Efficient Data Clustering with Link Approach

[1]*Y. Sireesha,* [2] CH. Srinivas, [3] K.C. Ravi Kumar.

[1]PG Scholar, Department of Computer Science and Engineering,
Sridevi Women's Engineering College
Hyderabad, A.P, India

[2] Associate Professor, Department of Computer Science and Engineering,
Sridevi Women's Engineering College
Hyderabad, A.P, India

[3] Head of the Department, Department of Computer Science and Engineering,
Sridevi Women's Engineering College
Hyderabad, A.P, India

Abstract:— Data clustering faces lots of studies and researches and at last the results being competitive to conventional algorithms, even though using these techniques finally we are getting an incomplete information. The existed partitioned-information matrix contains particular cluster-data point relations only, with lot entries which are not recognized. The paper explores researches that preferres this crisis decomposes the efficiency of the clustering result, and it contains a new link-based approach, which increases the conventional matrix by revealing the entries which are not recognized based upon the common things which are present both clusters and in ensemble. Often, a perfect link-based algorithm is invented and used for the underlying common assessment. After all those, to gain the maximum clustering outputs, a graph partitioning technique is used for a weighted bipartite graph that is formulated from the

refined matrix. Results on various real data sets suggest that the proposed link-based method mostly performs both conventional clustering algorithms for categorical data and also most common cluster ensemble techniques.

I. INTRODUCTION

To examine the data set we have different approaches through get down to its structure data clustering is the efficient way. Because of the beneficial characteristics of clustering like mining, machine learning and pattern recognition. Clustering deals with the data to get stick with similar ones. Those similar ones will get into group or cluster. There are sort of algorithms for clustering like k-means and PAM which are used for clustering the numerical data, these are used to

get the distance between feature vectors. The drawback is this not get inherited directly for clustering purpose on the categorical data, where domain values are discrete and have no ordering defined. As a result, many categorical data clustering algorithms have been introduced in recent years, with applications to interesting domains such as protein interaction data. The initial method was developed by making use of Gower's similarity coefficient. Following that, the k-modes algorithm in extended the conventional k-means with a simple matching dissimilarity measure and a frequency-based method to update centroids.

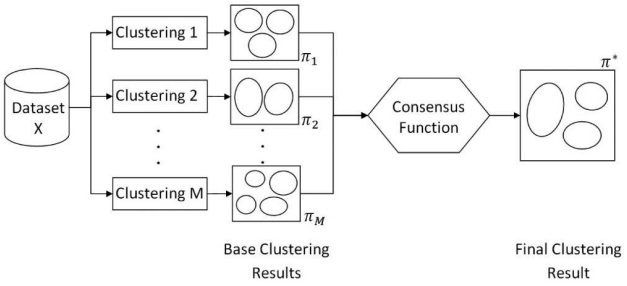


Fig 1. The basic process of cluster ensembles.

As a single-pass algorithm, Squeezer makes use of a prespecified similarity threshold to determine which of the existing clusters to which a data point under examination is assigned. LIMBO is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuples. The concepts of evolutionary computing and genetic algorithm have also been adopted by a partitioning method for categorical data, i.e., GAClust. Cobweb is a model-based method primarily exploited for categorical data sets. Different graph models have also been investigated by the STIRR, ROCK, and CLICK techniques. In addition, several density-based algorithms have also been devised for such purpose, for instance, CACTUS, COOLCAT, and CLOPE. Although, a large number of algorithms have been introduced for clustering categorical data, the No Free Lunch theorem suggests1 there is no single clustering algorithm that performs best for all data sets and can discover all types of cluster shapes and structures presented in data. Each algorithm has its own strengths and weaknesses. For a particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results. 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. Examples of well-known ensemble methods are:

1. The feature-based approach that transforms the problem of cluster ensembles to clustering categorical data.

2. The direct approach that finds the final partition through relabeling the base clustering results.

3. graph-based algorithms that employ a graph partitioning methodology, and

4. The pairwise-similarity approach that makes use of co-occurrence relations between data points.

Despite notable success, these methods generate the final data partition based on incomplete information of a cluster ensemble. The underlying ensemble-information matrix presents only cluster-data point relationships while completely ignores those among clusters. As a result, the performance of existing cluster ensemble techniques may consequently be degraded as many matrix entries are left unknown. This paper introduces a link-based approach to refining the aforementioned matrix, giving substantially less unknown entries. A link-based similarity measure is exploited to estimate unknown values from a link network of clusters. This research uniquely bridges the gap between the task of data clustering and that of link analysis. It also enhances the capability of ensemble methodology for categorical data, which has not received much attention in the literature. In addition to the problem of clustering categorical data that is investigated herein, the proposed framework is generic such that it can also be effectively applied to other data types.

II. DISCUSSION

The difficulty of categorical data analysis is characterized by the fact that there is no inherent distance (or similarity) between attribute values. The RM matrix that is generated within the LCE approach allows such measure between values of the same attribute to be systematically quantified. The concept of link analysis [34], [35], [36] is uniquely applied to discover the similarity among attribute values, which are modeled as vertices in an undirected graph. In particular, two vertices are similar if the neighboring contexts in which they appear are similar. In other words, their similarity is justified upon values of other attributes with which they co-occur. While the LCE methodology is novel for the problem of cluster ensemble, the concept of defining similarity among attribute values (especially with the case of “direct” ensemble, Type-I) has been analogously adopted by several categorical data clustering algorithms. Initially, the problem of defining a context-based similarity measure has been investigated in [61] and [62]. In particular, an iterative algorithm, called “Iterated Contextual Distances (ICD),” is introduced to compute the proximity between two values. Similar to LCE, the underlying distance metric is based on the occurrence statistics of attribute values. The WTQ algorithm is summarized below

ALGORITHM: WTQ(G, C_x, C_y)
$G = (V, W)$, a weighted graph, where $C_x, C_y \in V$;
$N_k \subset V$, a set of adjacent neighbors of $C_k \in V$;
$W_k = \sum_{\forall C_t \in N_k} w_{tk}$;
WTQ_{xy}, the WTQ measure of C_x {and} C_y;
(1) $WTQ_{xy} \leftarrow 0$
(2) **For each** $c \in N_x$
(3) **If** $c \in N_y$
(4) $WTQ_{xy} \leftarrow WTQ_{xy} + \frac{1}{W_c}$
(5) **Return** WTQ_{xy}

However, the fundamental information model that is used by ICD and LCE to capture the associations between data points and attribute values are notably different: a sequential probabilistic chain and a link network for ICD and LCE, respectively. Note that LCE makes use of WTQ that is a single-pass similarity algorithm, while ICD requires the chain model to be randomly initialized and iteratively updated to a fixed point.

III. SYSTEM DEVELOPMENT

The link based approach will be done in following steps

K-Means with Euclidian Distance (∏1)

- Transform categorical (text) into numerical value
- On Numerical data apply K-Means with Euclidian distance measure.
- The outcome of the ∏1 is {C11, C12, C13 …….. C1n} where C1 is the cluster 1 with ∏1 clustering algorithm.

K-Means with cosine similarity (∏2)

- Consider numerical values are categorical data
- On categorical data apply (∏2) for clustering.
- The outcome of the ∏2 is {C21, C22, C23…….. C2n}

K-Means with Jaccord's Coefficient (∏3)

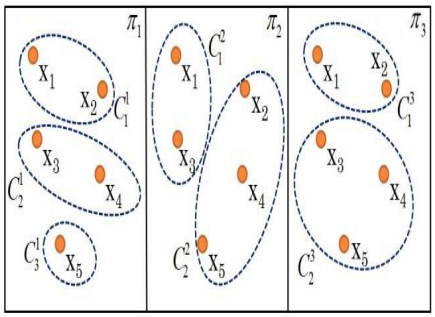
- Consider numerical values as categorical data
- On categorical data, apply (∏3) for clustering
- The outcome of the ∏3 is {C31, C32, C33, C34…… C3n}

Cluster Ensemble

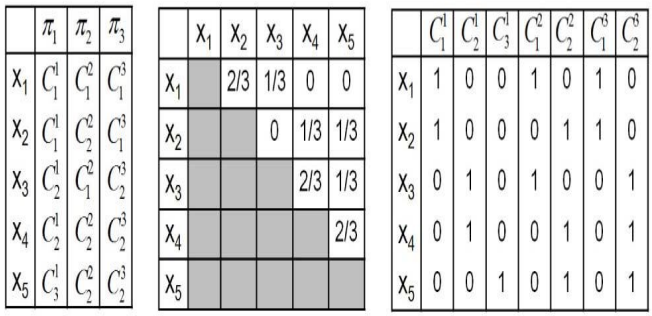
- Direct Ensemble
- Full Space Ensemble
- Subspace Ensemble

Generating Refined Matrix (RM)

- Prepare RM as a matrix where each record of the original dataset D as {x1, x2, x3…… xm}



(a)



(b) (c) (d)

Weighted Triple Quality (WTQ)

- Wxy = $\frac{Lx \,\Omega\, Ly}{Lx \,U\, Ly}$

Apply WTQ on RM, The output of the module is refined clusters.

IV. RELATED WORK

Despite pursuing an objective analogous to that of the LCE approach, several categorical data clustering methods have been developed using different mechanisms to specify a distance between attribute values: STIRR, ROCK, and CACTUS, for instance. STIRR is an iterative algorithm based on nonlinear dynamical systems. A database is encoded into a graph structure, where each weighted node stands for a specific attribute value. STIRR iteratively updates the weight configuration until a stable point (called "basin") is reached. This is achieved using a user-defined "combiner function" to estimate a node weight from those of others that associate to the same data records. Unlike LCE, the similarity between any node pair cannot be explicitly measured here. In fact, STIRR only divides nodes of each attribute into two groups (one with large positive weights and the other with small negative weights) that correspond to projections of clusters on the attribute. Yet, the post processing required to generate the actual clusters is nontrivial and not addressed in the original work. While LCE is generally robust to parameter settings, it is hard to analyze the stability of the STIRR system for any useful combiner function [63]. Rigorous experimentation and fine tuning of parameters are needed for the generation of a meaningful clustering [64]. ROCK [14] makes use of a link graph, in which nodes and links represent data points (or tuples) and their similarity, respectively. Two tuples are similar if they shared a large number of attribute values. Note that the link connecting two nodes is included only when the corresponding similarity exceeds a user-defined threshold. With tuples being initially regarded as singleton clusters, ROCK merges clusters in an agglomerative hierarchical fashion, while optimizing a cluster quality that is defined in terms of the number of links across clusters. Note that the graph models

used by ROCK and LCE are dissimilar—the graph of data points and that of attribute values (or clusters), respectively. Since the number of data points is normally greater than that of attribute values, ROCK is less efficient than LCE. As a result, it is unsuitable for large data sets. Also, the selection of a "smooth function" that is used to estimate a cluster quality is a delicate and difficult task for average users. CACTUS also relies on the co-occurrence among attribute values. In essence, two attribute values are strongly connected if their support (i.e., the proportion of tuples in which the values co-occur) exceeds a prespecified value. By extending this concept to all attributes, CACTUS searches for the "distinguishing sets," which are attribute value sets that uniquely occur within only one cluster. These sets correspond to cluster projections that can be combined to formulate the final clusters. Unlike LCE, the underlying problem is not designed using a graph based concept. It is also noteworthy that CACTUS and its recent extension assume each cluster to be identified by a set of attribute values that occur in no other cluster. While such conjecture may hold true for some data sets, it is unnatural and unnecessary for the clustering process. This rigid constraint is not implemented by the LCE method. Besides these approaches, traditional categorical data analysis also utilizes the "market-basket" numerical representation of the nominal data matrix. This transformed matrix is similar to the BM, which has been refined to the RM counterpart by LCE. A similar attempt in identifies the connection between "category utility" of the conceptual clustering (Cobweb) and the classical objective function of k-means.

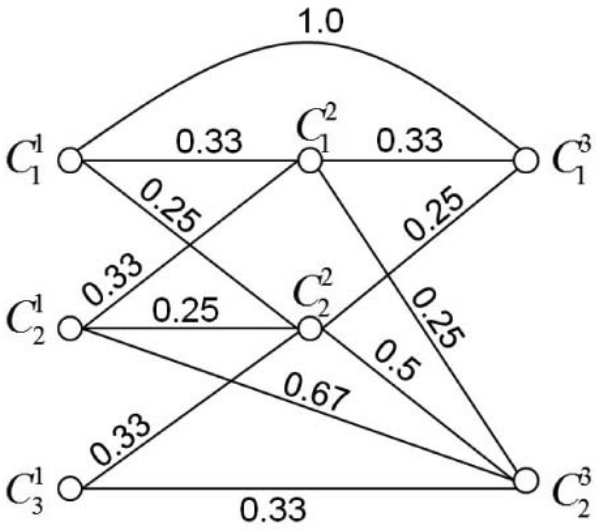


Fig 2. An example of a cluster network, where each edge is marked with its weight.

As a result, the so-called market-basket matrix used by the former is transformed to a variation that can be efficiently utilized by the latter. The intuitions of creating this rescaled matrix and the RM are fairly similar. However, the methods used to generate them are totally different. LCE discovers unknown entries (i.e., "0") in the original BM from known entries ("1"), which are preserved and left unchanged. On the other hand, the method in maps the attribute-value-specific "1" and "0" entries to the unique standardized values. Unlike the RM, this matrix does not

conserve the known fact ("1" entries), whose values are now different from one to another attribute. Despite the fact that many clustering algorithms and LCE are developed with the capability of comparing attribute values in mind, they achieve the desired metric differently, using specific information models. LCE uniquely and explicitly models the underlying problem as the evaluation of link-based similarity among graph vertices, which stand for specific attribute values (for Type-I ensemble) or generated clusters (for Type-II and Type-III). The resulting system is more efficient and robust, as compared to other clustering techniques emphasized thus far. In addition to SPEC, many other classical clustering techniques, k-means and PAM among others, can be directly used to generate the final data partition from the proposed RM. The LCE framework is generic such that it can be adopted for analyzing other types of data.

V. CONCLUSION

This paper presents a novel, highly effective link-based cluster ensemble approach to categorical data clustering. It transforms the original categorical data matrix to an information-preserving numerical variation (RM), to which an effective graph partitioning technique can be directly applied. The problem of constructing the RM is efficiently resolved by the similarity among categorical labels (or clusters), using the Weighted Triple-Quality similarity algorithm. The empirical study, with different ensemble types, validity measures, and data sets, suggests that the proposed link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and benchmark cluster ensemble techniques. The prominent future work includes an extensive study regarding the behavior of other link-based similarity measures within this problem context. Also, the new method will be applied to specific domains, including tourism and medical data sets.

REFERENCES

[1] D.S. Hochbaum and D.B. Shmoys, "A Best Possible Heuristic for the K-Center Problem," Math. of Operational Research, vol. 10, no. 2, pp. 180-184, 1985.

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

[3] A.K. Jain and R.C. Dubes, Algorithms for Clustering. Prentice-Hall, 1998.

[4] P. Zhang, X. Wang, and P.X. Song, "Clustering Categorical Data Based on Distance Vectors," The J. Am. Statistical Assoc., vol. 101, no. 473, pp. 355-367, 2006.

[5] J. Grambeier and A. Rudolph, "Techniques of Cluster Algorithms in Data Mining," Data Mining and Knowledge Discovery, vol. 6, pp. 303-360, 2002.

[6] K.C. Gowda and E. Diday, "Symbolic Clustering Using a New Dissimilarity Measure," Pattern Recognition, vol. 24, no. 6, pp. 567-578, 1991.

[7] J.C. Gower, "A General Coefficient of Similarity and Some of Its Properties," Biometrics, vol. 27, pp. 857-871, 1971.

[8] Z. Huang, "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, vol. 2, pp. 283-304, 1998.

[9] 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.

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

[11] 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.

[12] D.H. Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, vol. 2, pp. 139-172, 1987.