Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 9, September 2013, pg.220 – 226

RESEARCH ARTICLE

A Link-Based Cluster Collection Approach Combined Contagious Cluster With For Categorical Data Clustering

N.Premalatha[1], M.Chinnusamy[2]

1Research Scholar, Department of Computer Science, Vivekanandha College, Elayampalayam, Tiruchengode, Tamil Nadu, India

2Assistant Professor, Department of Computer Science, Vivekanandha College, Elayampalayam, Tiruchengode, Tamil Nadu, India

[1] premanatesan89@gmail.com; [2] chitrachinnu@rediffmail.com

ABSTRACT: - Data clustering is a challenging task in data mining technique. Various clustering algorithms are developed to cluster or categorize the datasets. Many algorithms are used to cluster the categorical data. Some algorithms cannot be directly applied for clustering of categorical data. Several attempts have been made to solve the problem of clustering categorical data via cluster ensembles. But these techniques generate a final data partition based on incomplete information. The ensemble information matrix represents cluster relations with many unknown entries. The link based ensemble approach has been established with the ability to discover unknown values and improve the accuracy of the data partition. Besides clustering, similarity based ranking approach, HITS link analysis is also proposed to enhance the categorical results. This enhanced link-based clustering and ranking method almost outperforms both predictable clustering algorithms for categorical data and contagious cluster ensemble techniques for grade.

Keywords- Uncertainty; index; range aggregate

I.INTRODUCTION

Data clustering is one of the challenging task in various applications. Data clustering is one of the fundamental tools to understand the structure of the data set. Clustering aims to categorize data into groups or clusters such that the data in the same cluster are more similar to each other than those in different clusters. Clustering is a data mining technique used to place similar data elements into related groups. A cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The notation of the cluster varies between different algorithms. The clusters found by different clustering algorithms are varying in their properties and structure. Clustering is used in many areas such as Statistical Data Analysis, Machine Learning, Data Mining, Pattern Recognition, Image Analysis, Bioinformatics, etc., The various clustering algorithms are Distance-based, Hierarchical, Partitioning, Probabilistic are proposed to cluster the datasets. These clustering algorithms are used to cluster the various data sets. Cluster ensembles provide a solution to challenges inherent to clustering. Cluster ensembles can find robust and stable solutions by leveraging the consensus across multiple clustering results. The cluster ensemble combines various clustering outputs into single consolidated cluster.

The cluster ensemble will differentiate various cluster outputs by using the clustering algorithms. The main goal of ensembles has been to improve the accuracy and robustness of a given classification or regression task, and spectacular improvements have been obtained for a wide variety of data sets. Cluster ensemble methods are presented under three categories: Probabilistic approaches, Approaches based on co-association, and Direct and other heuristic methods. Categorical variables represent types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level. Categorical data is a statistical data type consisting of categorical values used for observed data whose value is one of a fixed number of nominal categories, or for data that has been converted into that form. Categorical data are always nominal whereas nominal data need not be categorical. Clustering the categorical data is remaining a challenging task in many techniques. A critical problem in cluster ensemble research is how to combine multiple clustering's to yield a final superior clustering result. These problems are overcome by using different techniques. The link based similarity is used to improve the clustering result.

II.RELATED WORK

A. Clustering Categorical Data

A few algorithms have been proposed in recent years for clustering categorical data [5-24]. In [5], the problem of clustering customer transactions in a market database is addressed. STIRR, an iterative algorithm based on non-linear dynamical systems is presented in [6]. The approach used in [6] can be mapped to a certain type of non-linear systems. If the dynamical system converges, the categorical databases can be clustered. Another recent research [7] shows that the known dynamical systems cannot guarantee convergence, and proposes a revised dynamical system in which convergence can be guaranteed. K-modes, an algorithm extending the k-means paradigm to categorical domain is introduced

In [8, 9]. New dissimilarity measures to deal with categorical data is conducted to replace means with modes, and a frequency based method is used to update modes in the clustering process to minimize the clustering cost function. Based on k-modes algorithm, [10] proposes an adapted mixture model for categorical data, which gives a probabilistic interpretation of the criterion optimized by the k-modes algorithm. A fuzzy k-modes algorithm is presented in [11] and tabu search technique is applied in [12] to improve fuzzy k-modes algorithm. An iterative initial-points refinement algorithm for categorical data is presented in [13]. The work in [23] can be considered as the extensions of k-modes algorithm to transaction domain. In [14], the authors introduce a novel formalization of a cluster for categorical data by generalizing a definition of cluster for numerical data. A fast summarization based algorithm, CACTUS, is presented. CACTUS consists of three phases: summarization, clustering, and validation. ROCK, an adaptation of an agglomerative hierarchical clustering algorithm, is introduced in

Squeezer, a one-pass algorithm is proposed in [20]. Squeezer repeatedly read tuples from dataset one by one. When the first tuple arrives, it forms a cluster alone. The consequent tuples are either put into an existing cluster or rejected by all existing clusters to form a new cluster according to the given similarity function.

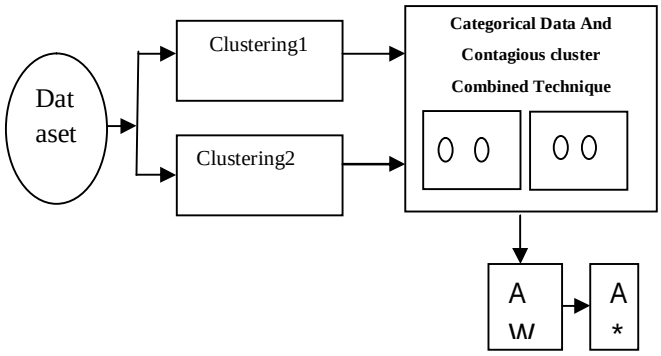
COOLCAT, an entropy-based algorithm for categorical clustering, is proposed in [21]. Starting from a heuristic method of increasing the height-to-width ratio of the cluster histogram, the authors in [22] develop the CLOPE algorithm. [24] introduce a distance measure between partitions based on the notion of generalized conditional entropy and a genetic algorithm approach is utilized for discovering the median partition.

B.Cluster Ensemble

In [25], the authors formally defined the CE problem as an optimization problem and propose combiners for solving it based on a hyper-graph model. A multi-clustering fusion method is presented in [27]. In that method, the results of several independent runs of the same clustering algorithm are appropriately combined to obtain a partition of the data that is not affected by initialization and overcomes the instabilities of clustering methods. After that, the fusion procedure starts with the clusters produced by the combining part and finds the optimal number of clusters according to some predefined criteria. The authors in [28] proposed a sequential combination method to improve the clustering performance. First, their algorithm uses the global criteria based clustering to produce an initial result, then use the local criteria based information to improve the initial result with a probabilistic relaxation algorithm or linear additive model.

III. CATEGORICAL DATA AND CONTIGIOUS CLUSTER COMBINED TECHNIQUE

cluster ensemble methods to categorical data analysis rely on the typical pair wise-similarity and binary cluster-association matrices [48], [49], which summarize the underlying ensemble information at a rather coarse level. Many matrix entries are left "unknown" and simply recorded as "0." Regardless of a consensus function, the quality of the final clustering result may be degraded. As a result, a link based method has been established with the ability to discover unknown values and, hence, improve the accuracy of the ultimate data partition [33]. In spite of promising findings, this initial framework is based on the data point data point pair wise-similarity matrix, which is highly expensive to obtain. The link-based similarity technique, SimRank [52], that is employed to estimate the similarity among data points is inapplicable to a large data set. To overcome these problems, a new link-based cluster ensemble (LCE) approach is introduced herein. It is more efficient than the former model, where a BM-like matrix is a new link-based cluster ensemble (LCE) approach is introduced herein. It is more efficient than the former model, where a BM-like matrix is used to represent the ensemble information. The focus has shifted from revealing the similarity among data points to estimating those between clusters. A new link-based algorithm has been specifically proposed to generate such measures in an accurate, inexpensive manner. The LCE methodology is illustrated in Fig. 3. It includes three major steps of: 1) creating base clustering's to form a cluster ensemble (_), 2) generating a refined cluster-association matrix (RM) using a link-based similarity algorithm, and 3) producing the final data partition by exploiting the spectral graph partitioning technique as a consensus function.



Fig,1 Categorical Data Analysis

IV.EXPERIMENT RESULTS

A.Performance Evaluation

This section presents the evaluation of the propose New link based method (NLCD), using a variety of validity indices and real data sets. The quality of data partitions generated by this technique is assessed against those created by different categorical data clustering algorithms and cluster ensemble techniques) Investigated Datasets the experimental evaluation is conducted over nine data sets. The "20Newsgroup" data set is a subset of the well-known text data collection—20- Newsgroups,2 while the others are obtained from the UCI Machine Learning Repository [22]. Their details are summarized in Table 1. Missing values (denoted as "?") in these data sets are simply treated as a new categorical value. The "20Newsgroup" data set contains 1,200 documents from two newsgroups, each of which is described by the occurrences of 6,084 different terms. In particular, the frequency (f ∈ 0, 1, . . . ,∞) that a key word appears in each document is transformed into a nominal value: "Yes"
if f > 0, "No" otherwise. Moreover, the "KDDCup99" data set used in this evaluation is a randomly selected subset of the original data. Each data point (or record) corresponds to a network connection and contains 42 attributes: some are nominal and the rest are continuous. Following the study in [17], numerical attributes are transformed to categorical using a simple discretization process. For each attribute, any value less than the median is assigned a label "0," otherwise "1." Note that the selected set of data records covers 20 different connection classes. These two data sets are specifically included to assess the performance of different clustering methods, with respect to the large numbers of dimensionality and data points, respectively. To fully evaluate the potential of the proposed method, it is compared to the baseline model (referred to as "Base" hereafter), which applies SPEC to the BM. This allows the quality of BM and RM tobe directly compared. In addition, five clustering techniques for categorical data and five methods developed for cluster ensemble problems are included in this evaluation. Details of these techniques are given below.

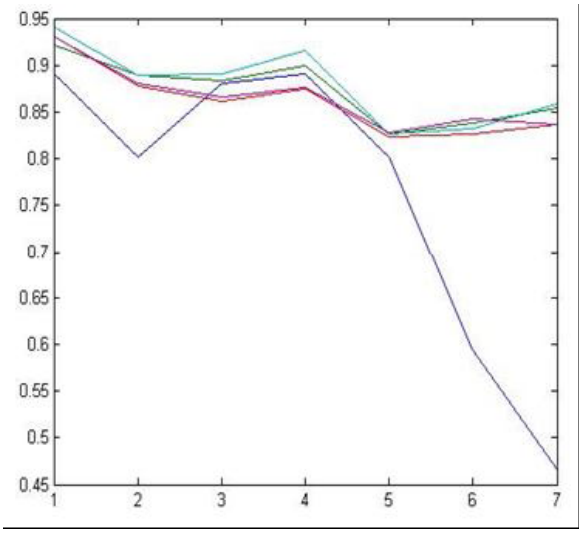
Dataset	N	d	A	K
Hospital	152	12	17	12
Zoo	101	16	36	7
Primary tumor	339	17	42	22
Soy bean	307	35	132	19
Breast Cancer	683	9	89	2
Mushroom	8124	22	117	2
20Newsgroup	1000	6084	12168	2
KDDcup	100,000	42	139	20
Iris	788	52	469	4

Clustering algorithms for categorical data is based on their notable performance reported in the literature and availability; five different algorithms are selected
to demonstrate the efficiency of conventional techniques to clustering categorical data: Squeezer, GAClust, k-modes, CLOPE, and Cobweb. Squeezer [9] is a single-pass algorithm that considers one data point at a time.

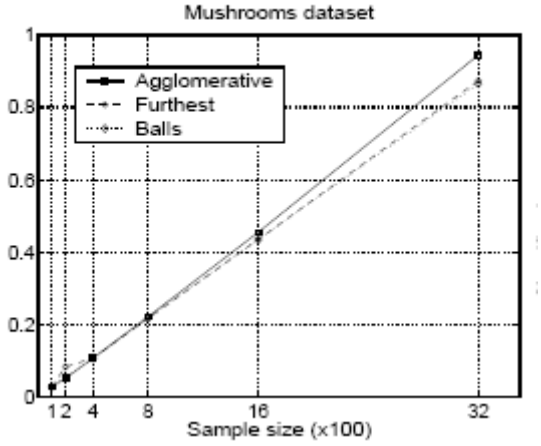
Each data point is either placed in one of the existing clusters if their distance is less than a given threshold, or used to form a new cluster. GAClust [11] searches for a data partition (referred to as the "median" partition), which has the minimum dissimilarity to those partitions generated by categorical attributes. Note that the similarity (or closeness) between two partitions is estimated by using a generalization of the classical conditional entropy. A genetic algorithm has been employed to make the underlying search process more efficient, with the partitions being represented by chromosomes.

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 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

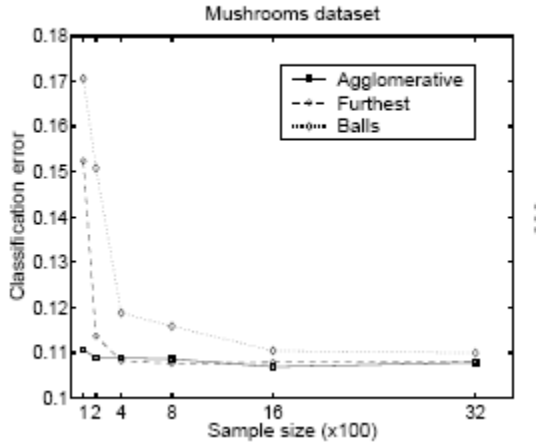
"direct" ensemble, Type-I) has been analogously adopted by several categorical data clustering algorithms. Cobweb [12] is a conceptual clustering method. It creates a classification tree, in which each node corresponds to a concept. Observations are incrementally integrated into the classification tree, along the path of best matching nodes. This is guided by the heuristic evaluation measure, called category utility. A given utility threshold determines the sibling nodes that are used to form the resulting data partition. Initially, the problem of defining a context-based similarity measure has been investigated in and. 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.



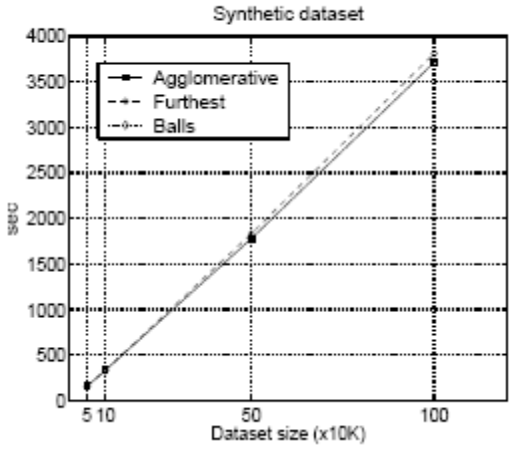
Fig, 1



Fig,2



Fig,3



Fig,4

V.CONCLUSION

Clustering ensembles have emerged as a categorical data and contiguous cluster combined for improving robustness, stability and accuracy of unsupervised classification solutions. So far, many contributions have been done to find consensus clustering. Firstly, we introduced clustering ensembles and research area and showed different representation of multiple partitions. There are several challenges for clustering ensemble that one of the major problems in clustering ensembles is the consensus function. We summarized clustering combination approaches and focused on consensus function method including: Hyper graph partitioning, Voting approach, Mutual information, Co-association based functions and Finite mixture model. We investigated some of the most important previous research works in each approach and compared their advantages, disadvantages and computational complexity. The comparison results show that robustness in all of the techniques is high. There are difficulties in implementing the algorithms especially in Hypergraph partitioning and Coassociation based functions techniques, thus, simplicity in their algorithms is necessary.

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 DataBased 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.
[13] 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.
[14] 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.
[15] 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.
[16] V. Ganti, J. Gehrke, and R. Ramakrishna, "*CACTUS: Clustering Categorical Data Using Summaries*," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.
[17] D. Barbara, Y. Li, and J. Couto, "*COOLCAT: An Entropy-Based Algorithm for Categorical Clustering*," Proc. In 'l Conf. Information and Knowledge Management (CIKM), pp. 582-589, 2002.
[18] 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- 687, 2002.
[19] D.H. Wolpert and W.G. Macready, "*No Free Lunch Theorems for Search*," Technical Report SFI-TR-95-02-010, Santa Fe Inst., 1995.
[20] L.I. Kuncheva and S.T. Hadjitodorov, "*Using Diversity in Cluster Ensembles*," Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, pp. 1214-1219, 2004.
[21] H.Xue, S. Chen, and Q. Yang, "*Discriminatively Regularized Least-Squares Classification*," Pattern Recognition, vol. 42, no. 1,pp. 93-104, 2009.

