IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 04, Issue 04 (April. 2014), ||V6|| PP 34-37

Different techniques designs for clustering of data, specially categorical data: An Overview

Kusha Bhatt [1], Pankaj Dalal [2]
[1] *(M.Tech Scholar, Software Engineering, Shrinathji Institute of Technology & Engg., Nathdwara)*
[2] *(Associate Professor,Deptt. Of CSE, Shrinathji Institute of Technology & Engg., Nathdwara)*

Abstract: - Mining of knowledge from a pool of data is known as data mining. One of the research potent field of research in area of data mining is clustering. The earlier work in clustering is mainly concentrated on quantitative and ordinal data or we can in general say numeric data which have an inherited order. But application of these algorithms on categorical data is not possible as these data items do not have any inherited order. Therefore it becomes necessary to have algorithms that can solve the problem of categorical data clustering. In this paper, an overview of various categorical data clustering techniques has been given.

Keywords: - *Clustering, categorical data, K-means, K-modes*

I. INTRODUCTION

The unsupervised learning that aims at partitioning a data set into groups of similar items is known as Clustering. The goal is to create clusters of data objects where the within-cluster similarity is maximized (intra-cluster similarity) and the between-cluster similarity is minimized (inter-cluster similarity). One of the stages in a clustering task is selecting a clustering strategy. A particular clustering algorithm is selected that is suitable for the data and the desired clustering type at this stage. Selecting a clustering algorithm is not an easy task and requires the consideration of several issues such as data types, data set size and dimensionality, data noise level, type or shape of expected clusters, and overall expected clustering quality. Many clustering algorithms have been proposed that employ a wide range of techniques such as iterative optimization, probability distribution functions, density-based concepts, information entropy, and spectral analysis in past few decades.

A. An Overview of Clustering

Here we are providing a small exploration of the elements that are to be taken under consideration when designing clustering algorithms. The elements are data types, proximity measures, and objective functions, the description of the elements is provided below.

Data Types

The value of an attribute can be classified as follows:

- Quantitative. These attributes contain continuous numerical quantities where a natural order exists between items of the same data type and an arithmetically based distance measure can be defined.
- Qualitative. These attributes contain discrete data whose domain is finite. We refer to the items in the domain of each attribute as categories. Qualitative data are further subdivided as:
- Nominal. Data items belonging to this group do not have any inherent order or proximity. We refer to these data as categorical data.
- Ordinal. These are ordered discrete items that do not have a distance relation
- Binary attributes are attributes that can take on only two values: 1 or 0.

Depending on the context, binary attributes can be qualitative or quantitative.

Proximity Measures

Proximity measures are metrics that define the similarity or dissimilarity between data objects for the purpose of determining how close or related the data objects are. To define proximity measure there are various approaches. Depending upon the data types and application area these approaches varies. For most algorithms, these proximity measures are used to construct a proximity matrix that reflects the distance or similarity between the data objects. These matrices are used as input for a clustering algorithm that clusters the data according to a partitioning criterion or an objective function.

B. Clustering Stages

Clustering cannot be a one-step process. In one of the seminal texts on Cluster Analysis, we can divide the clustering process in the following stages [2].

- ***Data Collection***
- ***Initial Screening***
- ***Representation***
- ***Clustering Tendency***
- ***Clustering Strategy***
- ***Validation.***
- ***Interpretation***

II. TYPES OF CLUSTERING

Types of clustering and algorithms.

Traditionally clustering is divided into two basic types of clustering:

- Hierarchical clustering methods
- Portioning methods.

Hierarchical clustering algorithms consist of assigning the objects to their own clusters and then pair of clusters is repeatedly merged until the whole tree is formed. [5].Some such algorithms are BIRCH [4], CURE [6] etc. BIRCH [4] is a data clustering method which is the first to handle noise (data points that are not part of an underlying pattern).It addresses the problems of large databases and minimization of I/O. BIRCH [4] addresses metric attributes. CURE [6] identifies clusters which are non-spherical in shape and which have wide variance in size. It employs the technique of random sampling and partitioning that allows it to handle large data sets. Partitional clustering decomposes the given data set into disjoint clusters. The common example for partitional clustering algorithm is K-mean. All the above algorithms are mainly used with numeric data.

III. REVIEW OF CATEGORICAL DATA CLUSTERING ALGORITHMS

A. Algorithms used earlier

Some of the algorithm which are used for clustering categorical data are ROCK [6], K-Modes, STIRR [7], CACTUS[8], COBWEB [9],and EM[10]. ROCK [6] proposes a novel concept of links to measure the similarity/proximity between a pair of data points. It employs links and not distances when merging clusters. ROCK [6] a robust clustering algorithm for categorical attributes. STIRR [7] it is based on an iterative method for assigning and propagating weights on the categorical values in a table; this facilitates a type of similarity measure arising from the co-occurrence of values in the dataset. CACTUS [8] introduces a normal formalization of a cluster for categorical attributes by generalizing the cluster definition for numerical attributes. CACTUS [8] clusters categorical attributes whose attribute values do not have any natural ordering. In this algorithm there is no post processing of the final output generated by the user to define the clusters. COBWEB [9] is an incremental clustering algorithm, based on probabilistic categorization trees. The search for a good clustering is guided by a quality measure for partitions of data. The algorithm reads unclassified examples, given in attribute-value representation. EM algorithm [10] is an iterative clustering technique. It starts with an initial clustering model for a data and iteratively refines the model to fit the data better. After indeterminate number of iterations it terminates at an optimal solution. K-modes [11] algorithm extends K-means to categorical domain. The K-modes method have three major modifications to K-means, it uses different dissimilarity measures. Replaces means with modes and uses a frequency based method to update modes. This algorithm defines the clusters based on how close they are to the centroid of cluster, distances between centroid of clusters is a poor measure of similarity between them.

B. Recently used algorithms

Tiejun et al. proposed a clustering algorithm processing data with mixed attributes, which aims to improve the speed and efficiency by means of restricting the searching direction of particles warm in the gradient space. Compared with the traditional algorithm processing data with mixed attributes such as K-prototype and K-mode, it is an unsupervised self-learning process. Furthermore it doesn’t need to pre-set parameter according to priori experience, which makes its results more directionally significant [12]. **Tripathy et al.** proposed an algorithm called SSDR which is more efficient than most of the earlier algorithms including MMR, MMeR and SDR, which are recent algorithms developed in this direction. It handles uncertain data using rough set theory. This takes both the numerical and categorical data simultaneously besides taking care of uncertainty. Also, this algorithm gives better performance while tested on well-known datasets [13]. **Reddy et al.** presented a clustering algorithm based on similarity weight and filter method paradigm that works well for data with mixed numeric and categorical features. They propose a modified description of cluster center to overcome the numeric data only limitation and provide a better characterization of clusters [14]. **Khan et al.** proposed an algorithm to compute fixed initial cluster centers for the K-modes clustering algorithm that exploits a multiple clustering approach that determines cluster structures from the attribute values of given attributes in a data. This algorithm is based on the experimental observations that some of the data objects do not change cluster membership irrespective of the choice of initial cluster centers and individual attributes may provide some information about the cluster structures [15]. **Chen et al.** Presented a method for performing multidimensional clustering on categorical data and show its superiority over unidimensional clustering. Complex data sets such as the ICAC data can usually be meaningfully partitioned in multiple ways, each being based on a subset of the attributes. Unidimensional clustering is unable to uncover such partitions because it seeks one partition that is jointly defined by all the attributes. They proposed a method for multidimensional clustering, namely latent tree analysis [16]. **Cao et al.** presented a weighting k-Modes algorithm for subspace clustering of categorical data and its corresponding time complexity is analysed as well. In this algorithm, an additional step is added to the k-Modes clustering process to automatically compute the weight of all dimensions in each cluster by using complement entropy. Furthermore, the attribute weight can be used to identify the subsets of important dimensions that categorize different clusters [17]. **Chen et al.** proposed a novel kernel-density-based definition using a Bayes-type probability estimator. Then, a new algorithm called k-centers is proposed for central clustering of categorical data, incorporating a new feature weighting scheme by which each attributes automatically assigned with a weight measuring its individual contribution for the clusters [18].

IV. CONCLUSION

In data mining clustering is the dynamic field of research. When the size of data set grows it is desirable to have algorithms which are able to discover highly correlated regions of objects. In this paper, we have discussed various clustering algorithm for categorical data. The algorithms mentions in this paper have their own advantages and disadvantages. After having a brief overview of the available techniques of clustering of categorical data we can try to develop a technique which emphasis on cluster center initialization for better results of clustering. This may be taken as the future scope of the paper.

REFERENCES

[1] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

[2] Anil K. Jain and Richard C. Dubes, “Algorithms for Clustering Data”, Prentice-Hall, 1988.

[3] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis, “Fundamentals of Data Warehouses”, Springer, 1999.

[4] Zhang, T., Ramakrishnan, R. and Livny, M., “BIRCH: An efficient data clustering method for very large databases”, International conference on Management of Data, Montreal, Canada, pp. 103 –144, June 1996.

[5] Ying Zhao and George Karypis, “Evaluation of Hierarchical Clustering Algorithms for DocumentDatasets” CIKM, 2002.

[6] Guha, S., Rastogi, R. and Shim, K., “Cure: An efficient clustering algorithm for large databases”, International conference on Management of Data, Seattle WA, pp. 73-83, June 1998.

[7] Gibson, D., Kleinberg, J. and Raghavan, P., “Clustering categorical data: an approach based ong dynamical systems”, Proceedings of the 24th Annual International Conference on Very Large Databases, VLDB Journal vol. 8, no.3-4, pp. 222-236, 2000.

[8] Ganti, V., Gehrke, J. and Ramakrishnan, R., “CACTUS – Clustering Categorical Data using Summaries”, Proceedings of the Fifth ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, San Diego, CA USA, pp. 73-83, Aug. 1999.

[9] Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: A new dataclustering algorithm and its applications”, Data Mining and Knowledge Discovery, pp. 141--182, 1997.

[10] Liu, C., and Sun, D. X. “Acceleration of EM algorithm for mixtures models using ECME”, ASA Proceedings of The Stat. Comp. Session. Pp. 109-114, 1997.

[11] Huang, Z., “Clustering large data sets with mixed numeric and categorical values”, Proceedings of First Pacific-Asia Conference on Knowledge Discovery & Data Mining, pp. 21-34, 1997

[12] Zhang Tiejuna, Yang Jinga, Zhang Jianpei, "Clustering Algorithm for Mixed Attributes Data Based on Restricted Particle Swarm Optimization", International Conference on Computer Science and Information Technology (ICCSIT), vol. 51, 2011.

[13] B. K. Tripathy and Adhir Ghosh, "SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory", Pelagia Research Library, Advances in Applied Science Research, pp. 314-326, 2011.

[14] M. V. Jagannatha Reddy and B. Kavitha, "Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method", International Journal of Database Theory and ApplicationVol. 5,No. 1, March, 2012.

[15] Shehroz S. Khan, Amir Ahmady, "Cluster Center Initialization for Categorical Data Using Multiple AttributeClustering", 3rdMultiClust Workshop: Discovering, Summarizing and Using Multiple Clustering, SIAM International Conference on Data Mining, Anaheim, California, USA, 2012.

[16] Tao Chen, Nevin L. Zhang, Tengfei Liu, Kin Man Poon, Yi Wang, "Model-based multidimensional clustering of categorical data", Artificial Intelligence, Published by Elsevier Ltd., pp. 2246–2269, Oct. 2012.

[17] Fuyuan Cao, Jiye Liang, Deyu Li, Xingwang Zhao, "A weighting k-Modes algorithm for subspace clusteringof categorical data", Neurocomputing Journal, Elsevier Science Publishers, Vol. 108, Pages 23–30, May, 2013.

[18] Lifei Chen, Shengrui Wang, "Central Clustering of Categorical Datawith Automated Feature Weighting"Proceedings of the 23rdInternational Joint Conference on Artificial Intelligence, Beijing, China, Aug. 2013.