International Journal of Contemporary Computer Research (IJCCR), Vol.2 Issue.1 (March, 2018)
ISSN: 2600-9048

Reviewing the Techniques of Data Clustering

Mohamed Yassin Abdelwahab [1], Wael M. S. Yafooz [2],

[1] *Faculty of Computer and Information Technology, Al-Madinah International University, Malaysia, mohcsc_shorouk@hotmail.com*
[2] *Faculty of Computer and Information Technology, Al-Madinah International University, Malaysia, wael.mohamed@mediu.edu.my*

Received 02 February 2018; accepted 21 March 2018

Abstract

Data mining is a field of intersection of computer science and statistics used to discover patterns in the information bank. The main aim of the data mining process is to extract the useful information from the dossier of data structure for future use. There are different process and techniques used to carry out data mining successfully. Data mining is the process of extracting hidden information and patterns from large database. Data mining play a vital role in the leading business environment. It helps to make decisions based on the past information gathered in the database. Data mining is used in various data enhancement processes. These enhancements help in decision-making. Many researchers have recognized mining information and knowledge from large databases as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity of major revenues.

Keywords: Data Cluster, Clustering Method, Clustering Model, Clustering Application.

1. Introduction

Data mining techniques are categorized into two major groups as supervised learning and unsupervised learning. Clustering is a process of grouping the similar data sets into groups. These groups should have two properties like dissimilarity between the groups and similarity within the group. Clustering is covered in the unsupervised learning category. There are no predefined class label exists for the data points. Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc. Clustering helps in gaining, overall distribution of patterns and correlation among data objects. This paper gives the overall idea on the methodologies used in the clustering technique. This paper is formatted in the manner to disclose the various data mining, clustering techniques.

Used and their methodologies. Clustering techniques included are Hierarchical clustering algorithms, K-means clustering algorithms, and Density Based Clustering Algorithm [1].

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups [2].

The notion of Data Mining has become very popular in recent years. Although there is not yet a unique.

Understanding what Data Mining means, the following definition seems to get more and more Accepted.

This information is filtered, prepared and classified so that it will be a valuable aid for decisions and strategies.

The list of techniques, which can be considered under such a definition, ranges from link analysis/associations, sequential patterns, analysis of time series, and classification by decision trees or neural networks, cluster analysis to scoring models [3].

2. What is a cluster?

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters.

Help users understand the natural grouping or structure in a data set. Clustering: unsupervised classification: no predefined classes. Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

Moreover, data compression, outlier's detection, understand human concept formation.

- A cluster is a subset of objects, which are "similar"
- A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it.
- A connected region of a multidimensional space containing a relatively high density of objects [4].

3. What is clustering analysis?

Clustering analysis 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). Following figure is an example of finding clusters of US population based on their income and debt [5].

4. Why do need clustering?

Analytics industry is dominated by objective modelling like decision tree and regression. If decision tree is capable of doing segmentation, do even need such an open-ended technique? The answer to this question is in one of the advantages of using clustering technique. Clustering generates natural clusters and is not dependent on any driving objective function. Hence, such a cluster can be used to analyse the portfolio on different target attributes [6, 7, 8, 9].

5. Types of Clustering

Clustering can be divided into two subgroups:

- Hard Clustering
- Soft Clustering

Types of clustering algorithms

Since the task of clustering is subjective, the means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the 'similarity' among data points. In fact, there are more than 100 clustering algorithms known. However, few of the algorithms are used popularly; let us look at them in detail:

- **Connectivity models:**

These models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases.

In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. In addition, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.

- **Centroid models:**

The notion of similarity is derived by the closeness of a data point to the centroid of the clusters in these iterative clustering algorithms. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.

- **Distribution models:**

These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution. These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm, which uses multivariate normal distributions.

Density Models:

These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS [10].

6. Applications of Clustering

Clustering has a large no. of applications spread across various domains. Some of the most popular applications of clustering are: [11]

- Recommendation engines
- Market segmentation
- Social network analysis
- Search result grouping
- Medical imaging
- Image segmentation
- Anomaly detection.

7. Examples of Clustering Applications

Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs. Land use:

- Identification of areas of similar land use in an earth observation database.
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.
- City planning: Identifying groups of houses according to their house type, value, and geographical location [12].

8. Type of attributes in clustering analysis

- Interval-scaled variables
- Binary variables
- Nominal, ordinal, and ratio variables
- Variables of mixed types
- Remark: variables. Attribute [13].

9. Hierarchical Clustering Weaknesses

The most commonly used type, single-link clustering, is particularly greedy [14].

- If two points from disjoint clusters happen to be near each other, the distinction between the clusters will be lost.
- On the other hand, average- and complete-link clustering methods are biased towards spherical clusters in the same way as k-means
- Does not really produce clusters; the user must decide where to split the tree into groups.
- Some automated tools exist for this
- As with k-means, sensitive to noise and outliers

10. Soft Clustering

Clustering typically assumes that each instance is given a "hard" assignment to exactly one cluster.

Does not allow uncertainty in class membership or for an instance to belong to more than one cluster.

Soft clustering gives probabilities that an instance belongs to each of a set of clusters.

Each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1) [15, 16].

- Requirements of Clustering in Data Mining.
- Scalability.
- Dealing with different types of attributes.
- Discovery of clusters with arbitrary shape.
- Minimal requirements for domain knowledge to determine input parameters.
- Able to deal with noise and outliers.
- Insensitive to order of input records.
- High dimensionality.
- Interpretability and usability.

References

[1] S.Anitha Elavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, January 2011, A Survey On Partition Clustering Algorithms.

[2] Survey of Clustering Data Mining Techniques, Pavel Berkhin, Accrue Software, Inc. www.ijarcsse.com

[3] Santos, J.M, de Sa, J.M, Alexandre, L.A , 2008. LEGClust- A Clustering Algorithm based on Layered Entropic subgraph. Pattern Analysis and Machine Intelligence, IEEE Transactions : 62-75.

[4] M. Livny, R.Ramakrishnan, T. Zhang, 1996. BIRCH: An Efficient Clustering Method for Very Large Databases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery :103-114.

[5] S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM Int'l Conf. Management of Data : 73-84.

[6] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.

[7] Yafooz, W. M., Abidin, S. Z., Omar, N., & Hilles, S. (2016, September). Interactive Big Data Visualization Model Based on Hot Issues (Online News Articles). In International Conference on Soft Computing in Data Science (pp. 89-99). Springer Singapore. "

[8] Yafooz, W. M., Abidin, S. Z., Omar, N., & Idrus, Z. (2013, December). Managing unstructured data in relational databases. In Systems, Process & Control (ICSPC), 2013 IEEE Conference on (pp. 198-203). IEEE.

[9] Yafooz, W. M., Abidin, S. Z., Omar, N., & Halim, R. A. (2013, August). Dynamic semantic textual document clustering using frequent terms and named entity. In System Engineering and Technology (ICSET), 2013 IEEE 3rd International Conference on (pp. 336-340). IEEE.

[10] Hwanjo Yu AND Jiong Yang AND Jiawei Han, "Classifying Large Data Sets Using SVM with Hierarchical Clusters

[11] INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH, VOL 3, ISSUE 11 17 ISSN 2347-4289 Copyright © 2015 IJTEEE.

[12] El-Ebiary, Y. (2015). 37-Data Analysis Techniques. Al-Madinah Technical Studies| مجلة جامعة المدينة العالمية للعلوم التقنية- ماليزيا, 1 (3).

[13] U. Boryczka, "Finding groups in data: Cluster analysis with ants," Applied Soft Computing Journal, vol. 9, pp. 61-70,2009.

[14] DESCRY: a Density Based Clustering Algorithm for Very Large Data Sets, Fabrizio Angiulli, Clara Pizzuti, Massimo Ruffolo

[15] M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96