International Journal of Multidisciplinary Research and Development

International Journal of Multidisciplinary Research and Development

Volume :2, Issue :5, 32-36
May 2015
www.allsubjectjournal.com
e-ISSN: 2349-4182
p-ISSN: 2349-5979
Impact Factor: 3.762

R.Tamilselvi
M.phil Research Scholar
Department of computer science, Vivekanandha college For Women, Unjanai, Tiruchengode, India

B.Sivasakthi
M.phil Research Scholar
Department of computer science, Vivekanandha college For Women, Unjanai, Tiruchengode, India

R.Kavitha
Assistant Professor
Department of computer science, Vivekanandha college For Women, Unjanai, Tiruchengode, India

Correspondence:
R.Tamilselvi
M.phil Research Scholar
Department of computer science, Vivekanandha college For Women, Unjanai, Tiruchengode, India

A comparison of various clustering methods and algorithms in data mining

R.Tamilselvi, B.Sivasakthi, R.Kavitha

Abstract
Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods.

Keywords: Hierarchical Methods, Partitional Methods, Density-based Methods, Grid-based Methods

1. Introduction
The goal of this survey is to provide a comprehensive review of different clustering techniques in data mining. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists
of objects that are similar between themselves and dissimilar to objects of other groups. Clustering methods as an optimization problem try to find the approximate or local optimum solution. An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. Common approach for all clustering techniques is to find clusters centre that will represent each cluster. Cluster centre will represent with input vector can tell which cluster this vector belong to by measuring a similarity metric between input vector and all cluster centre and determining which cluster is nearest or most similar one .Further data set can be numeric or categorical. Inherent geometric properties of numeric data can be exploited to naturally define distance function between data points. Whereas categorical data can be derived from either quantitative or qualitative data where observations are directly observed from counts.

2. Various Data Clustering Methods
There are many clustering methods available and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the data set. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters with or without overlap. These methods are sometimes divided into partitioning methods in which the classes are mutually exclusive and the less common clumping method in which overlap is allowed.

Partitioning Methods
The partitioning methods generally result in set of M clusters; each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster, this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroids may be required in

other cases,eg., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of clusters is large, the centroids cab be further clustered to produces hierarchy with in a dataset

Hierarchical Agglomerative Methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by two closest objects and merge them into a cluster and also find and merge the next two closest points where appoint is either an individual objects or cluster of objects. Individual methods are characterized by the definition used for identification of the closest pair of points and by the means used to describe the new cluster when two clusters are merged here are some general approaches to implementation of this algorithm, these being stored matrix and stored data are discussed. In the second matrix approach, an NXN matrix containing all pair wise distance values is first created and updated as new clusters are formed. This approach has at least an O(nxn) time requirement, rising to $O(n^3)$ if a simply serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomerative, a serious limitations for large N.

Single Link Method (SLINK)

The single Link Method is probably the best known of the hierarchical methods and operates by joining at each step, the two most similar objects which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

Complete Link Method (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity(so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster).This method is characterized by small, tightly bound clusters.

Group Average Method

The Group Average Method relies on the average value of the pair wise within a cluster rather than the maximum or minimum similarity as with the single linkage or the complete link methods. Since, all objects in a cluster contribute to the inter-cluster similarity; each object is on average more like every other member of its own cluster then the objects in any other cluster.

3 .Various Clustering Algorithms

Hierarchical Methods

- Agglomerative algorithms
- Divisive Algorithms

Partitioning Methods

- Relocation Algorithms
- Probabilistic Clustering
- K- Medoids methods
- K-Means Methods

Density-based Algorithms

- Density based connectivity clustering
- Density functions clustering

Grid-based Methods:

- Methods based on co-occurrence of categorical data
- Constraint-based clustering
- Clustering algorithms used in machine learning
- Gradient decent and artificial neural networks

Hierarchical Methods

In data mining, **hierarchical clustering** (also called **hierarchical cluster analysis** or **HCA**) is a method of clusterana lysis which seeks to build a hierarchy of clusters.

Agglomerative algorithm

Hierarchical clustering algorithms are either top-down or bottom-up.Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or *agglomerate*) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called *hierarchical agglomerative clustering* or *HAC*.

```
SIMPLEHAC(d_1, ..., d_N)
 1  for n ← 1 to N
 2  do for i ← 1 to N
 3     do C[n][i] ← SIM(d_n, d_i)
 4     I[n] ← 1  (keeps track of active clusters)
 5  A ← []  (assembles clustering as a sequence of merges)
 6  for k ← 1 to N − 1
 7  do ⟨i, m⟩ ← arg max_{⟨i,m⟩: i≠m ∧ I[i]=1 ∧ I[m]=1} C[i][m]
 8     A.APPEND(⟨i, m⟩)  (store merge)
 9     for j ← 1 to N
10     do C[i][j] ← SIM(i, m, j)
11        C[j][i] ← SIM(i, m, j)
12     I[m] ← 0  (deactivate cluster)
13  return A
```

Divisive Algorithms
Top-down clustering is conceptually more complex than bottom-up clustering since we need a second, flat clustering algorithm as a ``subroutine". It has the advantage of being more efficient if do not generate a complete hierarchy all the way down to individual document leaves. For a fixed number of top levels, using an efficient flat algorithm like K-means, top-down algorithms are linear in the number of documents and clusters.

Partitional Methods
Perhaps the most popular class of clustering algorithms is the combinatorial optimization algorithms and iterative relocation algorithms. These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained. In a basic iterative algorithm, such as K-means- or K-medoids, convergence is local and the globally optimal solution can not be guaranteed. Because the number of data points in any data set is always finite and, thereby, also the number of distinct partitions is finite, the problem of local minima could be avoided by using exhaustive search methods.

Relocation Algorithms
Data partitioning algorithms, which divide data into several subsets. Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. With appropriate data, this results in high quality clusters. Several methods of this type are often categorized as a partitioning cluster method (Non-hierarchical or flat methods). A general iterative relocation algorithm, which provides a baseline for partitioning-based (iterative relocation) clustering methods.

Input: The number of clusters K, and a database containing n objects from some

Output: A set of K clusters, which minimizes a criterion function $\mathcal{J}$.

Step 1. Begin with an initial K centers/distributions as the initial solution.

Step 2. (Re)compute memberships for the data points using the current cluster centers.

Step 3. Update some/all cluster centers/distributions according to new memberships of the data points.

Step 4. Repeat from Step 2. until no change to $\mathcal{J}$ or no data points change cluster.

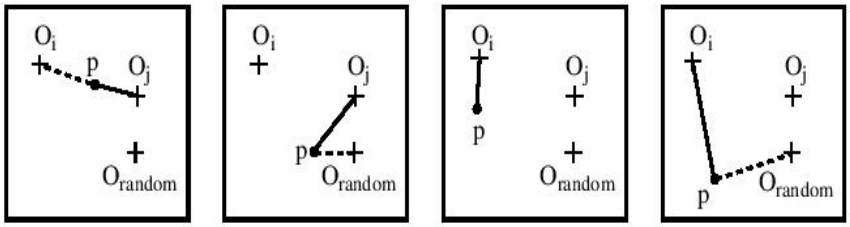
Using this framework, iterative methods compute the estimates for cluster centers, which are rather referred to as prototypes or centroids. The prototypes are meant to be the most representative points for the clusters. The mean and median

Probabilistic Clustering
One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found. More specifically, probabilistic models assume that the data comes from a mixture of several populations whose distributions and priors we want to find. In the *probabilistic approach*, data is considered to be a sample independently drawn from a mixture model of several probability distributions. It associates the cluster with the corresponding distribution's parameters such as mean, variance, etc. Each data point carries not only its (observable) attributes, but also a (hidden) cluster ID (class in pattern recognition). Each point x is assumed to belong to one and only one cluster. Probabilistic Clustering can be modified to handle recodes of complex structure and it can be stopped and resumed with sequence of data. It also results in easily interpretable cluster system.

K- Medoids Methods
The ***k*-medoids algorithm** is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm. Both the *k*-means and *k*-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize squared error, the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the *k*-means algorithm, *k*-medoids chooses data points as centers (medoids or exemplars).*K*-medoids is also a partitioning technique of clustering that clusters the data set of *n* objects into *k* clusters with *k* known *a priori*.

Algorithm

- Case 1: p currently belongs to representative object, o_j. If o_j is replaced by o_{random} as a representative object and p is closest to one of the other representative objects, o_i, $i \neq j$, then p is reassigned to o_i.
- Case 2: p currently belongs to representative object, o_j. If o_j is replaced by o_{random} as a representative object and p is closest to o_{random}, then p is reassigned to o_{random}.
- Case 3: p currently belongs to representative object, o_i, $i \neq j$. If o_j is replaced by o_{random} as a representative object and p is still closest to o_i, then the assignment does not change.



1. Reassigned to O_i 2. Reassigned to O_{random} 3. No change 4. Reassigned to O_{random}

- data object
- \+ cluster center
- — before swapping
- ▪▪▪ after swapping

Algorithm: *k*-**means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) **until** no change;

K-Means Methods

The *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured in regard to the *mean* value of the objects in a cluster, which can be viewed as the cluster's *centroid* or *center of gravity*.

Density-based Algorithms

An open set in the Euclidean space can be divided into a set of its connected components. A cluster, introduced as a connected dense component, grows in any direction that density leads.

Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes. Also this provides a natural defense against outliers.

Density-Based Connectivity clustering

The algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise)targeting low-dimensional spatial data is the major representative in this category. Two input parameters ε and

MinPts are used to define:

1) An ε –neighborhood Nε (x) = {y∈X |d (x, y) ≤ ε } of the point x

2) A core object (a point with a neighborhood consisting of more than MinPtspoints)

3) A concept of a point y density-reachable from a core object x (a finite sequence of core objects between x and y exists such that each next belongs to an ε - neighborhood of its predecessor)

4) A density-connectivity of two points x, y (they should be density-reachable from a common core object). So defined density-connectivity is a symmetric relation and all the points reachable from core objects can be considered as maximal connected components presenting as clusters. The points that are not connected to any core point are outliers (they are not wrapped by any cluster). The non-core points inside a cluster represent its boundary and core objects are internal points. There are any limitations on the dimension or attribute types because processing is out of data ordering.

One problem is by considering two parameters ε and MinPts, there is no straightforward way to fit them to data. Other representative algorithms are GDBSCAN , OPTICS, and DBCLASD.

Density functions clustering

It computes density functions defined over the underlying attribute space instead of computing densities pinned to data points. They introduced the algorithm DENCLUE (DENsity-based CLUstEring). It has a firm mathematical foundation Along with DBCLASD that uses a density function. DENCLUE focus on local maxima of density functions called density-attractors and uses a hill-climbing technique for finding them. It finds center-defined clusters and arbitrary-shape clusters that are defined as continuations along sequences of points whose local densities are more than threshold ξ .Also the algorithm can be considered as a grid-based method and it applied in high dimensional multimedia and molecular biology data.

Grid-based Methods

In the previous section vital concepts of density, connectivity, and boundary were described. Another concept of them is to inherit the topology from the underlying attribute space. To limit the search combinations, multi-rectangular segments are considered. Since some binning is for numerical attributes, methods partitioning space are frequently called grid-based methods. Our attention moved from data to space partitioning. Data partitioning is induced by points' membership in segments resulted from space partitioning, while space partitioning is based on grid-characteristics accumulated from input data. Grid-based clustering techniques are independent of data ordering. In contrast, relocation methods and all incremental algorithms are very sensitive to data ordering. While density-based partitioning methods work best with numerical attributes, grid-based methods work with attributes of different types. BANG-clustering improves the similar hierarchical algorithm GRIDCLUST.Grid-based segments summarize data. The segments are stored in a special BANG structure that is a grid-directory integrating different scales. Adjacent segments are neighbors. Nearest neighbors is a common face has maximum dimension. The density of a segment is a ratio between number of points in it and its volume. From the grid directory, a dendrogram is directly calculated. "The algorithm Wave Cluster works with numerical attributes and has an advanced multi-resolution". Wave Cluster is based on ideas of signal processing. It applies wavelet transforms to filter the data. It has also High quality of clusters, Ability to work well in relatively high dimensional spatial data, and successful handling of outliers.

4. Conclusion

In this study, the basic concept of clustering methods and algorithms are given. The process of grouping a set of physical or a abstract objects into classes of similar objects are named as clustering. These techniques are being used in many areas such as marketing,agriculture,biology and medical. This study concludes that clustering techniques and algorithms become a highly active research area in data mining research.

5. Reference

1. Yuni Xia, Bowei Xi ―Conceptual Clustering Categorical Data with Uncertainty‖ Indiana University – Purdue University Indianapolis Indianapolis, IN 46202, USA
2. Sanjoy Dasgupta ―Performance guarantees for hierarchicalclustering‖Department of Computer Science and Engineering University of California, San Diego
3. A. P. Dempster; N. M. Laird; D. B. Rubin ―Maximum Likelihood from Incomplete Data via the EM Algorithm‖ Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1.(1977), pp.1-38.
4. Fei Shao, Yanjiao Cao ―A New Real-time Clustering Algorithm‖Department of Computer Science and Technology, Chongqing University of Technology Chongqing 400050, China
5. Jinxin Gao, David B. Hitchcock ―James-Stein Shrinkage to Improve K-meansCluster Analysis‖ University of South Carolina,Department of Statistics November 30, 2009
6. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012.