A SURVEY ON ENHANCED APPROACH FOR CATEGORICAL LINK BASED CLUSTERING

C.M.Geetha[1], K.Sangeetha [2] ,Dr S.Karthik[3]

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 conventional clustering algorithms for categorical data and well-known cluster ensemble techniques for ranking.**

Index Terms—**Clustering, categorical data, cluster ensemble,ranking.**

I.INTRODUCTION

Data clustering is one of the challenging task in various appliactions. 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 (machine learning) 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 remain a challenging task in many techniques.

A critical problem in cluster ensemble research is how to combine multiple clusterings 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. *Link-Based Approach*

The categorical data is clustered and represented using the cluster ensembles. Cluster ensembles are used as best alternative to the standard cluster analysis. The data set has been clustered by using any of the well known cluster algorithm and represented as a cluster ensemble. The cluster ensembles generate a final data partition based on incomplete information and the information is not prefect to make use of it.

The novel link based approach is used to the cluster ensemble problem. The two consensus functions are

Manuscript received Jan, 2013.

Geetha.C.M, *Department Of Computer Science and Engineering, SNS College Of Technology, Coimbatore, India,9578441189*
Sangeetha.K, *Department Of Computer Science and Engineering, SNS College Of Technology, Coimbatore, India.*
Dr. Karthik. S, *Department Of Computer Science and Engineering, SNS College Of Technology, Coimbatore, India.*

generated from the RM: feature based partitioning and bipartite graph partitioning [7]. In link based framework, it first creates a refined matrix using links between the cluster points and the two consensus methods are applied to generate the final ensemble cluster.

The new link-based approach is used to improve quality of the conventional matrix. It achieves the result using the similarity between clusters and provides cluster view points of the cluster ensemble.

b. Similarity Weight and Filter Method

Many clustering algorithms work efficient either on pure numeric data or on pure categorical data, most of them perform poorly on mixed categorical and numerical data types. The clustering of mixed numeric and categorical data set is a challenging task. The scalability and memory constraint is a problem in clustering the large data sets. The clustering algorithm based on similarity weight and filter method paradigm [9] that works well for data with mixed numeric and categorical features.

The incremental clustering algorithm is used to cluster the categorical data. The incremental algorithm is more dynamic than other clustering algorithm. The algorithm uses The algorithm works efficiently even if the boundaries of clusters are irregular. The advantage is that we mix the different clustering datasets (labeled, unlabeled, nominal, and ordinal) with different algorithms.

c. Bayesian Concept

Compared to numerical clustering, categorical clustering seems to be more complicated. As the data mining deals with large data sets, the algorithms should be scalable. Due to the special properties of categorical data it seems more complicated than that of numerical data. Bayesian classifier is used to classify both numeric and nominal data.

The Bayesian concept is similar to K-modes concept. The naive Bayes concept is used to cluster the categorical data. The naïve Bayes concept in clustering are used with the assumption of K clusters, the objects are grouped based on the maximum posteriori probability. The process of clustering starts with K clusters each with one object as a member.

The Bayesian concept is efficient than K-modes and generates clusters with high purity and the results prove that this method is effective in the case of large data sets when compared to small data sets. The advantage of Bayesian concept are scalable, no repetitive execution is needed, produces efficient clusters for large data set, insensitive to the order of input.

d. Ensemble Fuzzy Clustering for Mixed Numeric and Categorical Data

The dataset sometimes may be in mixed nature that is it may consist of both numeric and categorical type of data. Naturally these two types of data may differ in their characteristics. Due to the differences in their characteristics, in order to group these types of mixed data, it is better to use the ensemble clustering method which uses split and merge approach to solve this problem.

The original mixed dataset is splitted into numeric dataset and categorical dataset and clustered using both traditional clustering algorithms (K-Means and K-Modes) and fuzzy clustering algorithms (Fuzzy C-Means and Fuzzy C-Modes). The resultant clusters are combined using ensemble clustering methods and evaluated.

e. Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure

Scalability and memory constraint is the challenging problem in clustering large data set. The new incremental algorithm [2] is used to cluster the categorical data. Categorical data is the one which cannot be ordered and with limited domains. Incremental algorithm finds clusters in less computation time.

In general the incremental algorithms generate large number of clusters; naturally the purity is also more, whereas the proposed measures generate less number of clusters with high purity. The proposed method is capable of clustering large data set.

f. Link-Based Cluster Ensemble Approach for Categorical Data Clustering

Clustering aims to categorize data into groups or clusters. Many well-established clustering algorithms, such as k-means have been designed for numerical data. However, these cannot be directly applied for clustering of categorical data.

To solve the problem of clustering categorical data via cluster ensembles with many entries being left unknown, the new link based similarity approach is proposed. In link based similarity approach finds the similarity between the data clusters in an ensemble and to form the refined matrix.

The LCE includes three steps: i)create a base cluster to from a cluster ensemble ii)generate a refined matrix using the link based similarity approach iii)produce a final data partition. It achieves superior clustering results compared to any cluster ensemble techniques. The ranking method is applied for the categorical data clustering to place the appropriate data points into the cluster ensembles.

III.CONCLUSION

Clustering is a challenging task in data mining. 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. The clustering result is improved by applying ranking to the cluster ensembles by finding similarity between the cluster data points.

ACKNOWLEDGMENT

The authors would like to thank the Editor-in-Chief, the Associate Editor and anonymous Referees for their comments.

REFERENCES

[1]Aranganayagi.S and Thangavel. K "Clustering Categorical Data using Bayesian Concept",International Journal of Computer Theory and Engineering, Vol. 1, No.2,June2009

[2]Aranganayagi.S and Thangavel.K "Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure" International Journal of Information and Mathematical Sciences 6:1 2010.

[3]Gan.G, Yang.Z, and Wu.J (2005), "A Genetic k-Modes Algorithm for Clustering for Categorical Data",ADMA, LNAI 3584, pp. 195–202.

[4]Guha. S, Rastogi.R, and Shim.K (2000). ROCK: A robust clustering algorithm for categorical attributes',Information System., vol. 25, no. 5, pp. 345–366.

[5]Hsu.C.C., & Huang,Y.P., "Incremental Clustering of Mixed Data Based on the Distance Hierarchy", Expert System with Applications,(2007),doi:10.1016/j/eswa 2007.08.049

[6]Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price " A Link-Based Approach to the Cluster Ensemble Problem",IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 12, December 2011

[7]Ohn Mar San, Van-Nam Huynh, Yoshiteru Nakamori, "An Alternative Extension of The K-Means algorithm For Clustering Categorical Data", J. Appl. Math. Comput. Sci, Vol. 14, No. 2,2004, 241–247.

[8]Periklis Andristos, "Clustering Categorical Data based On Information Loss Minimization", EDBT 2004: 123-146.

[9]Srinivasulu Asadi , Ch. D.V. Subba Rao , C. Kishore and Shreyash Raju "Clustering the Mixed Numerical and Categorical Datasets Using Similarity Weight and Filter Method",VSRD-IJCSIT, Vol. 2 (5), 2012 2012

[10]Visakh. R Lakshmipathi. B "Constraint based Cluster Ensemble to Detect Outliers in Medical Datasets" International Journal of Computer Applications,Volume 45– No.15, May 2012

[11]Zhexue Huang , A Fast Clustering Algorithm to cluster Very Large Categorical Datasets in Data Mining, In Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[12]Zhexue Huang, "Extensions to the K-means algorithm for clustering Large Data sets with categorical value", Data Mining and Knowledge Discovery 2, Kluwer Academic publishers, 1998. 283-304.

Ms C.M.Geetha currently pursuing her Master of Engineering in Software Engineering at SNS College of Technology, affiliated to Anna University Chennai, Tamilnadu, India. She received her B.Tech degree from Vivekanandha college of Engineering for women affiliated to Anna University Coimbatore. Her research interest includes data mining and networking.

Mrs. K.Sangeetha is presently Assistant Professor at Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Chennai, Tamilnadu, India. She received the B.E degree from Sasuri College of Engineering and M.E degree from Nandha Engineering college. Currently she is pursuing her doctoral degree in Anna University Chennai. Her research interests includes datamining and networking. She published papers in international journals , national and international conference.

Professor Dr.S.Karthik is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University-Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Anna University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.