step 3: P={{t1},…,{tn}};
Step 4: DG={<j, p ,P>};//originally dendrogram contains each object in its own cluster.
Step 5: Repeat
Step 6: Old k=k;
Step 7: j = j+1;
Step 8: M=vertex adjacency matrix for graph with threshold distance of d;
Step 9: <p ,P>= New Cluster(M , J);
Step 10: If old p =P then
Step 11: DG=DG U <j, p , P>; //new set of clusters added to dendrogram.
Step 12: Until k=1. [1]

3.2.2 Divisive Method

A divisive hierarchical clustering method works a top-down strategy. It starts by placing all objects in one cluster, which is the order's root. It then divides the root cluster into several smaller sub-clusters, and recursively partitions those clusters into smaller ones.

Top-down clustering is in theory more composite than bottom-up clustering in the meantime we need a second, flat clustering algorithm as a "subroutine". It has the advantage of being more efficient if we do not generate a complete order all the way down to individual document leaves.

3.2.3 BRICH Method

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is intended for clustering a large amount of numeric data by integrating this clustering and other clustering ways and means such as iterative partitioning.

It stops the two worries in agglomerative clustering methods: (1) scalability and (2) the disaster to undo what was done in the prior step. BIRCH uses the designs of clustering feature to review a cluster, and clustering feature tree (CF-tree) to denote a cluster hierarchy.

Algorithm:

Input:

N=set of elements;
T=threshold for CF tree construction;

Output:

C //set of clusters

Method:

Step 1: for each element that belongs to N Find correct leaf node for element insertion;
Step 2: if threshold condition is not violated than add element to cluster and update CF ;
Step 3: else make room to insert element then insert element as single cluster or update CF;
Step 4: else break leaf node and redistribute CF.

3.2.4 Chemeleon Method

It is an agglomerative hierarchical clustering established of rules that uses active modeling. It is a hierarchical method that methods the connection of two cluster based on dynamic model. The integration method using the dynamic model facilitates detection of expected and consistent clusters.

The set of rules scheme generally consist of two phases: first of all separating of data facts are done to system sub-clusters, using a graph partitioning, after that have to do frequently integration of sub clusters that come from prior step to obtain final clusters.

3.2.5 ROCK Method

Robust Clustering via associations is a robust agglomerative hierarchical-clustering set of rules created on the view of links. It is also applicable for management bulky data sets. For integration data points, this works associations between data points not the space between them.

ROCK method is divided into three parts are as follows:

1) First get a random sample of the data.
2) Get hold of the goodness extent by carrying out link agglomerative approach on data to get the fact which can be compound at each step. 3. Assigned the remaining data on disk by using these points which forms the clusters.

Pros and Cons of Hierarchical Clustering:

The main advantage of hierarchical clustering is it has no a-priori information about the number of clusters required and it is easy to implement and gives best result in some cases. The cons of the method that the set of rules cannot once loosen what was done formerly, no neutral function is in a straight line minimized and now and then it is difficult to recognize the accurate number of clusters by the dendrogram.[6]

3.3 Density Based Method

Density based clustering algorithm try to find clusters based on density of data points in s region. The basic hint of density based clustering is that for every case in point of cluster the locality of a given radius (Eps) has to have at least minimum number of instances (MinPts). One of the most well-known density based clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise). [1]

3.3.1 DBSCAN Method:

Density based clustering method based on connected regions is density based clustering method for handling spatial data with noise in application or database. It customs the in elevation density region for manufacture the cluster , and the extra region which have low density are kept outside the cluster by marks as outlier. There is no need to define the number of clusters in advanced.

“Density reachability” and “Density connectability” are the two concepts which are used during making the cluster which in turn have asymmetric and symmetric relation. “Minpt” and “e” are the two parameters ,if point k contains more “Minp”t than the e-neighborhood then a new cluster with core object will be created, then the DBSCAN will collect the density within reach object from these core objects.

This algorithm needs three input parameters:

- k, the neighbour list size;
- Eps, the radius that delimitate the neighbourhood area of a point (Eps neighbourhood);
- MinPts, the minimum number of points that must exist in the Eps-neighbourhood.

Algorithmic steps for DBSCAN Clustering

Let X = {x1, x2, x3, ..., xn} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

Step 1: Start with an arbitrary starting point that has not been visited.
Step 2: Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
Step 3: If there are sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise.
Step 4: If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points.
Step 5: A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
Step 6: This process continues until all points are marked as visited.

3.3.2 OPTICS Method

Ordering point to identify the clustering structure creates the liner ordering of objects in the database. Like the DBSCAN it use two parameter “e “and “Minpt” where e describe the maximum distance and “Minpt” describe the number of points or objects essential to make a cluster. Core distance and Reachability distance are needed define to ordering of objects in to the database.

The OPTICS algorithm finds clusters using the following steps:

Step 1: Generate an gathering of the objects in a database, storing the core-distance and an appropriate reachability-distance for each object. Clusters with uppermost density will be finished first
Step 2: Based on the ordering information produced by OPTICS, use another algorithm to extract clusters.
Step 3: Extract density-based clusters with respect to any distance e’ that is smaller than the distance e used in generating the order.

Pros and Cons of Density-Based Algorithm

The main advantage density-based clustering Algorithm does not require a-priori specification and able to identify noisy data while clustering. It fails in case of neck type of dataset and it does not work well in case of high dimensionality data. [6]

3.4 Grid Based Method

This method takes a space-driven methodology by subdividing the set in space into cells self-governing of the supply of the input objects. The grid-based clustering methodology routines a multi-resolution grid data structure. It quantizes the object space into a finite number of cells that system a grid structure on which all of the operations for clustering are performed. [6]

3.4.1 STING Method

STING is a grid-based multi-resolution clustering skill in which the drive in spatial area of the input objects is distributed into rectangular cells. The space can be divided in a hierarchical and recursive way. Several levels of such rectangular cells correspond to different levels of resolution and form a hierarchical structure:

Algorithm

Step 1: Determine a layer to begin with.
Step 2: For each cell of this layer, we calculate the confidence interval (or estimated range) of probability that this cell is relevant to the query.
Step 3: From the interval calculated above, we label the cell as relevant or not relevant.
Step 4: If this layer is the bottom layer, go to Step 6; otherwise, go to Step 5.
Step 5: We go down the hierarchy structure by one level. Go to Step 2 for those cells that form the relevant cells of the higher level layer.
Step 6: If the specification of the query is met, go to Step 8; otherwise, go to Step 7.
Step 7: Retrieve those data fall into the relevant cells and do further processing. Return the result that meets the requirement of the query.
Step 8: Find the regions of relevant cells. Return those regions that meet the requirement of the query.

3.4.2 CLIQUE Method

It is a simple grid-based method for discovery density based clusters in subspaces. CLIQUE dividers each dimension into non covering intervals, thus segregating the whole set in space of the data objects into cells. It routines a density threshold to recognize dense cells and spare ones. A cell is dense if the number of objects planned to it goes beyond the density threshold

Algorithm

Step 1: Based on m, the input feature space is split.
Step 2: The input is quantized to a particular grid.
Step 3: Initialize count (of elements) to 0 across all the grids in the feature space.
Step 4: count++
Step 5: For every attribute activate the regions of high density
Step 6 : Now take two attributes at a time and check for dense regions in the intersection of the dense regions in the individual attribute.
Step 7: Repeat the step 6 by adding one dimension with each iteration, and choosing all their possible combinations, until all the dimensions (attributes) in the data set are covered.
Step 8 : Label all the connected clusters with a label value.[1]

Pros and Cons of Grid-Based Algorithm

The main advantage is its processing time, which is typically independent of the number of objects. It fails when some potential clusters will be lost in the density-units.[6]

3.5 Model Based Method

Model-based clustering methods are based on the assumption that data are generated by a mixture of underlying probability distributions, and they optimize the fit between the data and some mathematical model. When facing an unknown data distribution, choosing a suitable one from the model based candidates is still a major challenge. On the additional, clustering based on odds suffers from high computational cost, especially when the scale of data is very huge. [6]

3.5.1 Expectation Maximization Method

EM is the most preferred iterative refinement method that is used to figure out the parameter estimates. Each cluster is defined by parametric probability distribution. Objects are assigned to cluster according to their mean value with some weight associated with objects. EM start with initial assumption of the parameter vector which is randomly chosen on the basis of clusters mean value and then the expectation step and maximization step are applied for the distribution of the given data. EM is simple and easy to implement.

Algorithm:

Step 1: Estimate the distribution of X, in sample space X, but we can only observe X indirectly through Y, in sample space Y.
Step 2: In many cases, there is a mapping $x \rightarrow (x)$ from X to Y, and x is only known to lie in a subset of X, denoted by X(y), which is determined by the equation y = y(x).
Step 3: The distribution of X is parameterized by a family of distributions f (x | θ), with parameters $\theta \in \Omega$ on x. The distribution of y, g(y | θ) is
$g(y \mid \theta) = \int_{x(y)} f(\mathbf{x} \mid \theta)\, dx$
Step 4: The EM algorithm aims at finding a _ that maximizes g(y | θ) given an observed y.
Step 5: Introduce the function

$$Q(\theta' \mid \theta) = E(\log f(x \mid \theta') \mid y, \theta);$$

that is, the expected value of log f (x | θ') according to the conditional distribution of x given y and parameter θ. The expectation is assumed to exist for all pairs (θ′, θ). In particular, it is assumed that f (x | θ) > 0 for $\theta \in \Omega$.[1]

3.5.2 Conceptual Clustering Method

Conceptual method is a unsupervised machine learning method for the classification of unknown classification. Concept based structure is used to separate the generated classes from the ordinary data. This concept based method is similar to decision tree in which a hierarchy is generated. The most well-known conceptual clustering method is COWEB method.

Algorithm:

INPUT:

N=Node; **I** = New instance

Method:
Train(N,I) = IF leaf(N)
THEN create_sub-tree(N,I)
ELSE
Incorporate(I,N); Update N's probabilities
//Compute score of placing I in each child of N
N1 = child with highest score = HIGH
N2 = child with second highest score
NEW = score when placing I as a new child of N
MERGE = score of merging N1 and N2 … and putting I in merged node
SPLIT = score of splitting N1 into its children
//IF highest score is:
HIGH: _ Train(N1,I)
NEW: _ Add I as a new child of N
MERGE: _ Train(merge(N1,N2,N),I)
SPLIT: _ Train(split(N1,N),I)

4. Conclusion

Clustering is present at the heart of data analysis and data mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties data interrelationships variation. Clustering is a imaginative method. The way out is not exclusive and it intensely be subject to upon the analyst's choices. Clustering is a unsupervised learning method which makes the cluster of objects or documents according to their similarity and dissimilarity bases. Objects which exhibits the same feature are placed into one cluster and those which are not similar are placed into other cluster. Clustering can be done by the various procedures such as hierarchical- based, partitioning-based, grid-based, density-based algorithms and model based algorithms.

References

[1] Kavita Nagar, "Data Mining Clustering Methods: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.
[2] Yogita Rani and Dr. Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal of Information and Computation Technology.ISSN 0974-2239 Volume 3, Number 11 (2013).
[3] L.V. Bijuraj, "Clustering and its Applications", Proceedings of National Conference on New Horizons in IT - NCNHIT 2013.
[4] T. Soni Madhulatha, "AN OVERVIEW ON CLUSTERING METHODS", IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725.
[5] Bipin Nair B J, "Generating Recurrent Patterns Using Clique Algorithm", International Journal of Software and Web Sciences (IJSWS).
[6] Brinda Gondaliya, "REVIEW PAPER ON CLUSTERING TECHNIQUES", International Journal of Engineering Technology, Management and Applied Sciences.
[7] Saroj, Tripti Chaudhary, "Study on Various Clustering Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3031-3033.
[8] S.R.Pande 1, Ms. S.S.Sambare 2, V.M.Thakre, "Data Clustering Using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 8, October 2012.
[9] Amandeep Kaur Mann, Navneet Kaur, "Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.
[10] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann, Third edition-2012.
[11] N.Deepika, Dr.N.Guruprasad, "Empirical Study of Combinatorial Learning Method For Data Clustering", International Journal of Informative and Futuristic Research, Volume 3, Issue 4, Dec-2015.

Author Profile

Lavanya N received B.E degree in Information science and Engineering from Cambridge Institute of Technology in 2016. Currently pursuing M.tech degree in Computer Science and Engineering.Her area of interests includes Data Mining,Software Engineering.

N.Deepika Sr. Asst. Professor having 15 years of experience in Academics has pursued her M.Tech from JNTU,Hyderabad and B.Tech from SVU.She is currently working In NHCE, Dept of CSE, Bangalore. She has guided many UG & PG students for their Projects. Her Research areas include Clustering techniques, Data Mining, Web Mining and Big Data Analysis.