IMPLEMENTATION OF CLUSTERING ALGORITHM AND SIMILARITY MEASURE FOR CATEGORICAL DATA

[1]Shravya Mandava, [2]Bhanu Chander
[1]Dept of ECE, Vardhaman College of Engineering, Hyderabad, Telangana, India

***ABSTRACT*:**
Data clustering could be a primary tool for understanding the structure of Data sets (information). The existed divided information matrix contains specific cluster-data purpose relations solely, with ton entries that aren't recognized. The ensemble data matrix represents cluster relations with many unknown entries. This paper presents a replacement link-based approach that improves the specific agglomeration by discovering unknown entries through similarity between clusters in Associate in nursing ensemble. A try wise similarity approach is applied to a weighted bipartite graph to get the ultimate agglomeration result. Therefore the link-based approach outperforms every typical agglomeration algorithms for categorical data and well-known cluster ensemble technique. This project proposes Associate in nursing formula cited as Average Weighted Quality (AWQ) that in addition uses k-means formula for basic agglomeration. Once the elemental agglomeration is finished by exploitation agreement functions we tend to square measure able to get cluster ensembles of categorical information. This categorical data is born-again to stylish matrix. This project introduces a link-based approach to purification a similar matrix, giving significantly less unknown entries.

INDEX TERMS—*clustering, ranking, categorical data, cluster ensemble.*

I. INTRODUCTION

Data agglomeration is one in the entire difficult task in numerous applications. Information agglomeration is one in all the basic tools to grasp the structure of the info set. Agglomeration aims to categories information into teams or clusters such the info within the same cluster square measure a lot of the same as one another than those in many clusters. Cluster may be a data processing technique accustomed place similar information components into connected teams. A cluster is a collection of objects that square measure "similar" between them and square measure "dissimilar" to the objects happiness to other clusters. The representation of the cluster varies between totally different algorithms. The clusters found by totally different agglomeration algorithms square measure varied in their properties and structure. agglomeration is employed in several areas like applied mathematics information Analysis, Machine Learning, data processing, Pattern Recognition, Image Analysis, Bioinformatics, etc., the varied agglomeration algorithms square measure Distance-based, Dividing, Probabilistic square measure planned to cluster the datasets. These clustering procedures rectangular measure accustomed cluster the varied information sets. Cluster ensembles give an answer to challenges characteristic to agglomeration. Cluster ensembles will realize strong and stable solutions by investment the accord across multiple agglomeration results. The cluster ensemble combines numerous agglomeration outputs into single shared cluster. The cluster joint can differentiate numerous cluster outputs by victimization the agglomeration algorithms. The most important goal of ensembles has been to boost the accuracy and strength of a given classification or regression task, and spectacular enhancements are obtained for a large kind of information sets. Cluster ensemble strategies square measure conferred underneath 3 categories:

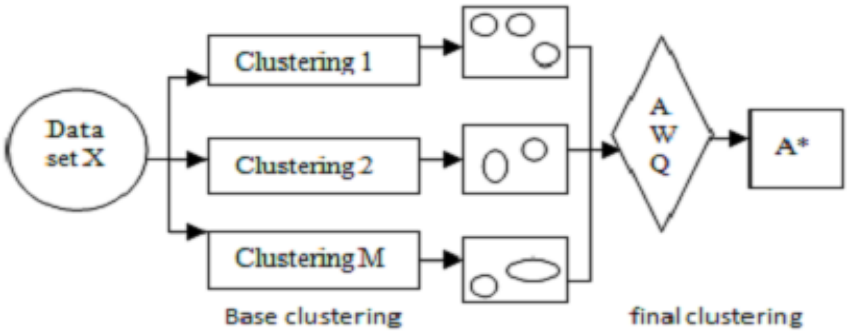

Fig 1: The Basic process of cluster Methodology

Probabilistic approaches, Approaches supported co-association, and Direct and alternative heuristic strategies. Categorical variables represent kinds of information which can be divided into teams. Samples of categorical variables square measure race, sex, age group, and academic level. Categorical information may be an applied mathematics information sort consisting of categorical worth's used for discovered information whose value is one in all a set range of nominal classes, or for information that has been regenerate into that type. Categorical information square measure forever nominal whereas nominal information needn't be categorical.Agglomeration the explicit information is remaining a difficult task in several techniques. An essential downside in cluster ensemble analysis is a way to mix multiple agglomeration's to yield a final superior clustering result. These issues square measure overcome by victimization totally different techniques. The link primarily based similarity is employed to boost the agglomeration result.

II. A NOVEL LINK BASED APPROACH

Existing cluster ensemble methods to categorical data analysis rely on the typical pairwise-similarity and binary Cluster association matrices, 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. In spite of promising findings, this initial framework is based on the data point data point pairwise-similarity matrix, which is highly expensive to obtain. The link-based similarity technique, SimRank 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 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 Figure 1b. It includes three major steps of:

1) Creating base clustering's to form a cluster ensemble (π),
2) Generating a refined cluster-association matrix (RM) using a link-based similarity algorithm, and
3) Producing the final data by exploiting the spectral graph partitioning technique as a consensus function.

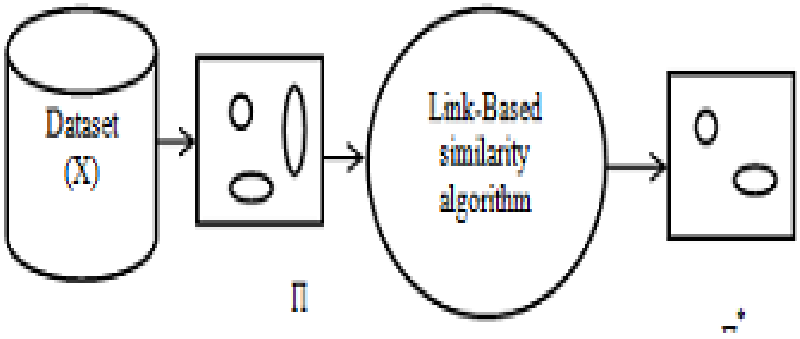

Figure 2: The link based cluster ensemble

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.

III. PROPOSED SOLUTION

The cluster ensemble results are measured using link based similarity measure for cluster ensemble methods, instead of random selection of cluster centroid values automatically select centroid values , clustering similarity measure part also need to improve using other similarity measurements. 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. Sometimes an image may contain text embedded on to it. Detecting and recognizing these characters can be very important, and removing these is important in the context of removing indirect advertisements, and for aesthetic reasons.
Evaluation of the proposed link based method (LCE), 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. In order to evaluate the quality of cluster ensemble methods previously identified, they are empirically compared, using the settings of cluster ensembles exhibited below. . Five types of cluster ensembles are investigated in this evaluation: Type-I, Type-II (Fixed-k), Type-II (Random-k), Type-III (Fixed), and Type-III (Random- k). The k-modes clusteringalgorithm is specifically used to generate the base clustering's with clustering methods such as Link-Based Cluster Ensemble(LCE) , Similarity matrix (CO) with single linkage (CO+SL), Similarity matrix (CO) with average linkage (CO+AL), Cluster-based Similarity Partitioning Algorithm (CSPA), Hyper-Graph Partitioning Algorithm (HGPA) and proposed optimization based clustering methods . Table 3 illustrates for each method the frequencies of significant better (B) performance, which are categorized in accordance with the evaluation indices Normalized Mutual Information (NMI), Adjusted Rand (AR) and Classification Accuracy (CA) .The results shown in this table indicate the superior effectiveness of the proposed link based methods, as compared to other clustering techniques included in this experiment .

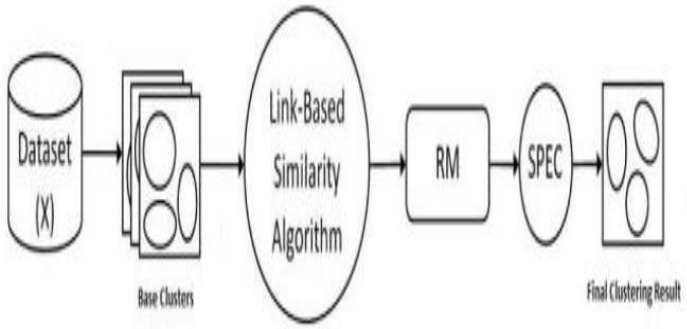

Fig. 3 Link-based cluster ensemble approach

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 aninformation 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) 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.

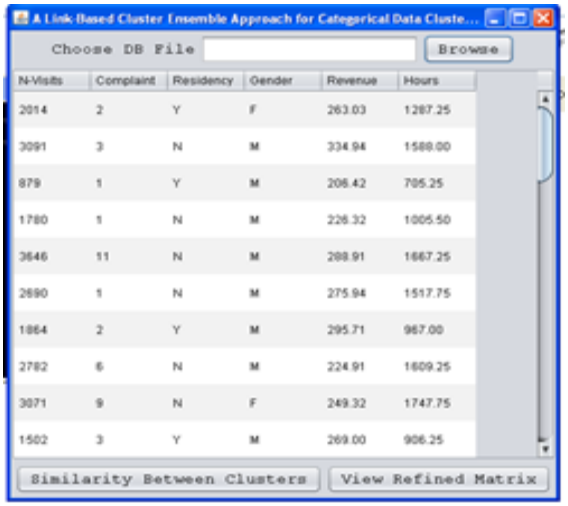

Fig 4: data clustering

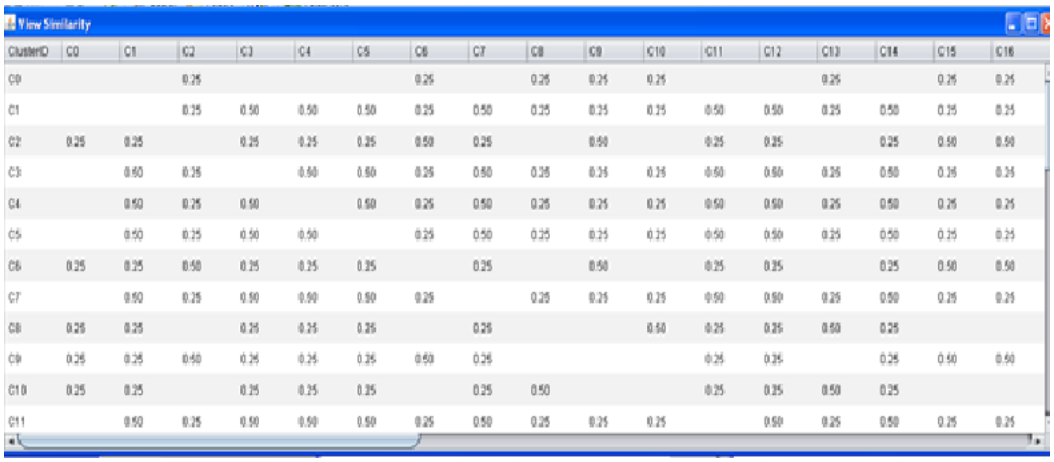

Fig 5: Similarity between all clusters

IV. FUTURE WORK

For future work, we are planning to take the data from cloud, and design k-means like clustering algorithms for categorical data that directly optimize the mutual information sharing based object function.In our evaluation methodology we have used one similarity measure across all attributes. Since different attributes in a data set can be of different nature, an alternative way is to use different measures for different attributes. This appears to be especially promising given the complimentary nature of several similarity measure.

V. CONCLUSION

This paper presents a completely unique, extremely effective link-based cluster ensemble approach to categorical information agglomeration. It transforms the initial categorical information matrix to Associate in nursing information-preserving numerical variation (RM), to which an effective graph partitioning technique will be directly applied. The matter of constructing the RM is expeditiously resolved by the similarity among categorical labels (or clusters), exploitation the Weighted Triple-Quality similarity algorithm. The empirical study, with completely different ensemble types, validity measures, and information sets, suggests that the proposed link-based technique sometimes achieves superior clustering results compared to those of the normal categorical information algorithms and benchmark cluster ensemble techniques. The distinguished future work includes an intensive study concerning the behavior of alternative link-based similarity measures inside this downside context.

Also, the new technique will be applied to specific domains, as well as commercial enterprise and medical information sets.

REFERENCES

[1] 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.
[2] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University.
[3]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.
[4]A.L.N. Fred and A.K. Jain, "Combining Multiple Clustering's Using Evidence Accumulation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, June 2005.
[5] 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
[6] L.A. Adamic and E. Adar, "Friends and Neighbors on the Web", Social Networks, vol. 25, no. 3, pp. 211-230, 2003 90.
[7] Aranganayagi.S and Thangavel.K, "Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure", International Journal of Information and Mathematical Sciences, Vol.6, No.1, pp.1-8, 2010.
[8] F. Gullo, C. Domeniconi, and A. Tagarelli, " Projective clustering ensembles", In IEEE International Conference on Data Mining, pp. 794- 799, 2009.
[9] 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.
[10] D. Liben-Nowell and J. Kleinberg, ―The Link Prediction Problem for Social Networks,‖ J. Am. Soc. for Information Science and Technology, vol. 58, no. 7, pp. 1019-1031, 2007.
[11] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS: Clustering Categorical Data Using Summaries," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.
[12] Cao, T.H., H.T. Do, D.T. Hong and T.T. Quan, 2008. Fuzzy named entity-based document clustering. Proceedings of IEEE International Conference on Fuzzy Systems, June 1-6, Hong Kong, pp: 2028-2034

AUTHORS BIOGRAPHY

Shravya Mandava, pursuing her B.tech from Vardhaman college of Engineering, Kacharam village, Shamshabad Mandal, Ranga Reddy District T.G, India. Jawaharlal Nehru Technological University (Autonomous), Hyderabad.Approved by AICTE, NEW DELHI. She also worked as an Intern in TEKsystems Global Services in the domain of Business Analytics / Business Intelligence prior to her actual employment in the same company.

Bhanu Chander,completed his B.tech from Aurora's Technological & Research Institute, Hyderabad, T.G, India. Jawaharlal Nehru Technological University.Approved by AICTE, NEW DELHI.He also worked as a Java developer in Nacre Software Service and Emobitise Technologies Pvt Ltd. Currently He is working as a Java trainer cum developer with Krest Technologies, Hyd from 2012 to till date.