International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(2), pp.024-028
DOI: http://dx.doi.org/10.21172/1.82.004
e-ISSN:2278-621X

CLUSTER ENSEMBLE - A TECHNICAL REVIEW

Tanushree Bhimanwar[1] and Gauri Chaudhary[2]

Abstract *:* In Data mining, Clustering is useful to discover distribution patterns in the underlying data. Clustering algorithms usually employ a distance metric based (e.g., Euclidean) similarity measure in order to partition the database such that data points belonging to same partition are more similar than points in different partitions. When the data is categorical, Clustering becomes more challenging problem, that is, when there is no inherent distance measure between data values. Various clustering algorithms are developed to cluster or categorize the datasets. Some algorithms cannot be directly applied for clustering of categorical data. The underlying ensemble information matrix presents only cluster data point relations, with many entries being left unknown. This paper presents an analysis that shows this problem degrades the quality of the clustering result, and it presents a new link-based approach, which improves the conventional matrix through similarity between clusters in an ensemble, by discovering unknown entries. In particular, an efficient link-based algorithm is used to measure the underlying similarity assessment. Hence, to obtain the final clustering result, a graph partitioning technique is applied to a weighted bipartite graph that is obtained from the refined matrix.

Keywords**-** Clustering, categorical data, cluster ensembles, data mining, similarity measures.

I. INTRODUCTION

One of fundamental tool for understanding the structure of a data set is Data Clustering which plays an important, foundational role in machine learning, data mining, information retrieval, and pattern recognition. Principally, clustering aims to categorize data into groups or clusters such that data in the same cluster are more similar to each other than to those in different clusters, with the underlying structure of real-world datasets containing a combination of shape, size and density. Every clustering algorithm implicitly or explicitly assumes a certain data model and it may produce results which may be erroneous or meaningless when these assumptions are not satisfied by the sample data. Thus, the previously available information about the data domain is necessary for successful clustering, though such information can be difficult to obtain, even from experts.

To establish the sample data's compliance with the underlying distributions or, at least, to the proper number of clusters [1] Identification of relevant subspaces or visualization may help. The investigative nature of clustering tasks demands efficient methods that would be beneficial from combining the strengths of many individual clustering algorithms. This is the focus of the research on cluster ensemble, look for the combination of multiple partitions that provides refined overall clustering of the given data.

Clustering ensembles is beyond what is typically achieved by a single clustering algorithm in several aspects:

[1] *Student, CSE Department, YCCE, Nagpur, Maharashatra, India*
[2] *Associate Professor, CSE Department YCCE, Nagpur, Maharashtra, India*

- Robustness: Across the domains and datasets, it has better average performance.

- Novelty: Finding a combined solution unreachable by any single clustering algorithm.

- Stability and confidence estimation: Clustering solutions have lower Sensitivity to noise, outliers, or sampling variations. Clustering uncertainty can be accessed using ensemble distributions.

- Scalability and Parallelization: Cluster ensemble have ability to combine solutions from multiple distributed sources of data or attributes (features) [1] Parallel clustering of data subsets with subsequent combination of results..

Recently, the cluster ensemble approach has come out as an effective solution that is able to overcome these problems. Cluster ensemble methods integrate multiple clustering of the same dataset to get a single overall clustering. It has been found that such a practice can improve robustness, as well as the quality of clustering results. Thus, the main objective of cluster ensembles is to combine different solutions of various clustering algorithms in such a way to achieve the accuracy superior to those of individual clustering.

[3-4]Despite of notable success, these methods generate the final data partition based upon incomplete information of a cluster ensemble. The unrevealed ensemble information matrix presents only cluster data point relationships while completely ignores relationship among clusters. Link based approach to refining the former matrix, giving substantially less unknown entries. This research uniquely bridges the gap between the task of data clustering and that of link analysis. It also strengthens the capability of ensemble methodology for categorical data, that has not received much attention in the literature. In this work, we study different existing cluster ensemble methods for clustering categorical data and briefly discuss among the merits and demerits to improve the results of categorical data.

II. LITERATURE REVIEW

In many real world application domains such as data mining, data compression and pattern recognition, Clustering analysis has been widely applied. However, there is a combinable optimization problem and no single clustering algorithm is available that can provide satisfactory clustering solutions for all types of data sets. Numbers of clustering algorithms are available; some of them often produce divergent clustering solutions. Cluster ensembles are supposed to be a robust and most perfect alternative to single clustering runs. It also provides for a visualization tool to test cluster number, membership, and boundaries. In this sense ensemble clustering is a potential approach to generate more accurate clusters that might be possible using an individual clustering approach [6].

[7]Knowledge based Cluster Ensemble technique mainly integrates the aforementioned knowledge of the information in the dataset into the cluster ensemble process. In particular the previous knowledge about the data is illustrated in the Pair wise constrains in which it helps to improve the quality and the accuracy of the clustering results. Hence this KCE method achieves the best performance for the cancer datasets, Novartis multi-tissue dataset, SRBCT dataset and St. Jude dataset.

[8]In this ensemble method Locally Adaptive Clustering algorithm was used and it finds clusters in subspaces spanned by different combinations of dimensions through local weightings of features. The Locally Adaptive clustering has benefit that it avoids the risk of loss of information detected in global dimensionality reduction techniques. A Weighted cluster is an ensemble method which is a subset of data points together with a vector of weights such that the points in the cluster are close to each other. Weighted Similarity Partitioning Algorithm [9] In this algorithm, it assigns only low similarity values to both pairs of a for data set, higher similarity values also very important for clustering data, it is not considered in this work, Many clustering algorithms may work efficient either on pure numeric data or on pure categorical data, but 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. In clustering the large data sets, the scalability and memory constraint is a problem. The clustering algorithm based on similarity weight and filter method model that works well for data with mixed numeric and categorical features. To cluster the categorical data, the incremental clustering algorithm is used. The incremental algorithm is more effective than other clustering algorithm. This incremental algorithm works efficiently even if the boundaries of clusters are irregular. The advantage is that we combine the different clustering datasets with different algorithms.

[13-14]This Squared Error Adjacent Matrix algorithm is mainly based upon the similarity matrix which is defined as the co-association matrix. It has the high possibility of finding the final data partition without previously knowing the number of clusters or any value of the thresholds when similarity matrix is given. In this algorithm the value of the similarity is assumed then the formation of cluster also less therefore assuming similarity values is not appropriate to all dataset.

[15]Hybrid fuzzy cluster ensemble method is appropriate method that is used for tumor clustering from the cancer gene expression datasets, and best clustering results are produced for cancer dataset only by using this method, and using this method for other types of dataset such intrusion detection dataset, 20 news group dataset, mushroom dataset, where identification of clusters data points becomes more complex and we get less clustering results.

[16]The challenging problem in clustering large data set is scalability and memory constraint. The new incremental algorithm is used to cluster the categorical data. Incremental algorithm takes less computational time to find clusters. Categorical data is the one which cannot be ordered and with limited domains. In general the incremental algorithms generate large number of clusters; naturally it has more purity, whereas the proposed measures generate less number of clusters with high purity. The major issue of this work is only applied to single clustering methods the entire data is given as input to process and clustering ensemble is not performed in this work, when compare to normal clustering methods clustering ensemble produces best clustering results. The accuracy of the clustering is less than the ensemble methods.

[9]Projective clustering aims to discover clusters which correspond to subsets of the input data and have different (possibly overlapping) dimensional subspaces associated to them. Francesco Gullo et al [14] problem of projective clustering ensembles (PCE) is addressed for the first time. The objective is to define methods for clustering ensembles that are able to deal with ensembles of projective clustering solutions and provide a projective consensus partition. In particular, focus on ensembles composed by axis-aligned projective clustering solutions. The projective consensus partition to be discovered is computed as a solution of an optimization problem formulated by exploiting information available from the input ensemble. Clustering similarity measurements is not performed in this work, so the clustering results are measured correctly, the efficiency of the clustering results degrades and more time complexity.

Co-clustering has emerged as an important technique for mining contingency data matrices. However, almost all existing co-clustering algorithms are hard partitioning, which assigns each row and column of the data matrix to one cluster. Recently a Bayesian co-clustering approach has been presented which allows a probability distribution [1] membership in row and column clusters. The approach uses variation inference for parameter estimation. Pu Wang et al proposed a nonparametric Bayesian approach to co-clustering ensembles is presented. Co-clustering ensembles combine various base co-clustering results to obtain a more robust consensus co-clustering, which is Similar to clustering ensembles. To avoid pre-specifying the number of co-clusters, specify independent Dirichlet process priors for the row and column clusters. Thus, the numbers of row and column-clusters are unbounded a priori; and the actual numbers of clusters can be learned a posteriori from observations. Next, we employ a Mondrian Process as a prior distribution over partitions of the data matrix, to model non-independence of row- and column-clusters.

III. INFERENCE FROM EXISTING SOLUTION

The main disadvantage of clustering ensemble methods for categorical data clustering is discussed below:

KCE method achieves the best

Performance in majority of the cancer datasets, along with the Novert is multitissue dataset, SRBCT dataset and St. Jude dataset. Other types of categorical dataset are support in this work

 Squared Error Adjacent Matrix , if value of the similarity is assumed then the formation of cluster also less hence assuming similarity values is not appropriate to all dataset.

 Projective clustering ensembles (PCE) Clustering similarity measurements is not performed in this work, so the clustering results are measured correctly, the efficiency of the clustering results degrades and more time complexity

 Hybrid Fuzzy Ensemble produce best clustering results for cancer dataset only and identification of clusters data points becomes more complex and have less clustering results in other types of dataset such as mushroom dataset, 20 news group dataset, intrusion detection dataset.

 Adjusted Rand Index measure is highly meaningful in examining the cluster performance without the underlying labels rather than having few similarity matrices only between the partitions. Duplicate data needs to reduce the cluster result or irrelevant data present in the system.

 The random selection of starting centers in this algorithm may show different clustering results and falling into less clustering results.

IV. THE BASIC PROCESS OF CLUSTER ENSEMBLE

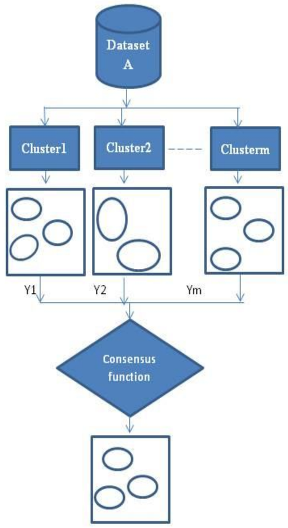

Fig.1 Cluster Ensembling

Let X={X1….Xn} be a set of N data points and Y={Y1…Ym) be a cluster ensemble with M base clustering, each of which is referred to as an ensemble member. Is the number of clusters in the ith clustering. The problem is to find a new partition of a data set X that condense the information from the cluster ensemble

Fig 1. shows the general framework of cluster ensembles. To form a final partition, solutions achieved from different base clustering are aggregated. This meta level methodology involves two major tasks of:

1) Generating a cluster ensemble, and

2) Producing the final partition normally referred to as a Consensus function.

V. CONCLUSION

Link based similarity measure are used to measure the cluster ensemble results for cluster ensemble methods, instead of random selection of cluster centroid values which automatically select centroid values, clustering similarity measure also need to improve using other similarity measurements. The main objective of cluster ensembles is to integrate different clustering decisions in such a way that achieve accuracy superior to that of any individual clustering.

REFERENCE

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

[2] P. Zhang, X. Wang, and P.X. Song, “Clustering Categorical Data Based on Distance Vectors,” in proceeding *The J. Am. Statistical Assoc.*, vol. 101, no. 473, 2006, pp. 355-367.

[3] J. Grambeier and A. Rudolph, “Techniques of Cluster Algorithms in Data Mining,” in proceeding *Data Mining and Knowledge Discovery*, vol. 6, 2002, pp. 303-360.

[4] Z. Huang, “Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values,” in proceeding *Data Mining and Knowledge Discovery*, vol. 2, 1998, pp. 283-304.

[5] L. Getoor and C.P. Diehl, “Link Mining: A Survey,” in proceeding *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, 2005, pp. 3-12.

[6] G. Das, H. Mannila, and P. Ronkainen, “Similarity of Attributes by External Probes,” in proceeding *ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD*), 1998 pp. 16-22,.

[7]Zhiwen Yu, Hau-San Wongb, Jane You, Qinmin Yang, and Hongying Liao, “ Knowledge based Cluster Ensemble for Cancer Discovery From Biomolecular Data” , *IEEE Transactions on Nanobioscience*, Vol.10, No. 2, 2011, pp.76-85.

[8]Domeniconi C, Al-Razgan M, “Weighted Cluster Ensembles: Methods and Analysis”, in proceeding *ACM Transactions on Knowledge Discovery from Data*, Vol.2, No.4, 2009, pp.1-40,.

[9] F. Gullo, C. Domeniconi, and A. Tagarelli, “ Projective clustering ensembles”, In proceeding *IEEE International Conference on Data Mining*, 2009, pp. 794799.

[10] P. Wang, C. Domeniconi, and K. Laskey, “Latent Dirichlet Bayesian coclustering”, In Proceedings *Springer of the European Conference on Machine Learning*, Vol.5782, Berlin Heidelberg 2009, pp. 522537.

[11]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”, in proceeding *VSRD-IJCSIT*, Vol. 2 ,No.5, 2012, pp.1-2.

[12]Yang Lili, Yu Jian, & JIA Caiyan, “A New method for Cluster Ensembles”, in proceeding *Programs Foundation of Ministry of Education of China*,2013.

[13] Sarumathi S, Shanthi N, Sharmila M, “ A Comparative Analysis of Different Categorical Data Clustering Ensemble Methods in Data Mining”, in procceding *International Journal of Computer Applications*, Vol. 81, No.4, 2013, pp.4656.

[14]Zhiwen Yu, Hantao Chen Jane You, Guoqiang Han Le Li ,“ Hybrid Fuzzy Cluster Ensemble Framework for Tumor Clustering from Biomolecular Data”, in proceeding *IEEE Transactions on computational biology and bioinformatics* , Vol.10 , No.3, 2013, pp. 657-670,.

[15]Aranganayagi.S and Thangavel.K, “Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure”, in proceeding *International Journal of Information and Mathematical Sciences*, Vol.6,No.1, 2010, pp.1-8.

[16]P. Wang, C. Domeniconi, and K. B. Laskey, "Nonparametric Bayesian Co-clustering Ensembles", in proceeding *Workshop on Nonparametric Bayes, held in conjunction with NIPS, Whistler, BC, Canada*, Vol.11, No.12, December 2009, pp.331-342.