Asian Journal of Information Technology 11 (1): 40-44, 2012
ISSN: 1682-3915

Clustering Methods and Algorithms in Data Mining: Review

[1]L. Arockiam, [1]S.S. Baskar and [2]L. Jeyasimman
[1]Department of Computer Science, St. Joseph's College, Trichirappalli, Tamil Nadu, India
[2]Department of Computer Applications, JJ College of Engineering and Technology,
Trichirappalli, Tamil Nadu, India

Abstract: Retrieval of information from the databases is now a day's significant issues. The thrust of information for decision making is challenging one. To overcome this problem, different techniques have been developed for this purpose. One of techniques is data clustering. In this study, data clustering methods are discussed along with its two traditional approaches and their algorithms. Some applications of data clustering like data mining using data clustering and similarity searching in medial image databases are also discussed.

Key words: Agglomerative, divisive, clustering, mining, databases, image

INTRODUCTION

Clustering is a one of the method in data mining. It means that process of grouping a set of physical or abstract objects into classes of similar objects (Arabie and Hubert, 1996). This study reviews the different comprehensive methods of clustering techniques. Clustering is grouping of data into similar groups with respect to similarity of the data. The cluster has the characteristics of more similarity within the group. While making the cluster, there is possibility of losing the fine details. But cluster achieves the simplification. Clustering is a process of unsupervised learning which results the data concept. It is otherwise described as unsupervised learning of a hidden data concept. Though, lots of techniques exist in data mining, the clustering techniques in large databases are reviewed here. Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data using pattern recognition technologies as well as statistical and mathematical techniques (Cadez *et al*., 2001). Data mining is a knowledge discovery process of extracting previously unknown, actionable information from very large databases (Cadez *et al*., 2001). Clustering is one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. The buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign. And also, a company which is selling a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking are examples for clustering. This will be done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

CLUSTERING GENESIS

More references are made in the clustering techniques. Studies carried on clustering include Fasulo (1999), Ghosh (2002), Jain *et al.* (1999), Kolatch (2001) and Klise and McKenna (2006). The brief details on clustering techniques can be seen in the book (Klise and McKenna, 2006). Clustering techniques are having close relationship with other disciplines. This technique has been used in statistics (Arabie and Hubert, 1996). In the agricultural science, data mining clustering techniques are found in grading apples before marketing (Fasulo, 1999) detecting weeds in precision agriculture (Tellaeche *et al*., 2008) and monitoring water quality changes (Klise and McKenna, 2006). Clustering techniques are widely used in data compression in image processing, it is otherwise known as vector quantization (Gersho and Gray, 1992). This study just briefs the clustering techniques of different areas. Clustering in data mining was brought to routine by greatest development in information retrieval

Corresponding Author: L. Arockiam, Department of Computer Science, St. Joseph's College, Trichirappalli, Tamil Nadu, India

and text mining (Dhillon *et al*., 2001; Steinbach *et al*., 2000). This techniques has been used in different areas such as spatial data base applications GIS or astronomical data (Ester *et al*., 2000). This techniques can also been seen in sequence and heterogeneous data analysis (Cadez *et al*., 2001) and web applications (Foss *et al*., 2001; Heer and Chi, 2001). References for these techniques can also be found in DNA analysis in computational biology (Ben-Dor *et al*., 1999).

CLUSTERING METHODS

There are many clustering methods available and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters with or without overlap.

These methods are sometimes divided into partitioning methods in which the classes are mutually exclusive and the less common clumping method in which overlap is allowed. Each object is a member of the cluster with which it is most similar however, the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy is build up in a series of N-1 agglomerations or fusion of pairs of objects, beginning with the un-clustered dataset. The less common divisive methods begin with all objects in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster.

PARTITIONING METHODS

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Single pass: A very simple partition method, the single pass method creates a partitioned dataset as follows: make the first object the centroid for the first cluster. For the next object, calculate the similarity, S with each existing cluster centroid, using some similarity coefficient. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name implies this method requires only one pass through the dataset; the time requirements are typically of order O(NlogN) for order O(logN) clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed with the first clusters formed usually being larger than those created later in the clustering run.

HIERARCHICAL AGGLOMERATIVE METHODS

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by two closest objects and merge them into a cluster and also find and merge the next two closest points where a point is either an individual object or a cluster of objects. Individual methods are characterized by the definition used for identification of the closest pair of points and by the means used to describe the new cluster when two clusters are merged here are some general approaches to implementation of this algorithm, these being stored matrix and stored data are discussed. In the second matrix approach, an N x N matrix containing all pair wise distance values is first created and updated as new clusters are formed. This approach has at least an $O(n \times n)$ time requirement, rising to $O(n^3)$ if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N. The stored data approach required the recalculation of pair wise dissimilarity values for each of the N-1 agglomerations and the O(N) space requirement is therefore achieved at the expense of an $O(N^3)$ time requirement.

Single Link Method (SLINK): The single link method is probably the best known of the hierarchical methods and operates by joining at each step, the two most similar objects which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

Complete Link Method (CLINK): The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

Group Average Method: The Group Average Method relies on the average value of the pair wise within a cluster rather than the maximum or minimum similarity as with the single link or the complete link methods. Since, all objects in a cluster contribute to the inter-cluster similarity, each object is on average more like every other member of its own cluster then the objects in any other cluster.

Text based documents: In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document.

Categorization of clustering algorithms: Algorithms are key step for solving the techniques. In these clustering techniques, various algorithms are currently in the life, still lot more are evolving. But in general the algorithm for clustering is neither straight nor canonical.

Clustering algorithms

Hierarchical methods:

- Agglomerative algorithms
- Divisive algorithms

Partitioning methods:

- Relocation algorithms
- Probabilistic clustering
- K-medoids Methods
- K-means Methods

Density-based algorithms:

- Density-based connectivity clustering
- Density functions clustering

Grid-based methods:

- Methods based on co-occurrence of categorical data
- Constraint-based clustering
- Clustering algorithms used in machine learning
- Gradient descent and artificial neural networks

Evolutionary methods:

- Scalable clustering algorithms
- Algorithms for high dimensional data
- Subspace clustering
- Projection techniques
- Co-clustering techniques

Normally clustering techniques are broadly divided in hierarchical and partitioning. Hierarchical clustering is further subdivided into agglomerative and divisive. The ground rules of hierarchical clustering include Lance-Williams formula, idea of conceptual clustering. Some of classic algorithms SLINK, COBWEB, CURE and CHAMELEON are under hierarchical clustering. While hierarchical algorithms build clusters gradually, partitioning algorithms learn clusters directly. These algorithms try to discover clusters by iteratively relocating points between subsets or try to identify clusters as areas highly populated with data. Algorithms of the first kind are surveyed in the section Partitioning Relocation Methods. They are further categorized into probabilistic clustering (EM framework, algorithms SNOB, AUTOCLASS, MCLUST), K-medoids Methods (algorithms PAM, CLARA, CLARANS and its extension) and K-means Methods (different schemes initialization, optimization, harmonic means, extensions). Such methods concentrate on how well points fit into their clusters and tend to build clusters of proper convex shapes.

APPLICATIONS

Data clustering has immense number of applications in every field of life. One has to cluster a lot of thing on the basis of similarity either consciously or unconsciously. So, the history of data clustering is old as the history of mankind. In computer field also, use of data clustering has its own value. Specially in the field of information retrieval data clustering plays an important role. Some of the applications are given.

In order to detect many diseases like Tumor, etc., the scanned pictures or the X-rays are compared with the existing ones and the dissimilarities are recognized.

Medical field have clusters of images of different parts of the body. For example, the images of the CT scan of brain are kept in one cluster. To further arrange things, the images in which the right side of the brain is damaged are kept in one cluster. The hierarchical clustering is used. The stored images have already been analyzed and a record is associated with each image. In this form a large database of images is maintained using the hierarchical clustering. Now when a new query image comes, it is firstly recognized that what particular cluster this image belongs and then by similarity matching with a healthy image of that specific cluster the main damaged portion or the diseased portion is recognized. Then, the image is sent to that specific cluster and matched with all the images in that particular cluster. Now the image with which the query image has the most similarities is retrieved and the record associated to that image is also associated to the query image. This means that now the disease of the query image has been detected.

Using this technique and some really precise methods for the pattern matching, diseases like really fine tumor can also be detected. So, by using clustering an enormous amount of time in finding the exact match from the database is reduced.

K-MEANS ALGORITHM

K-means algorithm introduced by J.B. MacQueen in 1967 is one of the most common clustering algorithms that groups data with similar characteristics or features together. These groups of data are called clusters. The data in a cluster will have similar features or characteristics which will be dissimilar from the data in other clusters. This is how K-means algorithm partitions a dataset into clusters: it accepts the number of clusters to group data into and the dataset to cluster as input values.

It then creates the first K initial clusters (K = Number of clusters needed) from the dataset by choosing K rows of data randomly from the dataset. For example if there are 10,000 rows of data in the dataset and 3 clusters need to be formed then the first K = 3 initial clusters will be created by selecting 3 records randomly from the dataset as the initial clusters. Each of the 3 initial clusters formed will have just one row of data.

The K-means algorithm calculates the arithmetic mean of each cluster formed in the dataset. The arithmetic mean of a cluster is the mean of all the individual records in the cluster. In each of the first K initial clusters, their is only one record. The arithmetic mean of a cluster with one record is the set of values that make up that record. For example if the dataset we are discussing is a set of height, weight and age measurements for students in a university where a record p in the dataset s is represented by a height, weight and age measurement, then P = (age, height, weight). Then, a record containing the measurements of a student John would be represented as John = (20, 170, 80) where John's Age = 20 years, Height = 1.70 m and Weight = 80 Pounds. Since, there is only one record in each initial cluster then the Arithmetic Mean of a cluster with only the record for John as a member = (20, 170, 80).

Next, K-means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity like the Euclidean Distance Measure or Manhattan/City-Block Distance Measure.

K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset. The arithmetic mean of a cluster is the arithmetic mean of all the records in that cluster. For example if a cluster contains two records where the record of the set of measurements for John = (20, 170, 80) and Henry = (30, 160, 120) then the arithmetic mean P_{mean} is represented as P_{mean} = (Age_{mean}, $Height_{mean}$, $Weight_{mean}$). Age_{mean} = (20+30)/2, $Height_{mean}$ = (170+160)/2 and $Weight_{mean}$ = (80+120)/2. The arithmetic mean of this cluster = (25, 165, 100). This new arithmetic mean becomes the center of this new cluster. Following the same procedure, new cluster centers are formed for all the existing clusters.

K-means re-assigns each record in the dataset to only one of the new clusters formed. A record or data point is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity like the Euclidean Distance Measure or Manhattan/City-Block Distance Measure.

The preceding steps are repeated until stable clusters are formed and the K-means clustering procedure is completed. Stable clusters are formed when new iterations or repetitions of the K-means clustering algorithm does not create new clusters as the cluster center or arithmetic mean of each cluster formed is the same as the old cluster center. There are different techniques for determining when a stable cluster is formed or when the K-means clustering algorithm procedure is completed.

CONCLUSION

In this study, the basic concept of clustering and clustering techniques are given. The processes of grouping a set of physical or a bstract objects into classes of similar objects are named as clustering. These

techniques are being used in many areas such as marketing, agriculture, biology and medical. This study concludes that clustering techniques become a highly active research area in data mining research.

REFERENCES

Arabie, P. and L.J. Hubert, 1996. An Overview of Combinatorial Data Analysis. In: Clustering and Classification, Arabie, P., L.J. Hubert and G.D. Soete (Eds.). World Scientific Publishing Co., New Jersey, USA., pp: 5-63.

Ben-Dor, A., R. Shamir and Z. Yakhini, 1999. Clustering gene expression patterns. J. Comput. Biol., 63: 281-297.

Cadez, I.V., P. Smyth and H. Mannila, 2001. Probilistic modeling of transactional data with application to profiling, visualization and prediction. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, California, pp: 37-46.

Dhillon, I., J. Fan and Y. Guan, 2001. Efficient Clustering of Very Large Document Collections. In: Data Mining for Scientific and Engineering Applications, Grossman, R.L., C. Kamath, P. Kegelmeyer, V. Kumar and R.R. Namburu (Eds.). Kluwer Academic Publishers, USA.

Ester, M., A. Frommelt, H.P. Kriegel and J. Sander, 2000. Spatial Data mining data bases primitives, algorithms and efficient DBMS support. Data Mining Knowledge Dis., 4: 193-216.

Fasulo, D., 1999. An analysis of recent work on clustering algorithms. Technical Report, University of Washington.

Foss, A., W. Wang and O. Zanne, 2001. A non parametric approach to web log analysis. 1st SIAM ICDM. Workshop on web mining. Chicago, pp: 51-58.

Gersho, A. and R.M. Gray, 1992. Vector Quantization and Signal Compression. In: Communications and Information Theory, Gersho, A. and R.M. Gray (Eds.). Kluwer Academic Publishers, Norwell, USA..

Ghosh, J., 2002. Scalable Clustering Methods for Data Mining. In: Handbook of Data Mining, Nong, Y. (Ed.). Erlbaum, Lawrence.

Heer, I. and E. Chi, 2001. Identification of web user traffic composition using multi-modal clustering and information scent. Proceedings of the 1st SIAM ICDM Workshop on Web Mining, April 5-7, 2001, Chicago, pp: 51-58.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surveys, 31: 264-323.

Klise, K.A. and S.A. McKenna, 2006. Water quality change detection: Multivariate algorithms. Proceedings of the SPIE (International Society for Optical Engineering), Defense and Security Symposium, April 18-20, 2006, Orlando. Florida, pp: 9-9.

Kolatch, E., 2001. Clustering algorithms for spatial data bases: A survey. Department of Computer Science, University of Maryland, College Park, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1145.

Steinbach, M., G. Karypis and V. Kumar, 2000. A comparison of document clustering techniques. Proceedings of the 6th ACM SIGKDD World Text Mining Conference, August 20-23, 2000, Boston, pp: 1-2.

Tellaeche, A., X.P. Burgos-Artizzu, G. Pajares and A. Ribeiro, 2008. A vision based hybrid classifier for weed detection in precision agriculture through the bayesian and fuzzy k-means paradigms. Adv. Soft Comput., 44: 72-79.