International Journal of Computer Applications Technology and Research
Volume 5–Issue 10, 671-674, 2016, ISSN:-2319–8656

Literature Survey: Clustering Technique

Ajinkya V. Jiman,
Marathwada Mitra Mandal's College of Engineering
Pune,India

Prof. Harmeet K. Khanuja,
Marathwada Mitra Mandal's College of Engineering
Pune,India

Abstract: Clustering is a partition of data into the groups of similar or dissimilar objects. Clustering is unsupervised learning technique helps to find out hidden patterns of Data Objects. These hidden patterns represent a data concept. Clustering is used in many data mining applications for data analysis by finding data patterns. There is a number of clustering techniques and algorithms are available to cluster the data object. According to the type of data object and structure appropriate clustering technique is selected. This survey focuses on the clustering techniques for their input attribute data type, their input parameters and output. The main objective is not to understand the actual working of clustering technique. Instead, the input data requirement and input parameters of clustering technique are focused.
Keywords: Clustering, Clustering techniques, Clustering algorithms, Cluster input parameters, Cluster input data type.

1. INTRODUCTION

We live in the world where huge amount of data are collected on daily basis. Analyzing such data to satisfy business or research need is very critical task. Data mining is the process of discovering interesting patterns and knowl- edge from large amount of data. Data mining provides techniques to extract the knowledge from huge data. The one of the interesting tech- nique introduced by data mining is Clustering. Clustering is a partition of data into a group of similar or dissimilar data points and each group is a set of data points called as clusters. Clustering is very useful in Data Mining to find out hidden patterns of data objects. These hid- den patterns are used for data analysis. Data objects can have qualitative or/and quantita- tive attributes. There are numbers of algorithms and techniques are available to cluster the data objects.[4]

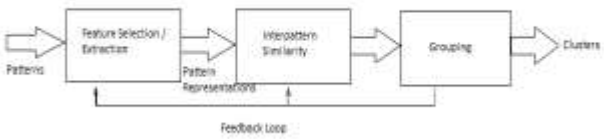

Figure 1. Cluster Analysis Phases

The most effective subset of the original features to use in clustering are identified in feature selection phase. The feature transformation of one or more input features to new salient feature is performed in Extraction phase. These techniques can be used together to obtain an appropriate set of features to use in clustering. In next phase the similarity or dissimilarity between data objects is calculated using similarity measures. According to the data type of data object and need the similarity measure is selected to cluster the data objects. Final stage is actual grouping of data objects, these grouping can be done by number of ways.

Following are some similarity measures used to calculate the similarity between the pair of object:[4]

1) Ecludian Distance
2) Minkowski Distance
3) Pearson Correlation
4) Cosine Similarity

Outlier Analysis: A data set may contain objects that do not comply with the general behavior or model of data. These data objects are outliers. Many data mining techniques remove the outliers as exceptions. In some situation, such rare events or exceptional data objects are useful. E.g. In Banking application to detect the fraud transactions.[4] In this paper the Clustering technique input data object attribute type, their input parameters and output are focused. So in next section the data objects type are described.

2. DATA OBJECT ATTRIBUTE TYPE

An attribute is a data field, representing a characteristic or feature of a data object.[4]

Nominal Attribute : The nominal attribute values consist of name of things or symbols. Each value shows the type of category or code. Nominal attribute are also referred as categorical attribute. E.g. Attribute hair_color, the possible values for hair_color are black, brown, blond, red, gray and white.[4]

Binary Attribute : This attribute is same as nominal attribute and it has only two categories or states: 0 or 1. 0 shows that attribute is absent and 1 means it is present. This type of attribute is also called as Boolean Attributes. E.g. Attribute of patient object, 1 indicates that patient smokes and 0 indicates that patient does not.[4]

Ordinal Attribute : In this attribute type, attribute values have a meaningful order or ranking among them. Here, the magnitude between successive values is not known. E.g. Attribute drink_size has three possible values small, medium and large.[4]

Numeric Attribute : This attributes has measurable quantity represented in integer or real values. The numeric attributes can be ratio-scaled or interval-scaled.[5]

1) Interval Scaled : These attributes are measured on a scale of equal-size units.

2) Ratio-Scaled :In these, the value of an attribute is multiple of another value.

Discrete and Continuous Attributes : A discrete attribute has a finite set of values. These attribute values may or may not be represented as integers. If attribute is not discrete then it is continuous. The continuous attributes are represented using floating point numbers.[4] So, the clustering techniques which are described in the next section have data objects with one of the attribute type described in the above section.

3. CLUSTERING TECHNIQUES

There are number of techniques and algorithms are available to cluster Data Objects.[4]

Partitional : A partitional clustering algorithm constructs partitions of the data. Partitional Clustering algorithms try to locally improve a certain criterion. First, they compute the values of the similarity or distance, they order the results, and pick the one that optimizes the criterion. Such algorithms has very high complexity.

Hierarchical : These techniques creates a hierarchical decomposition of the objects. They are either agglomerative (bottom-up) or divisive (top-down):

(a) Agglomerative algorithms start with each object being a separate cluster itself, and successively merge groups according to a distance measure. The clustering may stop when all objects are in a single group or at any other point the user wants.

(b) Divisive algorithms follow the opposite strategy. They start with one group of all objects and successively split groups into smaller ones, until each object falls in one cluster, or as desired. Divisive approaches divide the data objects in disjoint groups at every step, and follow the same pattern until all objects fall into a separate cluster.

Density-Based Clustering : These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter.

Grid-Based Clustering : The main focus of these algorithms is spatial data, i.e., data that model the geometric structure of objects in space, their relationships, properties and operations. The objective of these algorithms is to quantize the data set into a number of cells and then work with objects belonging to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. In this sense, they are closer to hierarchical algorithms but the merging of grids, and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.

Model-Based Clustering : These algorithms find good approximations of model parameters that best fit the data. They can be either partitional or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitionings. They are closer to density-based algorithms, in that they grow particular clusters so that the preconceived model is improved. However, they sometimes start with a fixed number of clusters and they do not use the same concept of density.

Categorical Data Clustering : These algorithms are specifically developed for data where Euclidean, or other numerical-oriented, distance measures cannot be applied. There are some techniques which are used for clustering, described in next section. Aside from the above categories of clustering methods, there are two classes of clustering task that requires special attention [4]. One is Clustering high-dimensional data, and other is Constraint based Clustering. Most of the clustering methods are designed for low-dimensional data and encounter challenges when the dimensionality of the data grows really high. This is because when the dimensionality increases, usually only small number of dimensions are relevant to certain clusters, but data in the

irrelevant dimensions may produce much noise and mask the real clusters to be discovered. When the data become really sparse, data points located at different dimensions can be considered as all equally distanced, and the distance measure, which is essential for cluster analysis, becomes meaningless. To overcome this difficulty two techniques are used, attribute transformation and attribute selection technique[4]. The attribute transformation summarizes data by creating linear combinations of attributes, and may discover hidden structure in data. This technique is problematic when there are large numbers of irrelevant attributes. The irrelevant information may mask the real clusters, even after transformation Moreover, the transformed attributes are often difficult to interpret, making the clustering result less useful. The attribute selection technique is commonly used for data reduction by removing irrelevant or redundant dimensions. Given a set of attributes, attribute subset selection finds the subset of attribute that are most relevant to the data mining task.

4. CLUSTERING ALGORITHMS

Some previous work is described here that are used to cluster data object[1][2][3][4].

1) Center-Based Partitional Clustering : Kmeans and K-medoid. Both these techniques are based on the idea that a center point can represent a cluster. For Kmeans the notion of a centroid, which is the mean or median point of a group of points.The problem with this technique is we need to specify the number of clusters we want to create. Works on Numeric data attribute.

2) PAM (Partitioning Around Medoids) : is a K-medoid based clustering algorithm that attempts to cluster a set of m points into K clusters. Requires numeric data attribute. Need to specify the number of clusters in advance.

3) CLARA (Clustering LARge Applications): is an adaptation of PAM for handling large data sets. It works by repeatedly sampling a set of data points, calculating the medoids of the sample, and evaluating the cost of the configuration that consists of these sample-derived medoids and the entire data set. Requires numeric data attribute. Need to specify the number of clusters in advance.

4) CLARANS : uses a randomized search approach to improve on both CLARA and PAM. Specifically for use in data mining spatial data mining. Requires numeric data attribute. Need to specify the number of clusters in advance.

5) CURE (Clustering Using Representatives) is a clustering algorithm that can handle large data sets, outliers, and clusters with non-spherical shapes and nonuniform sizes. It is Partitioning technique, and data is partitioned first then hierarchical clustering is used to create cluster. Requires numeric data attribute. Need to specify the number of clusters in advance.

6) Chameleon: uses a graph partitioning algorithm to cluster the data into a large number of relatively small sub-clusters. During the second phase, it uses an agglomerative hierarchical clustering algorithm to find the genuine clusters by repeatedly combining together these subclusters. Requires numeric data attribute. Need to specify the number of clusters in advance.

7) K-mode: Categorical data clustering technique. It requires number of cluster to be created in advance.

8) BIRCH : Numerical data clustering technique. It requires number of cluster to be created in advance.

9) PROCLUS : Numerical data clustering technique. It requires number of cluster to be created in advance.

10) DBSCAN : Numerical data clustering technique. This algorithm requires setting cluster radius in advance.

11) Expectation-Maximization : Numerical data clustering technique. It requires number of cluster to be created in advance.

12) CLIQUE : Numerical data clustering tech nique. This algorithm requires setting grid size and density threshold in advance.

13) ROCK : Categorical data clustering technique. It requires number of cluster to be created in advance. This algorithm doesnt work well in case of data objects with large number of attributes.

14) Wave-Cluster : Spatial-Data Clustering technique. Need to set the number of grid cells each dimension in advance.

15) STING : Spatial-Data Clustering technique. Need to set the number of objects in each cell in advance.

16) OPTICS : Numerical data clustering technique. It requires number of cluster to be created in advance.

17) FCM(Fuzzy C Means) : Numerical data clustering technique. It requires number of cluster to be created in advance. So, here we described most of the clustering algorithms with their required input parameters and input data attribute type.

5. CONCLUSION

In these paper most of the clustering techniques are described. Most of the techniques are developed for numeric data and that works well on low dimensions. Other techniques works on categorical data and spatial data. Most of the technique also requires setting some parameters in advance as input to the algorithm or technique. It may happen that wrong or less effective value to such input parameters will generate improper clusters.

6. ACKNOWLEDGMENTS

I am profoundly grateful to Prof. H. K. Khanuja, H.O.D., Computer Engineering Department for her expert guidance and continuous encouragement throughout to this research.

Also I must express my sincere heartfelt gratitude to all staff members of Computer Engineering Department and my family and friends who helped me directly or indirectly during this course of work.

REFERENCES

[1] M. Halkidi, Y. Batistakis, M. Vazirgiannis, “Clustering algorithms and validity measures”, Scientific and Statistical Database Management, 2001. SSDBM 2001.

[2] A.K. JAIN,M.N. MURTY AND P.J. FLYNN, ”Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999

[3] S.Anita Elavarasi, J. Akilanandeswari,“Survey on clustering algorithms and similarity measure for categorical data”,ICTACT journal on soft computing,JANUARY 2014, VOLUME: 04, ISSUE: 02

[4] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining : Concepts and Techniques : Concepts and Techniques (2nd Edition