A Review Paper on Clustering in Data Mining

Kuljit Kaur
Department of Computer Engineering
Punjabi University
Patiala
kskuljitkaur59@gmail.com
KanwalPreetSingh Attwal
Department of Computer Engineering
Punjabi University
Patiala
kanwalp78@yahoo.com

Abstract---Clustering is a process of keeping similar data into groups. Objects within the cluster/group have high similarity in comparison to one another but are very dissimilar to objects of other clusters. Clustering is an unsupervised learning technique as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. Types of clustering methods are—hierarchical and partitioning based. In this paper clustering and its methods are discussed.

Keywords---Data mining, clustering, partitioning method, hierarchical clustering, cluster distance.

I. INTRODUCTION

A. Data Mining

It is the process of discovering interesting knowledge from amounts of data stored in databases, data warehouses, or other information repositories [9]. It is an automatic process to find similar objects from a database, is a fundamental operation in data mining [8]. It is a process of keeping similar data into groups [7]. Data mining tools perform data analysis and may uncover important data patterns. Data mining is an essential step in knowledge discovery [3]. Data mining information can be of different types as shown in the below figure and there are different techniques of data mining for different data mining information.

Figure 1: Data mining Information types

One of the technique is discussed below:

B. Clustering

It is a process of grouping observation of similar kinds into smaller groups within the large population. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity [10]. Cluster of data objects may also be considered as a form of data compression [3]. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity [3].
We can identify dense and sparse regions in object space and therefore discover distribution patterns and interesting co-relations among data attributes.

1) Clustering is widely used in
   - Market research
   - Pattern recognition
   - Data analysis
   - Image processing

C. 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.

   Figure 2: well-separated clusters

   2) Center-based clusters: A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the “center” of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid.

   Figure 3: Center-based clusters

   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.

   Figure 4: contiguous clusters

   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.

   Figure 5: Density-based Clusters

   5) Shared Property or Conceptual Clusters: Finds clusters that share some common property or represent a particular concept.

   Figure 6: Conceptual Clusters

D. Typical requirements of clustering in data mining
   - Scalability
   - Ability to deal with different types of attributes
   - Discovery of clusters with arbitrary shape
   - Ability to deal with noisy data
   - Minimal requirements for domain knowledge to determine input parameters
   - High dimensionality:
   - Incremental clustering and insensitivity to the order of input records
   - Interpretability and usability
   - Constraint-based clustering

1) Distance Measures Used
   - Euclidean distance measure
     \[ d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}, \]
   - Manhattan distance measure
     \[ d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|. \]

E. Types of Data in Cluster Analysis
1) **Data Matrix (or object-by-variable structure)**
   - The rows and columns of the data matrix represent different entities.
   - It is often called is often called two-mode matrix.

   \[
   \begin{pmatrix}
   x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\
   \vdots & \ddots & \vdots & \ddots & \vdots \\
   x_{n1} & \cdots & x_{nf} & \cdots & x_{np}
   \end{pmatrix}
   \]

2) **Dissimilarity Matrix: (or object-by-object structure)**
   a. The rows and columns of the dissimilarity matrix represent the same entity.
   b. Dissimilarity matrix is called a one-mode matrix.
   c. Many clustering algorithms operate on a dissimilarity matrix. If the data are presented in the form of a data matrix, it can first be transformed into a dissimilarity matrix before applying such clustering algorithms.[3]

   \[
   \begin{pmatrix}
   0 & d(2,1) & 0 \\
   d(3,1) & d(3,2) & 0 \\
   \vdots & \vdots & \vdots \\
   d(n,1) & d(n,2) & \cdots & 0
   \end{pmatrix}
   \]

F. **Object dissimilarity can be computed for objects described by:**
   - interval-scaled variables
   - by binary variables
   - by categorical, ordinal and ratio-scaled variables
   - or combinations of these variable types[3].

G. **Components of a Clustering Task**

Typical pattern clustering activity involves the following steps:
   - pattern representation (optionally including feature extraction and/or selection),
   - definition of a pattern proximity measure appropriate to the data domain,
   - clustering or grouping,
   - data abstraction (if needed), and
   - Assessment of output (if needed)[4].

![Figure 7: Stages in Clustering](image)

Figure 7 shows a typical sequencing of the first three of these steps, including feedback path where the grouping process output could affect subsequent feature extraction and similarity computations. Various steps are explained below:

1) **Pattern representation:** refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm. Some of this information may not be controllable by the practitioner.

2) **Feature selection** is the process of identifying the most effective subset of the original features to use in clustering.

3) **Feature extraction** is the use of one or more transformations of the input features to produce new salient features. Either or both of these techniques can be used to obtain an appropriate set of features to use in clustering.[4]

4) **Cluster validity:** analysis, by contrast, is the assessment of a clustering procedure’s output. A clustering structure is valid if it cannot reasonably have occurred by chance or as an artifact of a clustering algorithm. When statistical approaches to clustering are used, validation is accomplished by carefully applying statistical methods and testing hypotheses[4].

Three types of validation studies are:
   - **Internal indices:** The internal indices generally evaluate the clusters produced by the clustering algorithm by comparing it with the data only.
   - **External indices:** The external indices evaluate the clustering results by using the prior knowledge, e.g., class labels.
   - **Relative indices:** As the name suggests, this criteria compares the results against various other results produced by the different algorithms [5].
Indices used for this comparison are discussed in detail in Jain and Dubes [1988] and Dubes [1993], and are not discussed further in this paper.

5) Pattern proximity: is usually measured by a distance function defined on pairs of patterns. A simple distance measure like Euclidean distance can often be used to reflect dissimilarity between two patterns, whereas other similarity measures can be used to characterize the conceptual similarity between patterns. [4]

6) Grouping: step can be performed in a number of ways. The output clustering can be hard (partition of the data into groups) or fuzzy (where each pattern has a variable degree of membership in each of the output clusters). Partitioned clustering algorithms identify the partition that optimizes (usually locally) a clustering criterion. [4]

7) Data abstraction: In the clustering context, a typical data abstraction is a compact description of each cluster, usually in terms of cluster prototypes or representative patterns such as the centroid. [4]

H. Advantages of Clustering

- Adaptable to changes.
- Helps single out features that distinguish different groups.
- Unlike in classification, labels are assigned to groups, otherwise assigning class labels to a large number of objects can be a very expensive process. [3]

II. CLUSTERING METHODS

A clustering method is a general strategy applied to solve a clustering problem. Various types of clustering methods are as shown in figure 8 below.

A. Partitioning Methods

As the name suggest, the partitioning methods, in general creates \( k \) partitions/groups of the datasets with \( n \) objects, each partition/group represent a cluster, where \( k \leq n \). It tries to divide the data into subset or partition based on some evaluation criteria. As checking of all possible partition is computationally infeasible, certain greedy heuristics are used in the form of iterative optimization.

1) Relocation based: 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, can be known as a probabilistic models or simply model based clustering. Here, a model assumes that the data comes from a mixture of several populations whose distributions and priors we want to find. The representative algorithms are EM, SNOB, AUTOCLASS and MCLUST. The other approach to partition is based on the objective function, in which the instead of pair-wise computations of the proximity measures, unique cluster representatives are constructed. Depending on how representatives are constructed iterative partitioning algorithms are divided into \( k \)-means and \( k \)-medoids. The partitioning algorithm in which each cluster is represented by the gravity of the centre is known as \( k \)-means algorithms. The one most efficient algorithm proposed under this scheme is named as \( k \)-means only. From the invention of \( k \)-means to till date large number of variations had been proposed, some of them can be listed as, ISODATA, Forgy, bisecting \( k \)-means, \( x \)-means, kernel \( k \)-means and so on. The partitioning algorithm in which cluster is represented by one of the objects located near its centre is called as a \( k \)-medoids. PAM, CLARA and CLARANS are three main algorithms proposed under the \( k \)-mediod method. [5]

2) Grid Based: Such methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the
quantized space. STING is an example of grid-based clustering [10].

3) Subspace clustering: Subspace clustering methods are designed with the aim to work with the high dimensional data. To do so the methods generally make use of the subspace of the actual dimension. The algorithms under this category have taken the idea from the number of other methods and thus fall into number of different categories. The representative algorithms are: CLIQUE, ENCLUS, MAFIA, PROCLUS and ORCLUS[5]

B. K-Means Clustering

It is a partition method, a technique which finds mutual exclusive clusters of spherical shape. It is an iterative method which assigns each point to the cluster whose centroid is the nearest[6]. K-Means algorithm organizes objects into k – partitions, where each partition represents a cluster.

1) K-Means Algorithm Properties
- There are always K clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters [7].

2) Strengths of K-Mean
- Simple - Easy to understand and to implement.
- Efficient: Time complexity: O(tk^n), where k is the number of clusters, n is the number of data points, and t is the number of iterations.
- k-Means is considered a linear algorithm. Since both k and t are small.

3) K-Means Clustering Algorithm
- Choose k cluster centers to coincide with k randomly-chosen patterns or randomly defined points inside the hypervolume containing the patterns set.
- Assign each pattern to the closest cluster center.
- Recompute the cluster centers using the current cluster memberships.
- If a convergence criterion is not met, go to step 2. Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centers, or minimal decrease in squared error.[4]

C. K-Medoids Clustering

Unlike k-means, in the k-medoids or Partitioning around Medoids (PAM) method a cluster is represented by its medoid that is the most centrally located object in the cluster. Medoids are more resistant to outliers and noise compared to centroids[12].

1) K-Medoid Algorithm
- Initialize: randomly select (without replacement) k of the n data points as the medoids.
- Associate each data point to the closest medoid. (“closest” here is defined using any valid distance metric, most commonly Euclidean distance and Manhattan distance.
- For each medoid m.
  o For each non-medoid data point o.
    o Swap m and o and compute the total cost of the configuration.
- Select the configuration with the lowest cost.
- Repeat steps 2 to 4 until there is no change in the medoid.

D. Hierarchical Methods

The hierarchical methods, in general try to decompose the dataset of n objects into a hierarchy of a groups [5].
It works by grouping data objects into a tree of clusters [11]. The tree structure diagram is called as a dendrogram; whose root node represents the whole dataset and each leaf node is a single object of the dataset [5]. The clustering results can be obtained by cutting the dendrogram at different level. Hierarchical clustering methods can be further classified as agglomerative or divisive, depending on whether hierarchical decomposition is formed in a bottom up or top down fashion [11].

**E. Agglomerative Based**

This is a bottom-up (merging) strategy that starts by placing each object in its own cluster and then merges these atomic clusters into large and larger clusters [3].

**Steps:**
- Start with the given objects as individual clusters.
- At each step, merge the closest pair of clusters until only one cluster (or K clusters) left or until a termination condition holds.

**F. Divisive based**

This is a top-down (splitting) strategy that does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller parts [3].

**Steps are as follows:**
- Start with one, all-inclusive cluster.
- At each step, split a cluster until each cluster contains a single object (or there are K clusters) or until a termination condition holds.

**REFERENCES**

Research in Computer Science and Software Engineering”.


Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques” Second Edition.


Categorization of Several Clustering Algorithms from Different Perspective: A Review Prof. Neha Soni and Prof. Amit Ganatra, Volume 2, Issue 8, August 2012 ISSN: 2277 128X.


Comparative Study of K-Means and Hierarchical Clustering Techniques Dr. Manju Kaushik, Mrs. Bhawana Mathur, ISSN No.2347-4890 Volume 2 Issue 6, June 2014.

A Survey of Grid Based Clustering Algorithms Mr. Illango and Dr. V Mohan, Vol. 2(8), 2010, 3441-3446.

Data Mining: Techniques and Algorithms, Divya Chaudhary, Volume 3, Issue 8, August 2013 ISSN: 2277 128X.

Exploring Clustering Algorithms Intended for Mining Data Streams, Dhanammati Jagi, Dr. Sunita Mahajan and Dr. N. Subhash Chandra, IT-NCNHT 2013.

An Improved Clustering Approach on Time Series Data Set, Pallavi and Sunila Godara, RTMC 2011, IJCA.

Clustering Techniques: A Brief Survey of Different Clustering Algorithms Deepti Sisodia, Lokesh Singh, Sheetal Sisodia and Khushboo Saxena, IJLTET.