International Journal of Engineering, Management, Humanities and Social Sciences Paradigms (IJEMHS)
Volume 30, Issue 02, Quarter 02 (April to June 2018)
An Indexed and Referred Journal with Impact Factor: 2.75
ISSN (online): 2347-601x
www.ijemhs.com

Enhancement of Data Mining using Clustering: A Review

Yashi Shrivastava[1] and Tarun Dalal[2]

[1]Department of Computer Science and Engineering, CBS Group of Institution, Jhajjar, Haryana (India)
[2]Assistant Professor, Department of Computer Science and Engineering, CBS Group of Institution, Jhajjar, Haryana (India)

Publishing Date: June 27, 2018

Abstract

Data mining consists of two terms: data and mining. Data term refers how we can collect and pre-process that data according to our purpose whereas mining require one analysis process which is called statistic. Data Mining is presented as a specialized discipline: "Applying statistics and pattern recognition to discover knowledge from data. The data mining is the technique which is applied to extract the useful information from the rough data. Data mining is the notion of all methods and techniques which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. The clustering is the efficient technique of data mining which will cluster the similar and dissimilar type of data. The clustering techniques are of many types like density based, hierarchal clustering etc. In this paper, various techniques of clustering has been reviewed and discussed in terms of various parameters.

Keywords: *Data Mining, Clustering, Hierarchical Methods.*

Introduction

Data mining is the process of extraction hidden knowledge from large volumes of raw data. Data mining has been defined as the nontrivial extraction of previously unknown, and potentially useful information from data. Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. Data mining is the notion of all methods and techniques which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics.

Cluster breakdown groups data objects into cluster such that thing belonging to the same cluster are similar, while those association to different ones are dissimilar. Cluster answer has been widely used in numerous applications, including market research, model recognition, information analysis, and image processing. In business, clustering can help marketers discover interests of their buyer based on purchasing shape and characterize groups of the customers. In biology, it can be used to derive action and animal taxonomies, categorize genes with similar functionality, and increment divination into structures inherent in populations. In geology, professionals tins employ clustering to identify areas of similar lands, similar houses in a city and silverware intelligence clustering tins also be helpful in labelling documents on the Web for information discovery.

Data Mining

Data mining is a popular technological innovation that converts piles of data into useful knowledge that can help the data owners/users make informed choices and take smart actions for their own benefit. In specific terms, data mining looks for hidden patterns amongst enormous sets of data that can help to understand, predict, and guide future behavior. A more technical explanation: Data Mining is the set of methodologies used in analyzing data from various dimensions and perspectives, finding previously unknown hidden patterns, classifying and grouping the data and summarizing the identified relationships.

The elements of data mining include extraction, transformation, and loading of data onto the data warehouse system, managing data in a multidimensional database system, providing access to business analysts and IT experts, analyzing the data by tools, and presenting the data in a useful format, such as a graph or table. This is achieved by identifying relationship using classes, clusters, associations, and sequential patterns by the use of statistical analysis, machine leaning and neural networks.

Data Mining Techniques

The art of data mining has been constantly evolving. There are a number of innovative and intuitive techniques that have emerged that fine-tune data mining concepts in a bid to give companies more comprehensive insight into their own data with useful future trends. Many techniques are employed by the data mining experts, some of which are listed below:

1. Seeking Out Incomplete Data**:**. Data mining relies on the actual data present, hence if data is incomplete, the results would be completely off-mark. Hence, it is imperative to have the intelligence to sniff out incomplete data if possible. Techniques such as Self-Organizing-Maps (SOM's), help to map missing data based by visualizing the model of multi-dimensional complex data. Multi-task learning for missing inputs, in which one existing and valid data set along with its procedures is compared with another compatible but incomplete data set is one way to seek out such data. Multi-dimensional preceptors using intelligent algorithms to build imputation techniques can address incomplete attributes of data
2. Dynamic Data Dashboards**:** This is a scoreboard, on a manager or supervisor's computer, fed with real-time from data as it flows in and out of various databases within the company's environment. Data mining techniques are applied to give live insight and monitoring of data to the stakeholders.
3. Database Analysis**:** Databases hold key data in a structured format, so algorithms built using their own language (such as SQL macros) to find hidden patterns within organized data is most useful. These algorithms are sometimes inbuilt into the data flows, e.g. tightly coupled with user-defined functions, and the findings presented in a ready-to-refer-to report with meaningful analysis.
4. Text Analysis**:** This concept is very helpful to automatically find patterns within the text embedded in hordes of text files, word-processed files, PDFs, and presentation files. The text-processing algorithms can for instance, find out repeated extracts of data, which is quite useful in the publishing business or universities for tracing plagiarism.
5. Efficient Handling of Complex and Relational Data**:** A data warehouse or large data stores must be supported with interactive and query-based data mining for all sorts of data mining functions such as classification, clustering, association, prediction. OLAP (Online Analytical Processing) is one such useful methodology. Other concepts that facilitate interactive data mining are analyzing graphs, aggregate querying, image classification, meta-rule guided mining, swap randomization, and multidimensional statistical analysis.
6. Relevance and Scalability of Chosen Data Mining Algorithms**:**
   While selecting or choosing data mining algorithms, it is imperative that enterprises keep in mind the business relevance of the predictions and the scalability to reduce costs in future. Multiple algorithms should be able to be executed in parallel for time efficiency, independently and without interfering with the transnational business applications, especially time-critical ones. There should be support to include SVMs on larger scale.
7. Popular Tools for Data Mining**:**
   There are many readymade tools available for data mining in the market today. Some of these have common functionalities packaged within, with provisions to add-on functionality by supporting building of business-specific analysis and intelligence.

Clustering

Cluster breakdown groups data objects into cluster such that thing belonging to the same cluster are similar, while those association to different ones are dissimilar. Cluster answer has been widely used in numerous applications, including market research, model recognition, information analysis, and image processing. In business, clustering can help marketers discover interests of their buyer based on purchasing shape and characterize groups of the customers. In biology, it can be used to derive action and animal taxonomies, categorize genes with similar functionality, and increment divination into structures inherent in populations. In geology, professionals tins employ clustering to identify areas of similar lands, similar houses in a city and silverware intelligence clustering tins also be helpful in labelling documents on the Web for information discovery. Data clustering is an unsupervised position method. This office aims at creating groups of goal or clusters in such a resources that objects in the same cluster are very similar and objects in different clusters are quite distinct. Cluster answer is one of the traditional topics in the data mining field. It is the first step in the influence of exciting awareness discovery. The

method of grouping data objects into a series of disjoint classes, called clusters is known as clustering. Now objects within a status have high similarities to each other in the meantime objects in separate classes are more unlike.

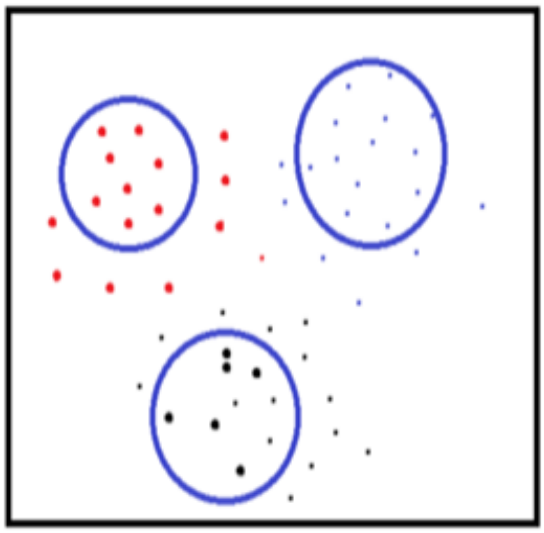

Figure 1: Output of Clustering

Data clustering is an unsupervised position method. This office aims at creating groups of goal or clusters in such a resources that objects in the same cluster are very similar and objects in different clusters are quite distinct. Cluster answer is one of the traditional topics in the data mining field. It is the first step in the influence of exciting awareness discovery. The method of grouping data objects into a series of disjoint classes, called clusters is known as clustering. Now objects within a status have high similarities to each other in the meantime objects in separate classes are more unlike.

Broadly clustering has two areas based on which it can be categorized as follows:

• Hard clustering: In hard clustering same object can belong to single cluster.

• Soft clustering: In this clustering same object can belong to different clusters.

Given there is set of input patterns Y= {y1, ... , yi, ... yN},

where yi= (yi1, ... yid)T € Rd and each is yjd known as variable, feature, dimension or attribute

. • Hard partitioning gives result:

C= {C1,.....,CK} where (K≤N) and i. Ci ≠ ϕ, i=1, 2,...., N

ii. ∪i=1 K $Ci = Y$

iii. $Ci \cap Cj = \emptyset, i, = 1,2, \ldots . .K \; and \; i \neq j$

• Hierarchical clustering has different perspective of representing the output that is tree like structure, partition of Y, = $P1, \ldots \ldots Pr \; where \; r \leq N \; and \; Ci \in Pl \; and \; Cj \in Pm$ and l > m imply $Ci \in Cj \; for \; all \; i,j \neq i,l \; m = 1,2, \ldots .$ }

Categorization of Clustering Methods

There is difference between clustering method and clustering algorithm [1]. A clustering method is a general strategy applied to solve a clustering problem, whereas a clustering algorithm is simply an instance of a method. As mentioned earlier no algorithm exist to satisfy all the requirements of clustering and therefore large numbers of clustering methods proposed till date, each with a particular intension like application or data types or to fulfil a specific requirement. All clustering algorithms basically can be categorized into two broad categories: partitioning and hierarchical, based on the properties of generated clusters . Different algorithms proposed may follows a good features of the different methodology and thus it is difficult to categorize them with the solid boundary. The detail categorization of the clustering algorithm is given in figure 2. Though we had tried to provide as much clarity as possible, there is still a scope of variation. The overview of each categorization is discussed below. A. Hierarchical Methods As the name suggest, the hierarchical methods, in general tries to decompose the dataset of n objects into a hierarchy of a groups. This hierarchical decomposition can be represented by a tree structure diagram called as a dendrogram; whose root node represents the whole dataset and each leaf node is a single object of the dataset. The clustering results.

A. Hierarchical Methods:

As the name suggest, the hierarchical methods, in general tries to decompose the dataset of n objects into a hierarchy of a groups. This hierarchical decomposition can be represented by a tree structure diagram called as a dendrogram; whose root node represents the whole dataset and each leaf node is a single object of the dataset.. There are two general approaches for the hierarchical

method: agglomerative (bottom-up) and divisive (top down).
-An agglomerative method starts with n leaf nodes(n clusters) that is by considering each object in the dataset as a single node(cluster) and in successive steps apply merge operation to reach to root node, which is a cluster containing all data objects. The merge operation is based on the distance between two clusters.
There are three different notions of distance: single link, average link, complete link.
-A divisive method, opposite to agglomerative, starts with a root node that is considering all data objects into a single cluster, and in successive steps tries to divide the dataset until reaches to a leaf node containing a single object. For a dataset having n objects there is 2n-1 – 1 possible two-subset divisions, which is very expensive in computation. Two divisive clustering algorithms: DIANA and MONA.

B. Partitioning Methods:

As the name suggest, the partitioning methods, in general creates k partitions of the datasets with n objects, each partition represent a cluster, where k<= 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-mediods .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[5][6]. The partitioning algorithm in which cluster is represented by one of the objects located near its centre is called as a kmediods. PAM, CLARA and CLARANS are three main algorithms proposed under the k-mediod method .
2) Grid Based: As the name suggest, grid based clustering methods uses a multidimensional grid data structure. It divides the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The representative algorithms based on this method are: STING, Wave Cluster, and CLIQUE.
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 .
4) Density Based: This method has been developed based on the notion of density that is the no of objects in the given cluster, in this context. The general idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold; that is for each data point within a given cluster; the neighbourhood of a given radius has to contain at least a minimum number of points. The density bases algorithms can further classified as: density based on connectivity of points and based on density function. The main representative algorithms in the former are DBSCAN and its extensions, OPTICS, DBCLASD, whereas under the latter category are DENCLUE and SNN

Comparison of Clustering Algorithms

Table 1: Type of Algorithms

Algorithm	Type	Space	Time	Notes
Single Link	Hierarchical	$O(n^2)$	$O(kn^2)$	Not incremental
Average Link	Hierarchical	$O(n^2)$	$O(kn^2)$	Not incremental
Complete Link	Hierarchical	$O(n^2)$	$O(kn^2)$	Not incremental
MST	Hierarchical/ Partitional	$O(n^2)$	$O(n^2)$	Not incremental
Squared Error	Partitional	$O(n)$	$O(tkn)$	Iterative
K-Means	Partitional	$O(n)$	$O(tkn)$	Iterative, No categorical
Nearest Neighbor	Partitional	$O(n^2)$	$O(n^2)$	Incremental
PAM	Partitional	$O(tn^3)$ or $O(tkn^2)$	$O(n^2)$	Iterative
BIRCH	Partitional	$O(n)$	$O(n)$	CF-Tree; Incremental; Outliers
CURE	Mixed	$O(n^2lgn)$	$O(n)$	Heap; k-D tree; Incremental; Outliers
ROCK	Agglomerative	$O(n^2lgn)$	$O(n^2)$	Sampling; Categorical; Links
DBSCAN	Mixed	$O(n^2)$	$O(n^2)$	Sampling; Outliers

Conclusions and Future Work

In this paper we had covered the detailed categorization of the different clustering methods with the representative algorithms under each. The future work planned is to perform a detailed analysis of major clustering algorithms and to do a comparative study.

References

[1] Review on Analysis of Clustering Techniques in Data Mining *Asha Devi1 ,Saurabh Sharma2, published : 01-08-2017
[2] A Review Paper on Data Mining Techniques and Algorithms Ashish Kumar Dogra1 , TanujWala2, published 05-05-2015
[3] Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity Mansi Gera, Shivani Goel, published 18-03-2015.
[4] researchgate.net/publication/49616224_DATA_MINING_TECHNIQUES_AND_APPLICATIONS
[5] Categorization of Several Clustering Algorithms from Different Perspective: A Review Prof. Neha Soni, Prof. Amit Ganatra, published on 8th august 2012
[6] https://www.invensis.net/blog/data-processing/12-data-mining-tools-techniques/
[7] http://slideplayer.com/slide/9105260/ data mining comparison by M vijayalakshmi