Review Paper On Data Clustering Of Categorical Data

Suchita S. Mesakar
Smt.Bhagwati Chaturvedi
College of Engg,Nagpur

Prof. M. S. Chaudhari
Smt.Bhagwati Chaturvedi
College of Engg,Nagpur

Abstract

Data mining refers to extracting or mining knowledge from large amounts of data. Clustering is the dynamic field of research in data mining.Most of the earlier work on clustering has mainly been focused on numerical data whose inherent geometric properties can be used to naturally define distance functions between data points. Many clustering algorithms are available to group objects having similar characteristics. But these algorithms cannot be applied to categorical data. Many of the algorithms have efficiency issues. This necessitated the development of some algorithms for clustering categorical data. This paper gives an overview of different clustering techniques.

1. Introduction

Data mining refers to extracting or mining knowledge from large amounts of data. The most distinct characteristic of data mining is that it deals with very large data sets. This requires the algorithms used in data mining to be scalable. However, most algorithms presently used in data mining do not scale well when applied to very large data sets because they were initially developed for other applications than data mining which involve small data sets.

Data clustering is one of the fundamental tools available, for understanding the structure of a data set. The process of grouping a set of physical or abstract objects into classes of similar objects is called as clustering. Clustering technique plays an important role in machine learning, data mining, information retrieval, web analysis, marketing, medical diagnostic and pattern recognition. Clustering can be considered the most important unsupervised learning problem it deals with finding a structure in a collection of unlabeled data. Clustering is often called an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given, which is the case in supervised learning .A general definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of data objects that are similar to one another within the sane cluster and are dissimilar to the objects in other clusters.

Clustering is the dynamic field of research in data mining. There exist a large number of clustering algorithms in the literature .The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. The major clustering methods can be categorized into

- Hierarchical Algorithms
- Partitional Algorithms
- Density Based Algorithms
- Grid Based Algorithms

The problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measure between data values.

Many algorithms for clustering are available like K-Means, Fuzzy c-means etc.but these cannot be directly applied for clustering of categorical data, where values are different and have no specific order.

An example of categorical attribute is shape whose values include circle, rectangle, ellipse, etc.Due to the special properties of categorical attributes; the clustering of categorical data seems more complicated than that of numerical data. Many algorithms focus on numerical data whose inherent geometric properties can be used naturally to define distance function between data points. However much of the data existed in the database is categorical where attributes values cannot be naturally ordered as numerical values.

Categorical data has a different structure than the numerical data. 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.

2. Review of Different Clustering Algorithm

Many algorithms are present for clustering data. The well known, K-means clustering has been a very popular technique for partitioning large data sets with numerical attributes. The K-means algorithm is a typical partition-based clustering method which is simple and unsupervised. It is the basic partition algorithm which aims to partition n observations into k clusters in which each observation 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. This algorithm does not work with categorical data.

K-modes, an algorithm extending the k-means paradigm to categorical domain were introduced K-modes [1,2] extend K-means and introduce a new dissimilarity measure for categorical data. The dissimilarity measure between two objects is calculated as the number of attributes whose values do not match. The K-modes algorithm then replaces the means of clusters with modes, using a frequency based method to update the modes in the clustering process to minimize the clustering cost function. K-modes generate local optimal solutions based on the initial modes and the order of objects in the data set. K-modes must be run multiple times with different starting values of modes to test the stability of the clustering solution.

As a single-pass algorithm, Squeezer[3] proposed by Z. He, X. Xu, and S. Deng makes use of a prespecified similarity threshold to determine which of the existing clusters (or a new cluster) to which a data point under examination is assigned. Squeezer algorithm for categorical data clustering clusters the data by giving greater weight to uncommon attribute value matches in similarity computations

P. Andritsos and V. Tzerpos proposed LIMBO[4] which is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuples. LIMBO has the advantage that it can produce clusterings of different sizes in a single execution. The IB framework is used to define a distance measure for categorical tuples.LIMBO handles large data sets by producing a memory bounded summary model for the data.

ROCK:A Robust Clustering Algorithm for Categorical Attributes [5] proposed by S. Guha, R. Rastogi, and K. Shim , is a agglomerative hierarchical clustering algorithm that explores the concept of links for data with categorical attributes.Traditional clustering algorithm for clustering categorical data use distance functions. Distance measure does not lead to high quality clusters when clustering categorical data. Rock algorithm considers the neighbourhood of individual pairs of points.ROCK algorithm starts by assigning each tuple to a separated cluster, and then clusters are merged repeatedly according to the closeness between clusters. The closeness between clusters is defined as the sum of the number of "links" between all pairs of tuples, where the number of "links" is computed as the number of common neighbors between two tuples.

V. Ganti, J. Gehrke, and R. Ramakrishnan proposed CACTUS [6] a fast summarization based algorithm for categorical clustering of data which 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 set of clusters are determined from the set of candidate clusters.

D. Barbara, Y. Li, and J. Couto proposed COOLCAT [7], 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. Given a set of clusters, COOLCAT will place the next point in the cluster where it minimizes the overall expected entropy. COOLCAT acts incrementally, and it is capable to cluster every new point without having to re-process the entire set. Therefore, COOLCAT is suited to cluster data streams.

Gibson et al. [8] proposed an algorithm called STIRR (Sieving Through Iterated Relational Reinforcement), a generalized spectral graph partitioning method forcategorical data. STIRR is an iterative approach, which maps categorical data to non-linear dynamic systems. If the dynamic system converges, the categorical data can be clustered.

CLOPE [9] proposed by Kaufman and Rousseeuw is a novel algorithm for categorical data clustering. CLOPE is proposed based on the intuitive idea of increasing the height-to-width ratio of the cluster histogram. The idea is generalized with a repulsion parameter that controls tightness of transactions in a cluster, and thus the resulting number of clusters. 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 [10] 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; no valid cluster is missed. It also merges overlapping cliques to report more meaningful clusters. It imposes no domain constraints and is scalable to high dimensions. CLICKS outperforms for high-dimensional datasets.

Krishnapuram et al. [11] have proposed a fuzzy clustering algorithm around medoid known as fuzzy c-medoids (FCMdd). The objective of the algorithm is based on selecting c representative objects (medoids) from the data set in such a way that the total dissimilarity within each cluster is minimized.

3. Conclusion

Clustering is the dynamic field of research in data mining. The ability to discover highly correlated regions of objects becomes desirable when the data set grows**.** In this paper, detailed literature about various data clustering algorithm for categorical data is mentioned. Various data clustering techniques have its own advantages and disadvantages.

References

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

[2]Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 1-8, 1997.

[3]Z. He, X. Xu, and S. Deng, “Squeezer: An Efficient Algorithm for Clustering Categorical Data,” J. Computer Science and Technology, vol. 17, no. 5, pp. 611-624, 2002

[4]P. Andritsos and V. Tzerpos, “Information-Theoretic Software Clustering,” IEEE Trans. Software Eng., vol. 31, no. 2, pp. 150-165, Feb. 2005.

[5] S. Guha, R. Rastogi, and K. Shim, “ROCK: A Robust Clustering Algorithm for Categorical Attributes,” Information Systems, vol. 25, no. 5, pp. 345-366, 2000.

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

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

[8] D. Gibson, J. Kleinberg, and P. Raghavan, “Clustering Categorical Data: An Approach Based on Dynamical Systems,” VLDB J., vol. 8, nos. 3-4, pp. 222-236, 2000.

[9] Y. Yang, S. Guan, and 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

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

[11] R. Krishnapuram, A. Joshi, and 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.

[12] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Brussels, Belgium: Wiley, 1990.

[13] D. Cristofor and D. Simovici, “Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms,” J. Universal Computer Science, vol. 8, no. 2, pp. 153-172, 2002.