Comparative study of data clustering techniques

[1]Sukhvir Kaur, [2]Charanjit Singh
[1]M.Tech Student, [2]Assistant Professor
RIMT-IET, Mandi Gobindgarh

Abstract: **- Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The main goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent. Clustering is often confused with classification, but there is some difference between the two.**

***Keywords*:-data minning,data warehouse,clusters,clustering,data minning techniques**

I.INTRODUCTION.

Data mining is a multi-step process. It requires collecting and converging data for a data mining algorithm, prospecting the data, evaluating the results and taking relevant action. Then the collected data can be stored in one or more operational databases, a data warehouse or a flat file[2]. Data analysis and data mining applications are important in clustering and classification

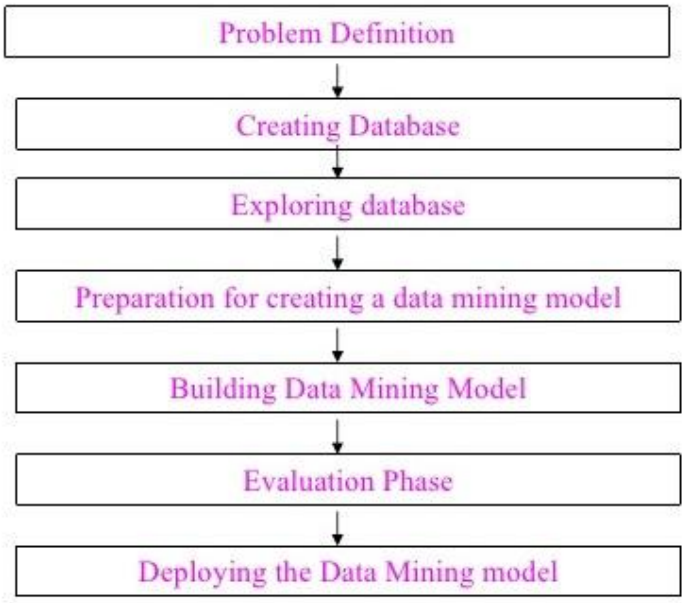

Figure: steps of data mining

Clustering is the technique of finding groups and structures in the data that are in some way or another "similar", without using known structures in the data. The main requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitive to the order of input records, etc .Mainly clustering is the method of grouping a set of objects in such a way that objects in the same group which is called a cluster are more similar to each other than to those in other groups means [3].It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including learning, pattern, image analysis, information retrieval, and bioinformatics. The main advantage of a clustered solution is that it recovers automatically from failure that is recovery without user intervention. But is has disadvantages also like clustering complexity and inability to recover from database corruption. In Clustering not one specific algorithm is used , 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.[4]There is no objectively "correct" clustering algorithm, but as it was noted, "clustering is in the eye of the beholder." The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is a mathematical reason to prefer one cluster model over another. It should be noted that an algorithm that is designed for one kind of model has no chance on a data set that contains a radically different kind of model. For example, k-means cannot find non-convex clusters.

II.COMPONENTS OF CLUSTERING TECHNIQUE

1) Pattern representation.

2) Definition of a pattern proximity measure appropriate to the data domain.

3) Clustering or grouping.

4) Data abstraction.

5) Assessment of output (if needed).

Cluster Analysis is divided into the following stages:-

***Data Collection*:** Includes the careful extraction of relevant data objects from the underlying data sources. In our context, data objects are distinguished by their individual values for a set of attributes or measures.

***Initial Screening*:** Refers to the massaging of data after its extraction from the source, or sources. This stage is closely connected to a process widely used in Data Warehousing, called Data Cleaning .

Representation: Includes the proper preparation of the data in order to become suitable for the clustering algorithm. Here, the similarity measure is chosen, the characteristics and dimensionality of the data is examined.

Clustering Tendency: Checks whether the data in hand has a natural tendency to cluster or not. This stage is often ignored, especially in the presence of large data sets.

***Clustering Strategy*:** Involves the careful choice of clustering algorithm and initial parameters.

Validation: This is one of the last and, in our opinion, most under-studied stages. Validation is often based on manual examination and visual techniques. However, as the amount of data and their dimensionality grow, we have no means to compare the results with preconceived ideas or other clusterings.

Interpretation: This stage includes the combination of clustering results with other studies, e.g., classification, in order to draw conclusions and suggest further analysis.

III. TYPES OF CLUSTERING TECHNIQUES

A)HIERARCHAL CLUSTERING

The hierarchical clustering produce a set of nested clusters in which each pair of objects or clusters is nested into a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods , the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the an un-clustered dataset. In the divisive methods begin with all objects in a single cluster and at each of N-1 steps divides some clusters into two smaller clusters, until each object resides in its own cluster.

i)*Agglomerative Hierarichal Clustering*

The hierarchical agglomerative clustering methods are the most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains , return to step 2.

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.

There are some general approaches to implementation of this algorithm, these being stored matrix and stored data, are discussed below

- In the second matrix approach , an N*N matrix containing all pair wise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to $O(n^3)$ if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.
- The stored data approach required the recalculation of pair wise dissimilarity values for each of the N-1 agglomerations, and the O(N) space requirement is therefore achieved at the expense of an $O(N^3)$ time requirement.

ii)*Divisive hierarichal clustering*

Divisive clustering starts with a single cluster that contains all data points and recursively splits the most appropriate cluster. The process repeats until a stopping criterion or frequently, the requested number k of clusters are achieved.

B) PARTIONING CLUSTERING

Partitioning [5] algorithm is a non-hierarchical method, it construct different partitions and then it evaluates them by some criterion . One such criterion function is minimizing square error criterion which is the following,

$$E = \Sigma \Sigma \| p - mi \|^2$$

Where p is the point in a cluster and mi is the mean of the cluster.In this clustering a cluster should exhibit two properties firstly is that each group must contain at least one object and the other one is each object must belong to exactly one group[review].
The main drawback of this algorithm is whenever a point is close to the center of another cluster; it gives poor result due to overlapping of data points [6].
The K-Means is the simplest clustering algorithm widely used for web proxy server [bath}.it is an algorithm based on finding data clusters in a data set such that a cost function (or an objection function) of dissimilarity (or distance) measure is minimized.{10.1.1} In most cases this dissimilarity measure is chosen as the Euclidean distance.

C)DENSITY-BASED CLUSTERING

Density [1] based clustering algorithm plays vital role in finding nonlinear shapes structure based upon the density. In Density-Based Clustering, clusters are defined as areas of higher density than the remainder of the data.The most popular one is probably DBSCAN (Density-Based Spatial Clustering of Applications with Noise)[4]. DBSCAN is the most well-known

density-based clustering algorithm. Due to its importance in both theory and applications .Unlike K-Means, DBSCAN does not require the number of clusters as a parameter. instead it infers the number of clusters based on the data, and it can discover clusters of arbitrary shape (for comparison, K-Means usually discovers spherical clusters).The ε-neighborhood is fundamental to DBSCAN to approximate local density, so the algorithm has two parameters:

- ε*:* The radius of our neighborhoods around a data point *p.*
- minPts*:* The minimum number of data points we want in a neighborhood to define a cluster.

It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

D)GRID-BASED CLUSTERING

The clustering methods discussed so far are data-driven.they partition the set of objects and adapt to the distribution of the objects in the embedding space. Alternatively**,** a grid-based clustering method takes a space-driven approach by partitioning the embedding space into cells independent of the distribution of the input objects. Grid-based clustering is used when the data space is quantized into finite number of cells which form the grid structure and it performs clustering on the grids{188}. The advantage of this method is its lower processing time. In this method Clustering complexity is based on the number of populated grid cells, and does not depend on the number of objects in the dataset. The major features of this algorithm is that no distance computations are in this method and Clustering is performed on summarized data points, Shapes are limited to the union of grid-cells, and the complexity of the algorithm is usually O(Number of populated grid-cells). An example of this algorithm is STING.

[1] Nanhay Singh, Arvind Panwar, and Ram Shringar Raw, "Enhancing the Performance of Web Proxy Server through Cluster Based Prefetching Techniques",International Conference on Advances in Computing, Communications and Informatics,pp.1158-1165,2013.
[2] V. Sathiyamoorthi, V. Murali Bhaskaran, " Optimizing the Web Cache Performance by Clustering based Pre-Fetching Technique using Modified ART1", International Journal of Computer Applications ,Volume 44,No.1, April 2012.
[3] Pavel Berkhin , "Survey of Clustering Data Mining Techniques", Accrue Software, Inc.
[4] Harminder Kaur, Jaspreet Singh, "Survey of Cluster Analysis and its Various Aspects", Volume 4, Issue 10, pp. 353-363, October 2015.
[5] K.Kameshwaran1, K.Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies, Volume 5 (2),pp. 2272-2276, 2014.
[6] S. Anitha Elavarasi, "A survey on partition clustering algorithms", International Journal of Enterprise Computing and Business Systems,Volume 1, Issue 1, January 2011.
[7] Khaled Hammouda,Prof. Fakhreddine Karray, "A Comparative Study of Data ClusteringTechniques".