Review of Clustering Algorithm for Categorical Data

Poonam M. Bhagat, Prasad S. Halgaonkar, Vijay M. Wadhai

Abstract: ***Clustering is a partition of data into a group of similar or dissimilar data points and each group is a set of data points called clusters. Clustering is an unsupervised learning with no predefined class label for the data points. Clustering is considered an important tool for data mining. Clustering has many applications such as pattern recognition, image processing, market analysis, World Wide Web and many others. Categorical data are groups of categories and each value represents some category. The problem of clustering categorical data is solved by the use of the cluster ensemble approach, but this technique generates a final data partition with imperfect information. The ensemble-information matrix that is the binary cluster association matrix content presents only cluster-data point relations with many entries being left unknown and which decrease the quality of the whole data partition. To avoid the degradation of the final data partition, a new approach of link-based is presented which includes the refined cluster association matrix. It maintains cluster to cluster relation and helps to improve quality of the final data partition result by determining the unknown entries through measuring similarity between clusters in an ensemble. The cluster ensemble combines multiple data partitions from different clustering algorithms into a single clustering solution to improve the robustness, accuracy and quality of the clustering result.***

Index Terms- Clustering, categorical, link-based, ensemble

I. INTRODUCTION

Clustering is a division of data into a group of data points, similar data points are in one group called cluster and dissimilar data points are in another cluster. The Fig.1 shows clustering, in which identify the three clusters into which the data can be divided. Here is the similarity criterion is distance two or more objects or data points belong to the same cluster if they are close according to a given distance then this is called distance-based clustering. While choosing any clustering algorithm some important requirements are required like:

- Robust
- Scalability
- Capability to deal with different types of attributes
- Handling outliers and noise
- High Dimensionality
- Usability and Compatibility

Categorical data is a collection of categories and each value represents some category, categorical data is also called as qualitative data which in the form of unordered manner. Categorical data further classified into two types that are nominal and ordinal. Nominal means related to names in which data points are in unordered categories such as marital status, hair color. Ordinal in which order is essential such as exam rank.

Manuscript received December, 2013

Poonam M. Bhagat, Department of Computer Engg. MITCOE-Pune University, India

Prasad S. Halgaonkar, Department of Computer Engg. MITCOE-Pune University, India

Vijay M. Wadhai, Principal MITCOE-Pune University, India

The idea of a cluster varies between algorithms and is one of the many decisions to take when choosing the appropriate algorithm for a particular problem. The clusters found by different algorithms have varied a lot in their belongings, and on the basis of belongings it helps to understand these various clusters differences between the various algorithms. The clustering is mainly used in data compression in image processing. Clustering of categorical data is a tremendously difficult task if the number of items or attributes involved increase.

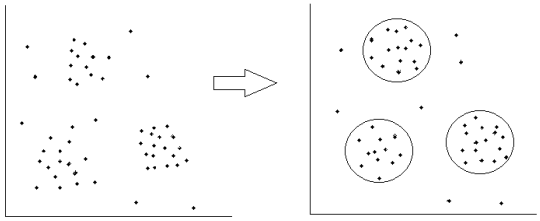

Fig. 1 Clustering

There are some issues with the existing clustering technique are mentioned as follows:

- Current clustering techniques do not refer all the requirements effectively.
- Dealing with a large number of dimensions and large number of data items can be difficult because of time complexity;
- The distance base clustering efficiency mainly depends on the definition of distance.
- Clustering algorithm final clustering results can be deduced in distinct ways.

There are various algorithms are introduced for clustering categorical data these algorithms are available for clustering categorical data but no single algorithm can achieve the best result for all the data sets. Many different clustering algorithms for categorical data are found to solve the problem from a different perspective that is based on the idea of co-occurrences between attributes and pairs defining a cluster and subspace algorithm locates clusters in different subspaces of the data set. It is a difficult task to cluster large amount of data to find a suitable partition in an unsupervised learning. Without any prior knowledge trying to maximize the similarity of objects belonging to the same cluster and minimizing the similarity among objects in different clusters. Each algorithm has its own advantages and disadvantages. These algorithms are executed with the specific data set and with the different or the same algorithm with the distinct parameters obtain diverse results. So it is difficult to decide that which algorithm works well. To avoid such confusion of algorithm that which algorithm would be good for available data set and to overcome the limitations of the algorithm there is a new approach of cluster ensemble which efficient result and it also improves the quality of the result. There is a detailed description of Cluster Ensemble method is given as follows: The clustering ensemble has emerged as a prominent method for improving the accuracy of unsupervised learning. It combines multiple

partitions of dataset generated by different clustering algorithms into a single clustering solution; the efficiency and accuracy of this method achieve using on a consensus function. Cluster ensembles are collections of clustering, which are all of the same kind of a set of data objects. Such ensembles can be obtained, for example, by varying the parameters of a base clustering algorithm by employing several different base clusters. The following Fig. 2 shows the cluster ensemble method. Every clustering ensemble method is made up of two steps:

• Generation
• Consensus Function

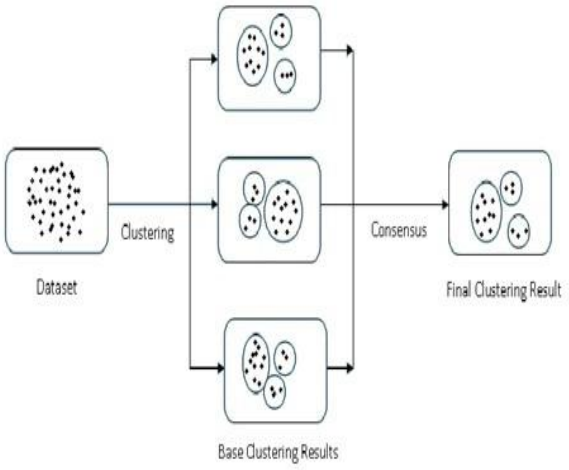

Fig. 2: Cluster Ensemble Method

The description of the first step in clustering ensemble methods is generation; in this step the set of clustering that will be combined is generated. In a particular problem, it is important to apply an appropriate generation process, because the final result will be conditioned by the initial clustering obtained in this step. There are various clustering ensemble methods, all the partitions should be obtained by applying the algorithm with different initializations for the number of clusters. In a general, in the generation step there are no constraints about how the partitions must be obtained. Therefore, in the generation process different clustering algorithms or the same algorithm with different parameter initialization can be applied. Even different object representations, different subsets of objects or projections of the objects on different subspaces could be used. In the generation step the weak clustering algorithms are also used. These algorithms make up a set of clustering using very simple and fast procedures. The weak clustering algorithms are capable of producing high quality consensus clustering in conjunction with a proper consensus function. There are several ensemble generation methods are available are as follows:

Direct ensemble [1], in which considering each value as a cluster in an ensemble. Categorical data can be directly transformed to cluster ensemble without applying any base clustering.

Homogeneous ensembles, in which base clusters are created using recursive runs for the single clustering algorithm such as k-means clustering algorithm.

Full space ensemble, in this methods base clustering are created from the data set .To select the number of clusters in each base clustering there are two methods that are used:

• Fixed k, in which numbers of clusters are fixed for each cluster ensemble.
• Random-k, in which randomly selecting the number of clusters for each cluster ensemble.

Data Subspace[2], it also creates a base cluster to generate a cluster ensemble from the different subset of the dataset. Heterogeneous ensembles, in which also base clusters are used to generate cluster ensembles by using different numbers of clustering algorithm. The next step is the consensus function is an important step in any clustering ensemble algorithm. To obtain the cluster ensemble at various forms of the consensus function has been used and obtains the final clustering result.

The consensus function makes use of the specific information matrix such as binary cluster association matrix (BM) in which the unknown data point belonging referred as 0 and known data point belonging refereed as 1. There are various consensus methods are available such as, Feature based method [3], in which each base cluster gives a cluster label depict as a new data point and make use of it in the final clustering result. Direct method, it depends on the basis of relabeling of ensemble member and searching for the new data partition that has the best match with the all other ensemble members.

Pairwise similarity [4], in which it creates matrix, containing the pairwise similarity. Graph based method , it uses the graph representation to get the cluster ensemble, in which the graph partitioned in to number of equal sized partitions to obtain final clustering results and graph corresponds to the similarity between data points is created from a pairwise similarity matrix. In this step, there are two approaches for consensus function, the first approach is object concurrence in which the how many times the data point belongs to one cluster of how many times two data points are belong together to the same cluster. Another approach is a median partition in which the finding median partition with respect to the cluster ensemble. These are some causes for why use consensus function:

- For each clustering technique there are possible limitations.
- When there is no knowledge about the number of clusters, it becomes complicated.
- Analysis of the result is complex in some instances.
- An essential problem in cluster analysis is the validation of the clustering results.
- Some algorithms can never invalidate what was done previously.

Which Algorithms are compared? Three distinct algorithms are chosen to explore study and compare them. The algorithms that are chosen are: A Link-Based Cluster Ensemble Approach for Categorical Data Clustering, Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data, ROCK: A Robust Clustering Algorithm for Categorical Attributes. The main reason for selecting these three algorithms is mainly work categorical high dimensional dataset. The remainder of this paper is organized as follows. Section 2 presents the related work of the project where the comparisons of different methods are done. Section 3 includes the discussion and future work. Section 4 draws some conclusions.

II. RELATED WORK

In [5], for clustering categorical data, a novel highly effective link-based cluster ensemble approach to categorical data clustering, this improves the conventional matrix by discovering unknown entries through similarity between clusters in an ensemble. An efficient link-based algorithm is proposed for the underlying similarity assessment. The link based similarity method, simply

measures the similarity among the data points is inappropriate for the large data set. Cluster ensemble approach to categorical data set mainly depends on the matching of similarity of among the neighbouring data points and binary cluster association matrix (BM) [6], which represents the ensemble information.

In binary matrix many entries are left being unknown and consider as 0. Due to this quality of the final clustering result not much better it degrades the result. To avoid such problems, a link based cluster ensemble approach is introduced. A link based approach find out the unknown values to improve accuracy of the final clustering result. The following Fig. 3 shows the link-based cluster ensemble approach.

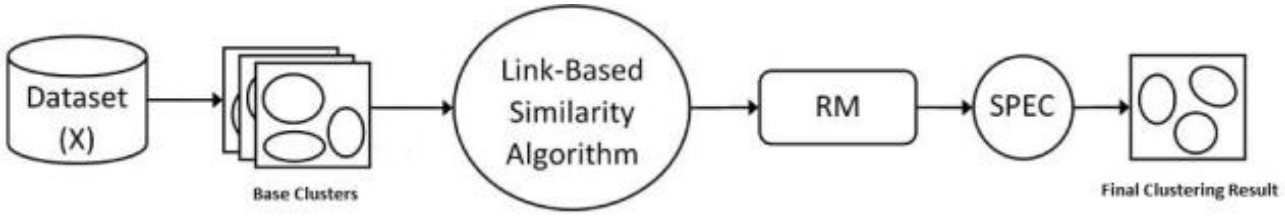

Fig. 3 Link-based cluster ensemble approach

The link based cluster ensemble process:

- Creating a base cluster of dataset to produce cluster ensemble,
- Using link based similarity algorithm and generate refined cluster association matrix,
- Finally generating final data partition clustering result.

Generating Refined Matrix (RM): The Refined cluster association matrix (RM) is an enhanced version of the Binary Matrix (BM) .In BM the unknown values are referred with the zeros (0) and known values referred to the one (1) but due to this association of references left larger unknown values in the clustering and its effects on the final clustering result. But RM is an enhanced version of the BM, it refers the known values as the one (1) and the unknown values are estimated as it measures the similarity between cluster labels which corresponding to a specific cluster of the clustering to which value belongs. Applying consensus function to RM: To obtain the final clustering result, refined cluster association matrix (RM) utilizes a graph based partitioning method.

The consensus function requires the basic original matrix to be initially transformed into a weighted bipartite graph. Given RM representing the relations between N data points and P clusters in an ensemble ,a weighted graph G=(V,W) where V is a set of vertices representing data points as well as clusters and W represents weighted edges. It transforms the original categorical data matrix to an information-preserving numerical variation to which an effective graph partitioning technique can be directly used. The problem of creating the refined matrix that is RM is sorted out by the similarity between categorical clusters, using the Weighted-Triple-Quality(WTQ)[7],[8]similarity algorithm.

The proposed link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and benchmark cluster ensemble techniques. Experimental results on multiple real data sets suggest that the proposed link-based method almost always outperforms both conventional clustering algorithms for categorical data and well known cluster ensemble techniques. The main advantage is, it obtains more accurate, finer and also improves the quality of final data partition clustering result. It's applicable for the large dataset. But its main drawback is time complexity is high.

In [9], Automatic Top-Down Clustering (AT-DC), which is a fully-automatic, parameter- free approach to clustering high-dimensional categorical data. The main idea of the approach is inspired from the top-down approach to decision-tree learning, which recursively partitions the existing data set on the basis of the increases transparency of the subsets with respect to the original data set. The technique supports to iterative procedure which works in a two stages, which attempts to improve the overall quality of the whole data partition. In the first stage, cluster allocations are given, and a new cluster is added to the data partition by identifying and splitting a low-grade cluster. In the second stage, the number of clusters is fixed, and an attempt to optimize cluster allocations is done. The advantage of this approach is it can efficiently and quickly search large amounts of high dimensional categorical data. It achieves optimal clustering result and improves overall quality of the final data partition these are advantages of this method.

In [10], a new concept of links to measure the similarity between a pair of data points with categorical attributes. ROCK (RObust Clustering using linKs), algorithm using links an adaptation of an agglomerative hierarchical clustering algorithm, proposes a new concept of neighbours and links to measure the similarity between a pair of data points and it uses links and not distances when merging clusters. ROCK employs the information about links between data points when making decisions on the data points to be merged into a single cluster. The following Fig. 4 shows the general idea of the ROCK algorithm.

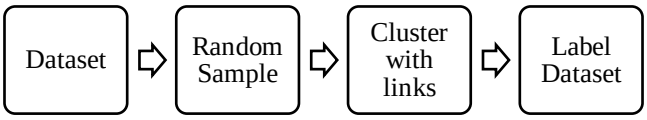

Fig. 4 Overview of ROCK

ROCK optimizes a criterion function defined in terms of the number of links between tuples. The number of links between two clusters is the number of common neighbours they have in the dataset. Starting with each data point in its own cluster, they repetitively merge the two closest clusters till the required number of clusters remains. The complexity of the algorithm is in cubic in the number of tuples available in the dataset, form cluster a sample from the available dataset, and then partition the whole dataset on the basis of clusters from the sample. Experiments on both synthetic and real datasets. It is not applicable for large datasets.

III. DISCUSSION AND FUTURE WORK

There is difficulty in the analysis of categorical data is categorized by the fact that there is no inherent similarity

between attribute values of categorical dataset. The clustering of categorical dataset is fully based on the available dataset. To cluster categorical dataset a link based cluster ensemble approach is used, in which initially the base clusters are created of the available dataset as input by applying the algorithm. From these base clusters a cluster ensemble is created. After that a refined cluster association matrix (RM) [1] is generated from the cluster ensemble using a link based similarity algorithm and finally the final data partition clustering result obtained by consensus function of the spectral graph partitioning algorithm. This system involves two major tasks of generating a cluster ensemble and producing final data partition referred as consensus function.

The main objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy superior to that of any individual clustering. It will generate a finer clustering result than the other ensemble technique and other several similarity algorithms. It also maps cluster to cluster relation which helps to improve final cluster results with fewer entries being left unknown.

To clustering categorical data an efficient link-based cluster ensemble approach is used. It intends to discover and makes use of the relationships between input clustering and it transforms the categorical data matrix to maintain information in the form of refined cluster association matrix (RM). To measure the similarity among clusters of giving ensemble with the help of the refined cluster association matrix (RM). Denoting these clusters as a link of networks and their similarity degrees can be efficiently assessment by link based similarity algorithm. The link-based approach obtains finer clustering results by maintaining cluster to cluster relation than the other several cluster ensemble techniques. The prominent future work includes a new link based similarity algorithm that is Weighted Quad Quality similarity algorithm will be applied to categorical dataset (Mushroom, Soybean, and Congressional Voting). It also provides the cluster to cluster relation which improves the final clustering result and with the less entries being left unknown. Table I presents the advantages and disadvantages of the works in clustering categorical algorithm. Investigate abilities and compare them based on robustness, simplicity and scalability.

TABLE I
SUMMARY OF CLUSTERING CATEGORICAL DATA ALGORITHM

Title	Method	Computational Complexity	Quality Measures	Advantages	Disadvantages
A Link-Based cluster Ensemble Approach for Categorical Data Clustering	• A link-based cluster ensemble approach • WTQ (Weighted - Triple Quality) similarity algorithm to cluster categorical data	• $O(P^2 l^2+NP)$	LCE utilizes three distinct quality measures, • Normalized Mutual Information (NMI) • Classification Accuracy (CA) and • Rand Index (RI)	• It's applicable for large datasets • It achieves accuracy & improves quality of clustering results	• Time and space Complexity is high
Top-Down Parameter Free Clustering of High-Dimensional Categorical Data	• A parameter-free, fully automatic approach based on decision-tree learning • AT-DC (Automatic Top-Down Clustering) to clustering High-Dimensional categorical data	• O(N)	AT-DC utilizes two different quality measures, • Local homogeneity within a cluster and • Global homogeneity of the partition	• It achieves optimal results • It improves overall quality of whole partitions	• No exact-match of conceptual similarity
ROCK: A Robust Clustering Algorithm for Categorical Attributes	• Proposes concept of links • ROCK(Robust Clustering using links) clustering algorithm	• $(O(n^2 \log n))$	ROCK uses to measure the quality or goodness of the cluster , • Criterion function	• It optimizes criterion function	• Not applicable for large dataset

IV. CONCLUSION

In this paper, review the algorithms for clustering categorical data with distinct approaches is done. The LCE approach measures the similarity among the clusters which are given by the ensemble and creates the refined cluster association matrix. The clusters formed as a link network, and then their similarity measure estimate is taken by link-based approach algorithm i.e. Weighted Triple Quality (WTQ) similarity algorithm. The LCE approach achieves better clustering results than the other several clustering algorithms. The ROCK algorithm is also based on link approach to measure the similarity between a pair of data points. But it cannot be applied for large datasets. The AT-DC is fully automatic parameter free approach for clustering categorical data. It achieves better result for clustering categorical data.

REFERENCES

[1] Z. He, X. Xu, and S. Deng , "A Cluster Ensemble Method for Clustering Categorical Data" , Information Fusion, vol. 6, no. 2, pp. 143-151, 2005

[2] A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions", J. Machine Learning Research, vol. 3, pp. 583- 617, 2002

[3] N. Nguyen and R. Caruana, "Consensus Clusterings," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 607-612, 2007.

[4] S. Monti, P. Tamayo, J.P. Mesirov, and T.R. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data", Machine Learning, vol. 52, nos. 1/2, pp. 91-118, 2003

[5] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering" , *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012*

[6] M. Al-Razgan, C. Domeniconi, and D. Barbara, "Random Subspace Ensembles for Clustering Categorical Data," Supervised and Unsupervised Ensemble Methods and Their Applications, pp. 31-48, Springer, 2008.

[7] P. Reuther and B. Walter, "Survey on Test Collections and Techniques for Personal Name Matching" , *Intl J. Metadata, Semantics and Ontologies*, vol. 1, no. 2, pp. 89-99, 2006

[8] L.A. Adamic and E. Adar, "Friends and Neighbors on the Web", Social Networks, vol. 25, no. 3, pp. 211-230, 2003 90

[9] Eugenio Cesario, Giuseppe Manco, and Riccardo Ortale , "Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data", I*EEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 12, DECEMBER 2007*

[10] S. Guha, R. Rastogi, and K. Shim," ROCK: A Robust Clustering Algorithm for Categorical Attributes", *15th International Conference on Data Engineering*, pp. 512-521, 2000