Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Clustering Techniques in Data Mining

Deepika Sharma
M.Tech Student
Department of CSE
Eternal University, Baru Sahib
Himachal Pradesh-173101, India

ABSTRACT- The main aim of Data mining process is to discover meaningful trends and patterns from the data hidden in repositories. For data analysis and data mining application, Clustering is important. It is a process or technique of grouping a set of objects that belong to the same class. Cluster analysis or Clustering has been widely used in several disciplines, such as statistics, software engineering, biology, psychology and other social sciences, in order to identify natural groups in large amounts of data. These data sets are constantly becoming larger, and their dimensionality prevents easy analysis and validation of the results. There are various clustering techniques like Simple K-Means, EM, Farthest First, Filtered Clustering, Hierarchical Clustering etc. In this research work, a brief introduction to cluster analysis is given.

KEYWORDS – Data mining, Clustering, Clustering Techniques, Clusters, Cluster analysis

I. Introduction

Data mining is the process of extracting interesting information from large amount of data stored in different databases or data warehouses. Data mining tools can be used to predict future in the field of business, knowledge driven systems. The data collection and management systems are already available in mid-range companies but the challenge is to convert this data into success. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). Several Data mining techniques are present like classification, association, clustering, etc. In this research paper clustering analysis is discussed. Clustering means identifying and making groups. A good clustering algorithm is able to identity clusters irrespective of their shapes. Cluster analysis itself is not one specific algorithm, but it can be achieved by several algorithms. Let's take some examples, in city planning; clustering technique helps in identifying groups of houses according to their house type, value and geographical location, in marketing, clustering technique help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.

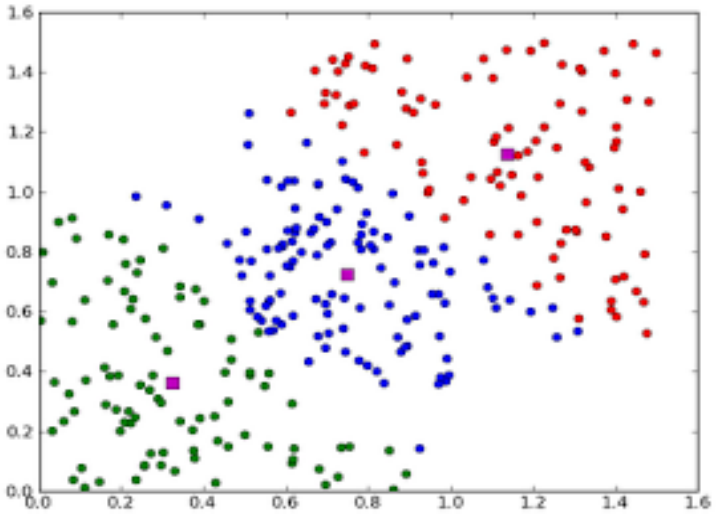

Fig1. An example of a data set with a clear cluster structure

II. TYPE OF CLUSTERS

1. Well-separated clusters: A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster.

2. Centre-based clusters: A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "centre" of a cluster, than to the centre of any other cluster. The centre of a cluster is often a centroid.

3. Contiguous clusters: A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

4. Density-based clusters: A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density.

5. Shared Property or Conceptual Clusters: Finds clusters that share some common property or characterize a particular concept.

III. APPLICATIONS OF CLUSTERING TECHNIQUES

Clustering techniques are applicable in many fields, such as:

- **Libraries:** Book ordering
- **Marketing:** Finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records
- **Biology:** Classification of plants and animals given their features
- **City-planning:** Identifying groups of houses according to their house type, value and geographical location
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost and identifying frauds
- **Spatial Data Analysis:** Create thematic maps in GIS by clustering feature spaces
- **WWW:** Document classification, clustering weblog data to discover groups of similar access patterns
- **Earthquake studies:** Clustering observed earthquake epicentres to identify dangerous zones

IV. CLUSTERING TECHNIQUES

1) Simple K-Means: The simple k-means algorithm is the most popular clustering technique used in technical and business applications. The K-Means algorithm is a technique to cluster objects based on their features into k partitions. It assumes that the *k* clusters show Gaussian distributions. It accepts that the object features form a vector space. The objective it tries to attain is to lessen total intra-cluster variance. Two versions of k-means iterative optimization are known. The first version is related to EM algorithm and comprises of two-step major iterations that (1) reassign all the points to their adjacent centroids, and (2) recomputed centroids of newly gathered groups. Iterations last until an ending criterion is attained. The second version of k-means iterative optimization reallocates points based on more detailed analysis of effects on the objective function caused by moving a point from its present cluster to a potentially new one. If a move has a positive effect, the point is repositioned and the two centroids are recomputed. It is not clear that this version is computationally possible, because the outlined examination needs an inner loop over all member points of involved clusters affected by centroids moves.

2) EM: The EM algorithm estimates the parameters of a model iteratively. Basically it starts with initial values for the parameters, and then it calculates the cluster probabilities for each instance. Re-estimation is done for the values of the parameters. In the end, repeat the process. An EM algorithm is iterative method for finding maximum likelihood or MAP estimates of parameters is statistical models, where the model depends on unobserved latent variables. It alternates between performing an expectation (E) step, which computes the expectation of the likelihood evaluated using the current estimate for parameters maximizing the expected log-likelihood found on the E Step. These parameters estimates are then used to determine the distribution of the latent variables in the next E Step.

3) Farthest First: Farthest first is a variant of K-Means that places each cluster centre in turn at the point farthest from the existing cluster centres. This point must lie within the data area. Farthest-point heuristic method is suitable for large-scale data mining applications. Farthest-point heuristic based method has the time complexity O(nk), where n is number of objects in the dataset and k is number of desired clusters.

4) Filtered Clustering: This algorithm is used for the purpose of filtering the information or pattern. In this the user provides the keywords or a sample set of relevant information. On the arrival of new information they are compared against the filtering profile and the information matching the keywords is presented to the user. The user is not provided with the details of filtering algorithms used by the system. Filtering of information or pattern by collaboration of multiple agents, data sources and viewpoints is referred to as collaborative filtering.

5) Hierarchical Clustering: Hierarchical Clustering produces a nested sequence of clusters, a tree, also called Dendrogram. There are two types of Hierarchical Clustering:

i. Agglomerative (bottom up) clustering: It builds the dendrogram (tree) from the bottom level, merges the most similar (or nearest) pair of clusters and stops when all the data points are merged into a single cluster (i.e., the root cluster).

ii. Divisive (top down) clustering: It starts with all data points in one cluster, the root. Then it splits the root into a set of child clusters. Each child cluster is recursively divided further and stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point.

6) Make Density Based Clustering: The cluster will be constructed based on the density properties of the database are derived from a human natural clustering approach. The clusters and consequently the classes are easily and readily identifiable because they have an increased density with respect to the points they possess. The elements of the database can be classified in two different types: the border points, the points located on the extremities of the cluster, and the core points, which are located on its inner region.

7) DBSCAN: Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity.

8) OPTICS: Although DBSCAN can cluster objects given input parameters such as ε and minPoints, it still leaves the user with the responsibility of selecting parameter values that will lead to the discovery of acceptable clusters. Actually, this is a problem associated with many other clustering algorithms. Such parameter settings are usually empirically set and difficult to determine, especially for real world & high dimensional data sets. Most algorithms are very sensitive to such parameter value: slightly different settings may lead to very different clustering of the data. To help overcome this difficulty, a cluster analysis method called OPTICS was proposed. Rather than producing a data set cluster explicitly, OPTICS computes an augmented cluster ordering for automatic and interactive cluster analysis. This ordering represents the density-based clustering obtained from a wide range of parameter settings. The cluster ordering can be used to extract basic clustering information as well as provide the intrinsic clustering structure.

V. CONCLUSION

Data Mining is a growing technology that combines techniques including statistical analysis, visualization, decision trees and neural network to explore large amount of data and discover relationship and patterns that shed light on business problems. Among other data mining techniques, clustering technique is of great use. It is an unsupervised learning method that attempts to find collections/groups of objects that are close to each other. Close is defined by a distance measure, and dissimilar clusters arise depending on the particulars of the distance measure. Selecting the number of clusters is a second task that has several different solutions. Sometimes the quantity of clusters can be specified within the business problem, other times an experiential method is used to estimate the suitable number. If the predictors and executives are aware of better measures, better clusters can be formed by describing a new variable that holds a specialist's allocated measure.

REFERENCES

[1] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques", *International Journal of Science, Engineering and Technology Research*, vol. 2, pp. 803-806, April 2013.

[2] Oded Maimon and Lior Rokach, "DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK", *Springer*, pp. 321-352, 2005.

[3] Pragati Shrivastava and Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data", *International Journal of Advanced Computer Research*, pp. 2249-7277, September 2012.

[4] Pradeep Rai and Shubha Singh, " A Survey of Clustering Techniques", *International Journal of Computer Applications*, October 2010.

[5] B.S. Everitt, S. Landau and M. Leese, "Cluster Analysis*", Oxford University* Press, fourth edition, 2001.

[6] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", *Printice Hall*, 1988.

[7] M. R. Anderberg, "Cluster Analysis for Applications", *Academic Press*, 1973.

[8] Tan P, Steinbach M and Kumar, "Introduction to Data Mining", *Addison Wesley*, vol.1, pp. 157-169, 2006.

[9] Er. Arpit Gupta , Er.Ankit Gupta and Er. Amit Mishra, "RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS*", International Journal of Advance Technology & Engineering Research*, vol. 1, pp. 39-47, 2011.

[10] PAULRAJ PONNIAH, "DATA WAREHOUSING FUNDAMENTALS FOR IT PROFESSIONALS", *A John Wiley & Sons, Publication*, 2nd Edition New Jersey.

[11] Kalyani M Raval, "Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, pp. 439-442, 2012.

[12] Bhoj Raj Sharma and Aman Paul (2013), "Clustering Algorithms: Study and Performance Evaluation Using Weka Tool", *International Journal of Current Engineering and Technology*, vol. 3, pp. 1094-. 1098, 2001.

