ISSN 2348 - 2249
Volume-3, Issue-1, January-March, 2015
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

A Survey on K-Mean Clustering in Data Mining

Gurpreet Kaur Cheena
Research Scholar, Department of CSE
RIMT-IET
Punjab Technical University, Jalandhar
India

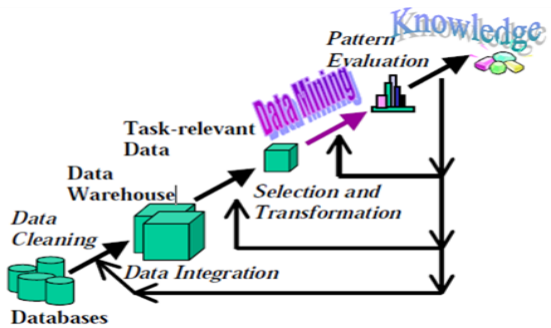
Rupinder Kaur Gurm
Department of CSE
RIMT-IET
Punjab Technical University, Jalandhar
India

Abstract **–** Cluster analysis or clustering is the task of assigning a set of objects into groups called clusters. Main task of clustering are explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval. Cluster analysis itself is not one specific algorithm, but is 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. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. The appropriate clustering algorithm and parameter settings including values such as the distance function to use, a density threshold or the number of expected clusters depend on the individual data set and intended use of the results

Keywords **-** Data Mining, Clustering , K-Mean Clustering.

I. INTRODUCTION

Data Mining is the process of extracting previously unknown but significant information from large databases. It is one of the ways to extract meaningful trends & patterns from huge amounts of data. Data mining is the core process of "KNOWLEDGE DISCOVERY IN DATABASE" it is also known as Knowledge Discovery in Databases (KDD). While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure shows data mining as a step in an iterative knowledge discovery process.



The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

II. DATA MINING TECHNIQUES

Association rule mining

Association rule mining is finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. Application- basket data analysis, clustering analysis, classification. Apriori algorithm is used in this mining.

Classification

Classification is the process of learning a model that describes different classes of data. The classes are predetermined. Once the model is built, then it can be used to classify new data. Example: In a banking application, customers who apply for a credit card may be classify as a "good risk", a "fair risk" or a "poor risk". Hence, this type of activity is also called supervised learning. The model that is produced is usually in the form of a decision tree or a set of rules. The classification methods used are-Bayesian classification, Decision tree ,attribute selection measure, genetic algorithm etc.

Clustering

It is task of grouping the set of objects in such a way that objects in same group are similar to each other than to those in other groups. It is used in many fields as machine learning,

ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

pattern recognition , image analysis , information retrieval and bio informatics. It is an iterative process of knowledge discovery or interactive multi objective optimization that involves trial and failure.

Hierarchical clustering

It is connectivity based clustering that is based on idea of objects being more related to nearby objects than to objects farther away. These algorithms connect objects to form clusters based on their distance. These do not provide a single partitioning of data sets but provide extensive hierarchy of clusters that merge with each other at certain distances.

Decision Trees

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining, machine learning. Leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

III. APPLICATIONS OF DATA MINING

Marketing / Retail - Data mining helps marketing companies to build models base

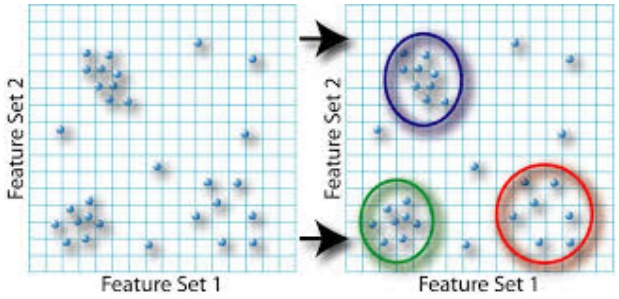
Finance / Banking - Data mining gives financial institutions information about loan information and credit reporting. By building a model from previous customer's data with common characteristics, the bank and financial can estimate what are the good and/or bad loans and its risk level.

Manufacturing - By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters.

Governments - Data mining helps government agency by digging and analysing records of financial transaction to build patterns that can detect money laundering or criminal activity

IV. CLUSTERING

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. Cluster is collection of data objects which are similar to one another in same cluster and dissimilar to objects in other clusters.



There are no pre defined classes so it is an unsupervised learning.

The following are typical requirements of clustering in data mining :

1. Scalability: Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions of objects. Clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.

2. Ability to deal with different types of attributes: Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.

3. Discovery of clusters with arbitrary shape: Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape.

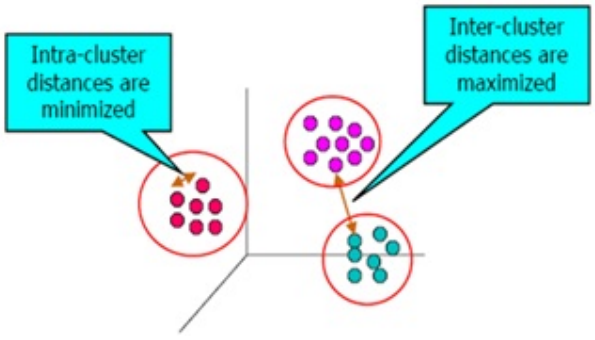
4. Minimal requirements for domain knowledge to determine input parameters: Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters. Parameters are often difficult to determine, especially for data sets containing high-dimensional objects. This not only burdens users, but it also makes the quality of clustering difficult to control.

5. Ability to deal with noisy data: Most real-world databases contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.

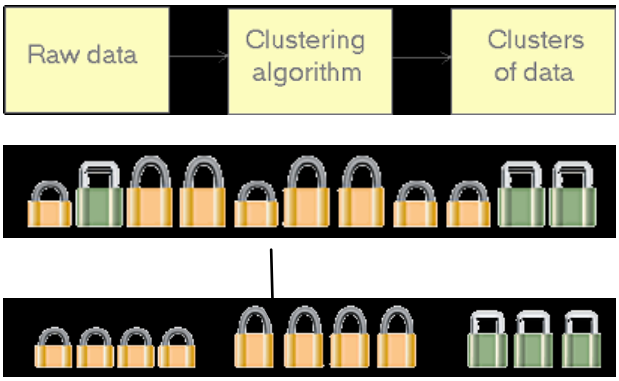
6. Interpretability and usability: Users expect clustering results to be interpretable, comprehensible, and usable. That is, clustering may need to be tied to specific semantic interpretations and applications. It is important to study how an application goal may influence the selection of clustering features and methods.

Cluster analysis

Finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups.



ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)



V. K-MEAN – THE ALGORITHM

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point need to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid.

The algorithm is composed of the following steps:

Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Assign each object to the group that has the closest centroid.

When all objects have been assigned recalculate the positions of k centroids. Repeat Steps 2 and 3 until the centroids no longer move.

Weaknesses of K-means clustering

1. The number of cluster, K, must be determined beforehand.

2. The real cluster never know, using the same data, if it is inputted in a different order may produce different cluster if the number of data is a few.

3. Sensitive to initial condition. Different initial condition may produce different result of cluster. The algorithm may be trapped in the local optimum.

4. Never know which attribute contributes more to the grouping process since assume that each attribute has the same weight.

5. Weakness of arithmetic mean is not robust to outliers. Very far data from the centroid may pull the centroid away from the real one.

6. The result is circular cluster shape because based on distance

VI. CONCLUSION

The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). In k – mean clustering The user has to choose the value of K, the number of clusters. it is not so for higher dimension data, and there are usually no clues as to what number of clusters might be appropriate. Choosing an inappropriate number of clusters will lead to a meaningless typology.

REFRENCES

[1] J. Han, M. Kamber, “Data Mining: Concepts and Techniques,” Second Edition, Elsevier Inc., Rajkamal Electric Press, 2006.
[2] Ó.R. Zaïane, “Introduction to Data Mining,” CMPUT690 Principles of Knowledge Discovery in Databases, University of Alberta, 2009..
[3] J. Gu, X. Chen, J. Zhou, “An Enhancement of K-means Clustering Algorithm” 2009 International Conference on Business Intelligence and Financial Engineering,2009.
[4] O. Altun, N. Dursunoglu, and M. F. Amasyali. Clustering application benchmark. In To Appear in Proceedings of the International Symposium on Workload Characterization (IISWC), Oct. 2006.
[5] Data Mining: Concepts and Techniques, 2nd Edition, Jiawei Han and.heline Kamber, Morgan Kauffman.
[6] John Wiley & Sons Giudici (Paolo), Applied Data Mining Statistical Methods for Business and Industry, England, p.2, 2005.
[7] Two Crows Corporation, Introduction to Data Mining and Knowledge, 2005.
[8] Kantardzic, Mehmed, John Wiley & Sons, Data Mining, Concepts, Models, Methods, and Algorithms, ISBN, 2003.
[9] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, and R. Srikant. The Quest data mining system. In Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining.
[10] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining, pages 307–328,
[11] K. Albayraktaroglu, A. Jaleel, X. Wu, M. Franklin, B. Jacob, C. Tseng, and D. Yeung. BioBench: A benchmark suite of bioinformatics applications. In Proceedings of The 5th International Symposium on Performance Analysis of Systems and Software (ISPASS), Mar. 2005.
[12] T. Pang, M. Steinbach, V. Kumar “Introduction to Data Mining”, Pearson Education, Boston, 2006.
[13] Sonali guglani, Gaurav Gupta mamta dhanda, ways to improve k-means algorithm by using various attributes, IJAEST-2011,VOL-7, issue-2, pages-330-336.
[14] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.

