A Review of Clustering Algorithms

[1]Suchita S. Mesakar, [2]M. S. Chaudhari
[1,2]Dept. of CSE, Smt.Bhagwati Chaturvedi College of Engg., Nagpur, India

Abstract

Data mining is the process of extracting meaningful data or knowledge from large amount of data. Clustering is the dynamic field of research in data mining. Data clustering is used in variety of applications like pattern matching, machine learning, image segmentation and information retrieval. The aim of clustering is to group data into clusters or groups, so that data in the same cluster are more similar to each other than to those in other clusters. There is large amount of data available in the database; fast retrieval of data from database is always required. So clustering the data will ease the task of retrieval of data from database. This paper presents an overview of various clustering algorithms used for clustering numerical and categorical data.

Keywords

Clustering, Clustering Algorithm, Categorical Data, Data Mining

I. Introduction

Clustering is a data mining technique and it plays a vital role in classification of data. The database consists of large amount of data. This large amount of data can be grouped into meaningful data for further analysis or management. Clustering plays important role in management and analysis of data. Fast retrieval of data from database is always a need and if large amount of data is classified into meaningful groups or clusters then it will be easier and faster to access the data from the database. The goal of clustering is to cluster the data into groups or cluster depending on the similarity and dissimilarity measures.

Data clustering is the process of organizing objects into groups whose members are similar in some way. Clustering algorithm partitions the data into certain number of clusters. Clustering is used in many areas like machine learning, pattern recognition together with data mining, document retrieval, image segmentation. Learning can be classified as supervised learning and unsupervised learning. Clustering can be considered as unsupervised learning problem as it deals with finding a structure in a collection of unlabeled data. In unsupervised learning for given set of patterns, a collection of clusters is to be discovered and additional patterns are assigned to correct cluster. In supervised learning set of classes (clusters) are given, new pattern (point) are assigned to proper cluster, and are labeled with label of its cluster. Clustering is often called an unsupervised learning task because no class values are given which denotes an a priori grouping of the data instances.

Clustering is the dynamic field of research in data mining. There exist a large number of clustering algorithms in the literature. The clustering algorithms partition data into certain number of clusters based on similarity and dissimilarity. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. The clustering can be performed on numerical data and categorical data.

The numeric data can be ordered naturally and the properties can be used to apply distance measures to the attribute values. The examples of numeric attributes are age,cost,weight But in case of categorical data the attribute values are not numeric and have no specific order, so distance measures cannot be applied directly to categorical data. The example of categorical attribute is color = {red, green, blue} or shape= {circle, rectangle, square, ellipse}. Clustering can be classified as hard clustering and soft clustering. In hard clustering the object belongs to exactly one cluster. Soft clustering is also referred as fuzzy clustering. In fuzzy clustering method, the objects can belong to several clusters simultaneously, depending on the degrees of membership associated with each object. The major clustering algorithm can be categorized into

1. Hierarchical Algorithms
   - Agglomerative
   - Divisive
2. Partitional Algorithms
3. Density Based Algorithms
4. Grid Based Algorithms

Hierarchical clustering algorithms organize data into hierarchical structure. Hierarchical clustering is set of nested clusters that are organized in the form of tree. It starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. Hierarchical algorithms are classified into agglomerative methods and divisive methods. An agglomerative approach is also called as bottom up approach. It considers each pattern as a distinct cluster, and successively merges clusters together until a stopping criterion is satisfied. A divisive method also called as top down begins with all patterns in a single cluster and performs splitting until a stopping criterion is met [16].

A partitional clustering algorithm divides the set of data objects into disjoint clusters. Such that each data object belongs to only one cluter.Partitional clustering algorithm splits the data points into k partition, where each partition represents a separate cluster.

Density-based algorithms identify clusters as dense regions of objects in the data space separated by regions of low density.

Grid based methods first divide space into grids, and then performs clustering on the grids. The main advantage of Grid based method is its fast processing time which depends on number of grids in each dimension in quantized space. The density-based partitioning methods work best with numerical attributes, and grid-based methods work with attributes of different types.

II. Review of Different Clustering Algorithms

Many clustering algorithms are present for clustering numeric data. The numeric data consists of numeric attributes. Finding similarity between numerical objects usually relies on common distance measures such as Euclidean, Manhattan, Minkowski and Mahalanobis distances. The overview of various clustering algorithm for numerical data is given below.

K-means [1] clustering has been a very popular technique for partitioning large data sets with numerical attributes. The K-means algorithm is a partition-based clustering method which is simple and unsupervised. The aim of K-means is to partition data objects d into k clusters in which each object belongs to the cluster with the nearest mean. K in k-means is the number of cluster which is user input to the algorithm. It is iterative in nature. The K-means algorithm is simple and fast. The K-Means algorithm is applicable to numerical data.

A hierarchical algorithm is proposed in BIRCH [3].It is used for clustering large numerical datasets in Euclidean spaces. BIRCH can only deal with metric attributes. A metric attribute is one whose values can be represented by explicit coordinates in a Euclidean space. An advantage of Birch is its ability to incrementally and dynamically cluster incoming, metric data points in order to produce the best quality clustering for a given set of resources.

CURE [4] is hierarchical agglomerative algorithm that clusters numerical datasets. Traditional clustering algorithms favors either clusters with spherical shapes and similar sizes. CURE identifies clusters having non-spherical shapes and wide variances in size. CURE employs combination of random sampling and partitioning in order to process large database.

Chameleon [5] proposed by G. Karypis, E. Han, and V. Kumar, is agglomerative hierarchical clustering algorithm. Chameleon finds the clusters in the data set by using a two-phase algorithm .In the first phase, Chameleon uses a graph partitioning algorithm to cluster the data items into several relatively small sub clusters and in the second phase, it uses an algorithm to find the clusters by repeatedly combining these sub clusters.

A density-based clustering algorithm is proposed in DBSCAN [6]. It is a density-based clustering algorithm because it finds a number of clusters based on density distribution of corresponding nodes. A data point x is density-reachable from a data point y, if x is in the neighborhood of y and also y is surrounded by more than a certain number of data points.Then,it can be considered that x and y are in same cluster. In that case, we can consider that p and q are in the same cluster. This method is used to discover arbitrarily shaped clusters, DBSCAN does not perform any pre- clustering and it is executed directly on the entire database.

Many algorithms are present for clustering categorical data. Categorical data has a different structure than the numerical data. Categorical data contains attribute values which do not have specific order like numerical attributes. Categorical data contain non numeric attributes. The distance functions in the numerical data might not be applicable to the categorical data. Algorithms for clustering numerical data cannot be applied to categorical data.

K-modes, [2] algorithm extends the K-means algorithm with respect to categorical domain.K-modes is extension of K-means which introduces a new dissimilarity measure for categorical data it uses modes instead of means for clusters, and a frequency based method is used to update modes in the clustering process in order to minimize the clustering cost function.

Squeezer [7] algorithm is used for clustering of categorical data. The squeezer algorithm reads each tuple sequentially, either assigning it to existing cluster, or creating new cluster, which is determined by similarities between tuples and clusters. The squeezer repeatedly reads tuples from dataset, when the first tuple occurs it forms cluster and the consequent tuples are either put into existing cluster or a new cluster is formed depending on the similarity function defined between tuple and cluster.

LIMBO [8] which is a hierarchical clustering algorithm for categorical data. Limbo uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuple, and allows clustering of various sizes in single execution.

ROCK: A Robust Clustering Algorithm for Categorical Attributes [9] proposed by S. Guha, R. Rastogi, and K. Shim , is a agglomerative hierarchical clustering algorithm for data with categorical attributes. ROCK explores the concept of links to measure the similarity between a pair of data points and uses links and not distances while merging the clusters.

V. Ganti, J. Gehrke, and R. Ramakrishnan proposed CACTUS [10] a fast summarization based algorithm for categorical clustering of data. The idea implemented in CACTUS is that summary constructed from dataset is sufficient for discovering clusters. CACTUS consists of three phases: summarization, clustering, and validation. In the summarization phase, the summary information from the dataset is computed. In the clustering phase, the summary information is used to discover a set of candidate clusters. In the validation phase, the actual sets of clusters are determined from the set of candidate clusters.

D. Barbara, Y. Li, and J. Couto proposed COOLCAT [11], a new method which uses the notion of entropy to group records. COOLCAT is an incremental algorithm that aims to minimize the expected entropy of the clusters. COOLCAT works incrementally that aims to minimize the entropy of the clusters for the given set of clusters, and it is capable to cluster every new point without having to re-process the entire set. COOLCAT will place the next point in the cluster where it minimizes the overall expected entropy.

D. Gibson, J. Kleinberg, and P. Raghavan [12] proposed an algorithm called STIRR (Sieving Through Iterated Relational Reinforcement), for clustering of categorical data. It converts dataset into weighted graph and propagates these weights in iterative manner; this corresponds to a similarity measure based on co-occurrence of values in the dataset.STIRR maps categorical data to non-linear dynamic systems and it is an iterative approach.

CLOPE [13] proposed Y. Yang, S. Guan, and J. You is a novel algorithm for categorical data clustering. Clope proposes a global criterion function that tries to increase the intra-cluster overlapping of transaction items by increasing the height-to-width ratio of the cluster histogram. The simple idea behind CLOPE makes it fast, scalable, and memory saving in clustering large, sparse transactional databases with high dimensions.

M.J. Zaki and M. Peters proposed CLICKS [14] that finds clusters in categorical datasets based on a search for k-partite maximal cliques. CLICKS uses a selective vertical expansion approach to guarantee complete search. CLICK outperforms and scale better on high dimensional datasets.

D. Cristofor and D. Simovici [15] proposed a genetic algorithm for clustering of categorical data. The approach finds partition of the rows of a given database that is as close as possible to the partitions associated to each attribute. The closeness of two partitions is evaluated by using a generalization of the classical conditional entropy. A partition (referred to as the median partition) is constructed such that the sum of the dissimilarities between this partition and all the partitions determined by the attributes of the database is minimal. To search more efficiently the large space of possible genetic algorithm is used where the partitions are represented by chromosomes.

III. Conclusion

Data mining is the process of extracting meaningful data from huge amount of database. Clustering plays an important role in data mining. Clustering is kind of unsupervised learning, and deals with finding a structure in a collection of unlabeled data. The ability to discover highly correlated regions of objects becomes desirable when the data set grows. Clustering is an important data mining technique used for grouping of the data. This paper describes various clustering algorithms for available for clustering numeric data and categorical data.

References

[1] J.MacQueen,"Some methods for classification and analysis of multivariate observations, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press Berkeley, CA 1967, pp. 281–297.

[2] Z. Huang,"Extensions to the k-means algorithm for clustering large datasets with categorical values", Data Mining and knowledge Discovery 2, 1998, 2(3), pp. 283-304.

[3] R.Ramakrishnan, T. Zhang, M.Livny, Birch: A efficient data clustering algorithm for very large databases, SIGMOD Record25, 1996, pp. 103–114.

[4] R.Rastogi, S.Guha, K.Shim, Cure: An efficient clustering algorithms for large databases, in: ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998, pp.73–84.

[5] G. Karypis, E. Han, V. Kumar,"Chameleon: Hierarchical clustering using dynamic modeling", IEEE Computer, Vol. 32, No. 8, pp. 68–75, Aug. 1999.

[6] J.S.M.Ester, H.P.Kriegel, X.Xu,"A density-based algorithm for discovering clusters in large spatial database with noise", in: International Conference on Knowledge Discovery in Databases and Data Mining (KDD-96), AAAI Press, Portland, Oregon, 1996, pp. 226–231.

[7] Z. He, X. Xu, S. Deng,"Squeezer: An Efficient Algorithm for Clustering Categorical Data", J. Computer Science and Technology, Vol. 17, No. 5, pp. 611-624, 2002.

[8] P. Andritsos, V. Tzerpos,"Information-Theoretic Software Clustering," IEEE Trans. Software Eng., Vol. 31, No. 2, pp. 150-165, Feb. 2005.

[9] S. Guha, R. Rastogi, K. Shim,"ROCK: A Robust Clustering Algorithm for Categorical Attributes", Information Systems, Vol. 25, No. 5, pp. 345-366, 2000.

[10] V. Ganti, J. Gehrke, R. Ramakrishnan,"CACTUS: Clustering Categorical Data Using Summaries", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.

[11] D. Barbara, Y. Li, J. Couto,"COOLCAT: An Entropy-Based Algorithm for Categorical Clustering", Proc. Int'l Conf. Information and Knowledge Management (CIKM), pp. 582-589, 2002.

[12] D. Gibson, J. Kleinberg, P. Raghavan,"Clustering Categorical Data: An Approach Based on Dynamical Systems", VLDB J., Vol. 8, No. 3-4, pp. 222-236, 2000.

[13] Y. Yang, S. Guan, J. You,"CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 682

[14] M.J. Zaki, M. Peters,"Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques", Proc. Int'l Conf. Data Eng. (ICDE), pp. 355-356, 2005.

[15] D. Cristofor, D. Simovici,"Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms", J. Universal Computer Science, Vol. 8, No. 2, pp. 153-172, 2002.

[16] James C. Bezdek, Robert Ehrlich, William Full, "FCM: The Fuzzy c-Means Clustering Algorithm", Computers & Geosciences Vol. 10, No. 2-3, pp. 191-203, 1984

[17] R. Krishnapuram, A. Joshi, L. Yi,"A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering", in Proc. FUZZ-IEEE, 1999, pp. 1281–1286.

[18] A.K.Jain, M.N.Murty, P.J.Flynn. Data Clustering: A Review. ACM Computing Surveys, 1999, Vol. 31, No. 3, 264-323.