Importance of Clustering in Data Mining

Prof. M. A. Deshmukh
Assistant professor in Computer Science and Engineering
Prof. Ram Meghe Institute of Technology & Research,
Badnera.Maharashtra,India
meghnadeshmukh9@gmail.com

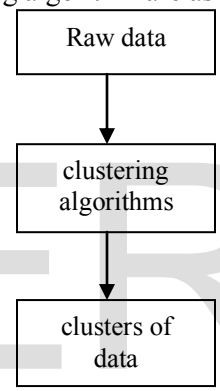
Prof. R. A. Gulhane
Assistant Professor in Computer Science and Engineering
Prof. Ram Meghe Institute of Technology & Research,
Badnera. Maharashtra,India
gulhanerutuja@gmail.com

Abstract— Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the ―better" or more distinct the clustering. Data mining is the process of analysing data from different viewpoints and summerising it into useful information. Data mining is one of the top research areas in recent days. Cluster analysis in data mining is an important research field it has its own unique position in a large number of data analysis and processing.

I. Introduction

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, clustering analysis is done. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is important in data analysis and data mining applications[1]. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. A good clustering algorithm is able to identity clusters irrespective of their shapes. The stages involved in clustering algorithm are as follows,



II. Literature Review

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to

modify data preprocessing and model parameters until the result achieves the desired properties…..

Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goal.

Why clustering?

• Organizing data into clusters shows internal structure of the data
– Ex. Clusty and clustering genes above
• Sometimes the partitioning is the goal
– Ex. Market segmentation
• Prepare for other AI techniques
– Ex. Summarize news (cluster and then find centroid)
• Techniques for clustering is useful in knowledge discovery in data
– Ex. Underlying rules, reoccurring patterns, topics, etc.

Methods:

Basic Agglomerative Hierarchical Clustering Algorithm

1) Compute the proximity graph, if necessary. (Sometimes the proximity graph is all that is available.)
2) Merge the closest (most similar) two clusters.
3) Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
4) Repeat steps 3 and 4 until only a single cluster remains.

The key step of the previous algorithm is the calculation of the proximity between two clusters, and this is where the various agglomerative hierarchical techniques differ.

Any of the cluster proximities that we discuss in this section can be viewed as a choice of different parameters (in the Lance-Williams formula) for the proximity between clusters *Q* and *R*, where *R* is formed by merging clusters *A* and *B*.

$p(R, Q) = \alpha_A p(A, Q) + \alpha_B p(B, Q) + \beta p(A, Q) + \gamma |p(A, Q) – p(B, Q)|$

In words, this formula says that after you merge clusters *A* and *B* to form cluster *R*, then the distance of the new cluster, *R*, to an existing cluster, *Q*, is a linear function of the distances of *Q* from the original clusters *A* and *B*.

Any hierarchical technique that can be phrased in this way does not need the original points, only the proximity matrix, which is updated as clustering occurs.

However, while a general formula is nice, it is often easier to understand the different hierarchical methods by looking directly at the definition of cluster distance that each method uses, and that is the approach that we shall take here. [DJ88] and [KR90] both give a table that describes each method in terms of the Lance-Williams formula

.

Mutual Nearest Neighbor Clustering

Mutual nearest neighbor clustering is described in [GK77]. It is based on the idea of the "mutual neighborhood value (*mnv*)" of two points, which is the sum of the ranks of the two points in each other's sorted nearest-neighbor lists. Two points are then said to be mutual nearest neighbors if they are the closest pair of points with that *mnv*.

Clusters are built up by starting with points as singleton clusters and then merging the closest pair of clusters, where close is defined in terms of the *mnv*. The *mnv* between two clusters is the maximum *mnv* between any pair of points in the combined cluster. If there are ties in *mnv* between pairs of clusters, they are resolved by looking at the original distances between points. Thus, the algorithm for mutual nearest neighbor clustering works in the following way.

a) First the k-nearest neighbors of all points are found. In graph terms this can be regarded as breaking all but the k strongest links from a point to other points in the proximity graph.

b) For each of the k points in a particular point's k-nearest neighbor list, calculate the *mnv* value for the two points. It can happen that a point is in one point's knearest neighbor list, but not vice-versa. In that case, set the *mnv* value to some value larger than 2k.

c) Merge the pair of clusters having the lowest *mnv* (and the lowest distance in case of ties).

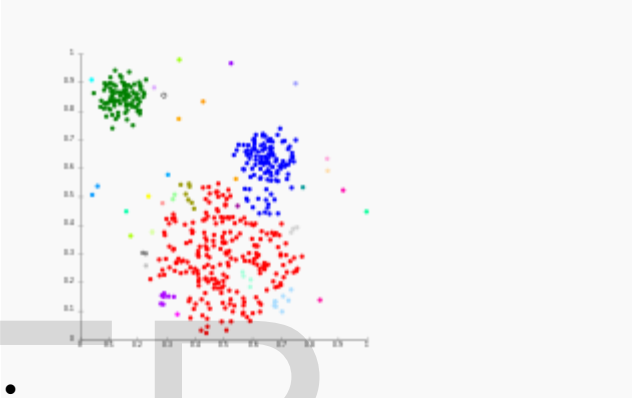
d) Repeat step (c) until the desired number of clusters is reached or until the only
clusters remaining cannot be merged. The latter case will occur when no points in
different clusters are k-nearest neighbors of each other.
The mutual nearest neighbor technique has behavior similar to the shared nearest
neighbor technique in that it can handle clusters of varying density, size, and shape.
However, it is basically hierarchical in nature while the shared nearest neighbor approach
is partitional in nature.

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

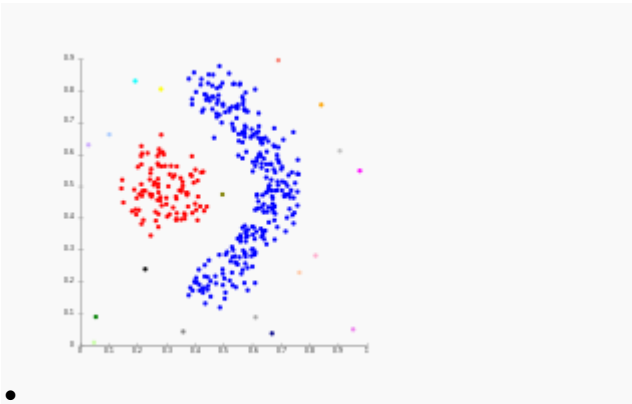
Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance to) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

These methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (known as "chaining phenomenon", in particular with single-linkage clustering). In the general case, the complexity is $\mathcal{O}(n^3)$ for agglomerative clustering and $\mathcal{O}(2^{n-1})$ for divisive clustering, which makes them too slow for large data sets. For some special cases, optimal efficient methods (of complexity $\mathcal{O}(n^2)$) are known: SLINK for single-linkage and CLINK for complete-linkage clustering. In the data mining community these methods are recognized as a theoretical foundation of cluster analysis, but often considered obsolete. They did however provide inspiration for many later methods such as density based clustering.

- Linkage clustering examples



- Single-linkage on Gaussian data. At 35 clusters, the biggest cluster starts fragmenting into smaller parts, while before it was still connected to the second largest due to the single-link effect.



- Single-linkage on density-based clusters. 20 clusters extracted, most of which contain single elements, since linkage clustering does not have a notion of "noise".

General Types of Clusters

1. Well-separated clusters

A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster.

2. Centre-based clusters

A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the ―centre" of a cluster, than to the centre of any other cluster [2]. The centre of a cluster is often a centroid.

3. Contiguous clusters

A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

4. Density-based clusters

A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density.

5. Shared Property or Conceptual Clusters

Finds clusters that share some common property or represent a particular concept.

III. Analysis of Clustering Algorithm

Clustering is the main task of Data Mining and it is done by the number of algorithms. The most commonly used algorithms in Clustering are Hierarchical, Partitioning and Grid based algorithms [1].

1. Hierarchical Algorithms

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster [3].

IV. The Application of Cluster Analysis in Data Mining

The application of cluster analysis in data mining has two main aspects: first, clustering analysis can be used as a pre-processing step for the other algorithms such as features and classification algorithm, and also can be used for further correlation analysis. Second, it can be used as a stand-alone tool in order to get the data distribution, to observe each cluster features, then focus on a specific cluster for some further analysis [5]. Cluster analysis can be available in market segmentation, target customer orientation, performance assessment, biological species etc.

Some Relative Applications

The cluster analysis has been applied to many occasions. For example, in commercial, cluster analysis was used to find the different customer groups, and summarize different customer group characteristics through the buying habits; in biotechnology, cluster analysis was used to categorized animal and plant populations according to population and to obtain the latent structure of knowledge; in geography, clustering can help biologists to determinate the relationship of the different species and different geographical climate; in the banking sector, by using cluster analysis to bank customers to refine a user group; in the insurance industry, according to the type of residence, around the business district, the geographical location, cluster analysis can be used to complete an automatic grouping of regional real estate, to reduce the manpower cost and insurance company industry risk; in the Internet, cluster analysis was used for document classification and information retrieval etc.

V.conclusion

The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different no. of algorithms such as hierarchical, partitioning and grid algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering; the value of k-mean is set. The application of cluster analysis is more and more urgent; the requirements are also getting higher and higher. With the development of modern technology, in the near future, cluster areas will achieve a critical breakthrough.

References

[1] https://en.wikipedia.org/wiki/Cluster_analysis

[2] http://www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf

[3] http://www.academia.edu/7764213/Analysis_and_Application_of_Clustering_Techniques_in_Data_Mining

[4] Yan-Ying Chen, An-Jung Cheng, ―Travel Recommendation by Mining People Attributes and Travel Group Types From Community-Contributed Photos" IEEE Transactions on Multimedia, Vol. 15, No. 6, October 2013.

.