International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244

CLUSTERING CATEGORICAL DATA USING ROUGH SETS: A REVIEW

Dr. Jyoti*

Abstract: *In the present day scenario, there are large numbers of clustering algorithms available to group objects having similar characteristics. But the implementations of many of those algorithms are challenging when dealing with categorical data. While some of the algorithms available at present cannot handle categorical data, the others are unable to handle uncertainty. Many of them have the stability problem and also have efficiency issues. This paper provides a review of the various algorithms that try to cluster the categorical databases.*

Keywords: *Categorical data, Clustering, Uncertainty, Fuzzy sets, Rough sets.*

***Asst. Prof. (CE), YMCA University of Science and Technology, Faridabad, Haryana, India

I. INTRODUCTION

The main objective of clustering is to group data or objects having the similar characteristics in the same cluster and having dissimilar characteristics in separate 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

Clustering as a technique has been used in various data mining tasks such as unsupervised classification and data summation. It is also used in segmentation of large heterogeneous data sets into smaller homogeneous subsets which is easily managed, separately modeled and analyzed [1]. The basic goal in cluster analysis is to discover natural groupings of objects [2]. Clustering techniques are used in many areas such as manufacturing, medicine, nuclear science, radar scanning and research and also in development.

Various researchers have proposed algorithms that primarily operate on numerical datasets. These are those datasets whose attributes belong to the numerical domain. For example, Wu et al. [3] developed a clustering algorithm specifically designed for handling the complexity of gene data. Jiang et al. [4] analyze a variety of cluster techniques, which can be applied for gene expression data. Wong et al. [5] presented an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET). Haimov et al. [6] used cluster analysis to segment radar signals in scanning land and marine objects. Mathieu and Gibson [7] used the cluster analysis as a part of a decision support tool for large scale research and development planning to identify programs to participate in and to determine resource allocation. The basic reason for dealing with numerical attributes is the fact that one can easily define similarity on them and hence are easy to handle. This, similarity can be defined as common objects, common values for the attributes and the association between two. In such cases horizontal co-occurrences (common value for the

objects) as well as the vertical co-occurrences (common value for the attributes) can be examined in [3].

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.

II. LITERATURE REVIEW

A. Clustering Algorithms for Categorical Datasets

This section presents the literature for the various existing algorithms for clustering categorical data. Dempster et al. [8] presents a partitional clustering method, called the Expectation-Maximization (EM) algorithm. EM first randomly assigned different probabilities to each class or category, for each cluster. These probabilities were then successively adjusted to maximize the likelihood of the data given the specified number of clusters. Since the EM algorithm computes the classification probabilities, each observation belonged to each cluster with a certain probability. The actual assignment of observations to a cluster was determined based on the largest classification probability. After a large number of iterations, EM terminates at a locally optimal solution. Han et al. [9] proposed a clustering algorithm to cluster related items in a market database based on an association rule hyper-graph. A hyper-graph is used as a model for relatedness. The approach targets binary transactional data. It assumed item sets that defined clusters were disjoint and there was no overlap amongst them. However, this assumption may not hold in practice as transactions in different clusters may have a few common items. K-modes [8] extended K-means and introduced a new dissimilarity measure for categorical data. The dissimilarity measure between two objects was calculated as the number of attributes whose values did not match. The K-modes algorithm then replaced 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. One advantage of K-modes is it is useful in interpreting the results [8]. However, 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. Ralambondrainy [10] proposed a method to convert multiple category attributes into binary attributes using 0 and 1 to represent either a category absence or presence, and to treat the binary attributes as numeric in the K-means algorithm. Huang [8] also proposds the K-prototypes algorithm, which allows clustering of objects described by a combination of numeric and categorical data. CACTUS (Clustering Categorical Data Using Summaries) [11] is a summarization based algorithm. In CACTUS, the authors clustered categorical data by generalizing the definition of a cluster for numerical attributes. Summary information constructed from the data set was assumed to be sufficient for discovering well-defined clusters. CACTUS finds clusters in subsets of all attributes and thus performed a subspace clustering of the data. Guha et al. [12] proposed a hierarchical clustering method termed ROCK (Robust Clustering using Links), which can measure the similarity or proximity between a pair of objects. Using ROCK, the number of "links" was computed as the number of common neighbors between two objects. An agglomerative hierarchical clustering algorithm is then applied which was

- First, the algorithm assigned each object to a separate cluster.
- Clusters were then merged repeatedly according to the closeness between clusters, where the closeness was defined as the sum of the number of "links" between all pairs of objects.

Gibson et al. [4] proposed an algorithm called STIRR (Sieving Through Iterated Relational Reinforcement), a generalized spectral graph partitioning method for categorical data. STIRR was an iterative approach, which mapped categorical data to non-linear dynamic systems. If the dynamic system converges, the categorical data can be clustered. Clustering naturally lends itself to combinatorial formulation. However, STIRR required a non-trivial post-preprocessing step to identify sets of closely related attribute values [11]. Additionally, certain classes of clusters were not discovered by STIRR [11]. Moreover, Zhang et al. [14] argued that STIRR cannot guarantee convergence and therefore proposed a revised dynamic system algorithm that assured convergence. He et al. [15] proposed an algorithm called

"Squeezer", which was a one-pass algorithm. Squeezer puts the first-tuple in a cluster and then the subsequent-tuples were either put into an existing cluster or rejected to form a new cluster based on a given similarity function. He et al. [16] explored categorical data clustering "CDC" and link clustering "LC" problems and proposed a "LCBCDC" (Link Clustering Based Categorical Data Clustering), and compared the results with Squeezer and K-mode.

In reviewing these algorithms, some of the methods such as STIRR and EM algorithms cannot guarantee the convergence while others have scalability issues. In addition, all of the algorithms have one common assumption which is "Each object can be classified into only one cluster and all objects have the same degree of confidence when grouped into a cluster [17]". However, in real world applications, it is difficult to draw clear boundaries between the clusters. Therefore, the uncertainty of the objects belonging to the cluster needs to be considered.

B. Clustering algorithms for categorical datasets with uncertainty amongst the objects

One of the first attempts to handle uncertainty is fuzzy K-means [18]. In this algorithm, each pattern or object was allowed to have membership functions to all clusters rather than having a distinct membership to exactly one cluster. Krishnapuram and Keller [19] proposed a probabilistic approach to clustering in which the membership of a feature vector in a class had nothing to do with its membership in other classes and modified clustering methods were used to generate membership distributions. Krishnapuram et al. [20] presented several fuzzy and probabilistic algorithms to detect linear and quadratic shell clusters.

It is important to note here that some of the initial work in handling uncertainty was based on numerical data. Huang [8] proposed a fuzzy K-modes algorithm with a new procedure to generate the fuzzy partition matrix from categorical data within the framework of the fuzzy K-means algorithm. The method found fuzzy cluster modes when a simple matching dissimilarity measure is used for categorical objects. By assigning confidence to objects in different clusters, the core and boundary objects of the clusters could be decided. This helped in providing more useful information for dealing with boundary objects. More recently, Kim et al. [21] have extended the fuzzy K-modes algorithm by using fuzzy centroid to represent the clusters of categorical data instead of the hard-type centroid used in the fuzzy K-modes algorithm. The use of fuzzy centroid made it possible to fully exploit the

power of fuzzy sets in representing the uncertainty in the classification of categorical data. However, fuzzy K-modes and fuzzy centroid algorithms suffered from the same problem as K-modes, i.e. they required multiple runs with different starting values of modes to test the stability of the clustering solution. In addition, these algorithms had to adjust one control parameter for membership fuzziness to obtain better solutions. This necessitated the effort for multiple runs of these algorithms to determine an acceptable value of this parameter. Therefore, there was a need for a categorical data clustering method, having the ability to handle uncertainty in the clustering process while providing stable results.

One methodology with potential for handling uncertainty is Rough Set Theory (RST) which has received considerable attention in the computational intelligence literature since its development by Pawlak in the 1980s. Unlike fuzzy set based approaches, rough sets have no requirement on domain expertise to assign the fuzzy membership. Still, it may provide satisfactory results for rough clustering. In 2007, an algorithm, termed MMR was proposed [22], which used the rough set theory concepts to deal with the above problems in clustering categorical data. Later in 2009, this algorithm was further improved to develop the algorithm MMeR [23] and it could handle hybrid data. Again, very recently in 2011 MMeR was again improved to develop an algorithm called SDR [24], which can also handle hybrid data. The last two algorithms can handle both uncertainties as well as deal with categorical data at the same time but SDR has more efficiency over MMeR and MMR. Again in 2011, authors of [25] improved the SDR algorithm and presented SSDR. This took both the numerical and categorical data simultaneously besides taking care of uncertainty.

Till the development of MMR, the only algorithms which aimed at handling uncertainty in the clustering process were based upon fuzzy set theory [26].These algorithms based on fuzzy set theory included fuzzy K-modes, fuzzy centroids. The K-modes algorithm replaced the means of the clusters (K-means) with modes and uses a frequency based method to update the modes in the clustering process to minimize the clustering cost function. Fuzzy K-modes generates a fuzzy partition matrix from categorical data. By assigning a confidence to objects in different clusters, the core and boundary objects of the clusters were determined for clustering purposes. The fuzzy centroids algorithm uses the concept of fuzzy set theory to derive fuzzy centroids to create clusters of objects which have categorical attributes. But in MMR, MMeR and in SDR, authors have used rough sets concept to build

those algorithms but as compared to efficiency MMeR is more efficient than MMR and less efficient than SDR but SSDR is much more efficient than other.

III. 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 algorithms for categorical data is mentioned. This also includes algorithms that have considered categorical data in the shades of uncertainty. Various data clustering techniques have their own advantages and disadvantages.

REFERENCES

[1] Z. Huang, Data Mining and Knowledge Discovery 2 (3), 283-304,1998.

[2] R. Johnson, W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, New York, 2002.

[3] S. Wu, A. Liew, H. Yan, M. Yang, “Cluster analysis of gene expression data based on self-splitting and merging competitive learning”,IEEE Transactions on Information Technology in BioMedicine 8 (1),5-15,2005.

[4] D. Jiang, C. Tang, A. Zhang IEEE Transactions on Knowledge and Data Engineering 16 (11), 1370-1386, 2004.

[5] K. Wong, D. Feng, S. Meikle, M. Fulham, IEEE Transactions on Nuclear Science 49 (1), 200-207, 2002.

[6] S. Haimov, M. Michalev, A. Savchenko, O. Yordanov, IEEE Transactions on Geo Science and Remote Sensing 8 (1), 606–610, 1989.

[7] R. Mathieu, J. Gibson, IEEE Transactions on Engineering Management 40 (3), 283-292,2002.

[8] Z. Huang, Data Mining and Knowledge Discovery 2 (3), 283-304, 1998.

[9] E. Han, G. Karypis, V. Kumar, B. Mobasher, Clustering based on association rule hypergraphs, in: Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 9-13,1997.

[10] H. Ralambondrainy, Pattern Recognition Letters 16 (11) (1995) 1147-1157.

[11] V., Ganti, J. Gehrke, R. Ramakrishnan, CACTUS – clustering categorical data using summaries, in Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 73-83,1999.

[12] S. Guha, R. Rastogi, K. Shim, Information Systems 25 (5) (2000) 345-366.

[13] D. Gibson, J. Kleinberg, P. Raghavan, The Very Large Data Bases Journal 8 (3-4), 222–236, 2002.

[14] Y. Zhang, A. Fu, C. Cai, P. Heng, Clustering categorical data, in: Proceedings of the 16th International Conference on Data Engineering, pp. 305–324, 2000.

[15] Z. He, X. Xu, S. Deng, Journal of Computer Science & Technology, 17 (5), 611-624, 2002.

[16] Z. He, X. Xu, S. Deng, A link clustering based approach for clustering categorical data, Proceedings of the WAIM Conference, (2004).

[17] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Journal of Intelligent Information Systems 17 (2-3), 107–145, 2001.

[18] E. Ruspini, Information Control 15 (1), 22–32, 1969.

[19] R. Krishnapuram, J. Keller, IEEE Transactions on Fuzzy Systems 1 (2), 98–110, 1993.

[20] R. Krishnapuram, H. Frigui, O. Nasraoui, IEEE Transactions on Fuzzy Systems 3 (1), 29–60, 1995.

[21] D. Kim, K. Lee, D. Lee, Pattern Recognition Letters 25 (11), 1263–1271.Mkm, 2004.

[22] D Parmar, Teresa Wu, Jennifer B, Data & Knowledge Engineering (2007).

[23] B.K.Tripathy and M S Prakash Kumar, in the Proc. of International Journal of Rapid Manufacturing (special issue on Data Mining) (Switzerland),vol.1, no.2, pp.189-207,2009.

[24] Tripathy, B.K. and A.Ghosh, "SDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory", Communicated to the International IEEE conference to be held in Kerala, 2011.

[25] B. K. Tripathy and A. Ghosh, "SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory", in the Proc. of Advances in Applied Science Research, 2 (3): 314-326, 2011.

[26] E. Han, G. Karypis, V. Kumar, B. Mobasher, Clustering based on association rule hypergraphs, in: Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 9-13, 1997.