ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 3 , March 2015

A Thorough Investigation of Link-Based Cluster Ensemble Approach for Data Clustering

N. Yuvaraj , Dr. C .Suresh Gnana Dhas

Research Scholar, Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, St.Peter's University, Avadi , Chennai.

Professor & Head, Vivekanandha College of Engineering for Women, Tiruchencode, India

ABSTRACT: Clustering, in data mining, 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 in the same partition are more similar than points in different partitions. The problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measure between data values. 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. Cluster ensemble has proved to be a good alternative when facing cluster analysis problems. It consists of generating a set of clustering's from the same dataset and combining them into a final clustering. The goal of this combination process is to improve the quality of individual data clustering's. This paper presents an overview of clustering ensemble methods that can be very useful for the categorical data clustering. The characteristics of several methods are discussed, which may help in the selection of the most appropriate one to solve a problem at hand. 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. Secondly, we describe several clustering ensembles methods such as including Cluster Ensemble technique, Squared Error Adjacent Matrix algorithm, Hybrid Fuzzy Ensemble, and next explain their advantages, disadvantages and computational complexity. Finally, we compare the characteristics of clustering ensembles algorithms such as computational complexity, simplicity and accuracy on different datasets in previous techniques.

KEYWORDS: Data clustering cluster ensembles, link-based similarity and categorical data.

I. INTRODUCTION

Data clustering is a common task, which plays a crucial role in various application domains such as machine learning, data mining, information retrieval, pattern recognition and bioinformatics. 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 bewildering combination of shape, size and density. Every clustering algorithm implicitly or explicitly assumes a certain data model and it may produce erroneous or meaningless results when these assumptions are not satisfied by the sample data. Thus, the availability of prior information about the data domain is crucial for successful clustering, though such information can be hard to obtain, even from experts. Identification of relevant subspaces or visualization may help to establish the sample data's conformity to the underlying distributions or, at least, to the proper number of clusters [1]. The exploratory nature of clustering tasks demands efficient methods that would benefit from combining the strengths of many individual clustering algorithms. This is the focus of the research on clustering ensembles, seeking a combination of multiple partitions that provides improved overall clustering of the given data. Clustering ensembles can go beyond what is typically achieved by a single clustering algorithm in several respects:

- Robustness: Better average performance across the domains and datasets.
- Novelty: Finding a combined solution unattainable by any single clustering algorithm.
- Stability and confidence estimation: Clustering solutions with lower sensitivity to noise, outliers, or sampling variations. Clustering uncertainty can be accessed from ensemble distributions.
- Parallelization and Scalability: Parallel clustering of data subsets with subsequent combination of results. Ability to integrate solutions from multiple distributed sources of data or attributes (features) [1].

Recently, the cluster ensemble approach has emerged as an effective solution that is able to overcome these problems. Cluster ensemble methods combine multiple clustering of the same dataset to yield 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 decisions of various clustering algorithms in such a way to achieve the accuracy superior to those of individual clustering. Examples of well- known ensemble methods are: (i) the feature-based approach that transforms the problem of cluster ensembles to clustering categorical data, i.e., cluster labels [2], (ii) graph-based algorithms that employ a graph partitioning methodology [3-4] .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.

Link-based approach to refining the aforementioned matrix, giving substantially less unknown entries. A link-based similarity measure [5] 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 this work, we study different existing cluster ensemble methods for categorical data clustering and briefly discuss among the advantages and disadvantages to enhance the results of categorical data.

A. SURVEY OF LINK BASED CLUSTERING ENSEMBLE METHODS

Clustering analysis has been widely applied in many real world application domains such as data compression, data mining and pattern recognition. However, it is in fact an ill posed combinatory optimization problem and no single clustering algorithm is able to achieve satisfactory clustering solutions for all types of data sets. Numbers of clustering algorithms exist so far and some of them often produce contradictory 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 examine cluster number, membership, and boundaries. In this sense ensemble clustering is a potential approach to generate more accurate clusters than might be possible using an individual clustering approach [6]. It generally involves two major tasks as Generation step in which generating several clustering solutions by applying clustering algorithm are done and the Consensus step through which final cluster partition is produced.

Knowledge based Cluster Ensemble technique [2] mainly integrates the prior knowledge of the information in the dataset into the cluster ensemble process. In particular the prior knowledge about the data is illustrated in the Pair wise constrains in which it helps in enhancing the quality and the accuracy of the clustering results. Hence this KCE method achieves the best performance in majority of the cancer datasets, along with the Novertis multi-tissue dataset, SRBCT dataset and St. Jude dataset. Other types of categorical dataset are support in this work.

A Weighted cluster is a subset of data points together with a vector of weights such that the points in the cluster are close to each other. In this ensemble method [7] Locally Adaptive Clustering algorithm was used and it discovers clusters in subspaces spanned by different combinations of dimensions through local weightings of features. The major benefit of this Locally Adaptive clustering was that it avoids the risk of loss of information encountered in global dimensionality reduction techniques. Weighted Similarity Partitioning Algorithm [7] in which it assigns only low similarity values to both pairs of a data set, higher similarity values also important for clustering data, it is not considered in this work .

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 [8] 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 works efficiently even if the boundaries of clusters are irregular. The advantage is that we mix the different clustering datasets with different algorithms.

How to combine the multiple data partitions to get a consistent partition for a given data set using the information obtained in the different clustering results. This Squared Error Adjacent Matrix algorithm [9-10] is mainly based upon the similarity matrix which is defined as the co-association matrix. It has the high potential of finding the final data partition without predefining the number of clusters or any value of the thresholds when similarity matrix is given. The value of the similarity is assumed then the formation of cluster also less since assuming similarity values is not appropriate to all dataset.

Cluster ensemble is a promising technique for improving the clustering results. An alternative to generate the cluster ensemble is to use different representations of the data and different similarity measures between objects. This way, it is produced a cluster ensemble conformed by heterogeneous partitions obtained with different point of views of the faced problem. This diversity enhances the cluster ensemble but, it restricts the combination process since it makes difficult the use of the original data.

S. Vega-Pons et al [11] proposes a unified representation of the objects taking into account the whole information in the cluster ensemble. This representation allows working with the original data of the problem regardless of the used generation mechanism. Also, this new representation is embedded in the WKF algorithm making a more robust cluster ensemble method. The main goal of the construction of the matrix WC is for obtaining a new representation of the objects that allows using the WKF method without any constraint in the generation step. This algorithm work well to discrete attributes and it doesn't support for categorical data clustering, it reduces the accuracy of the cluster ensemble data.

Hybrid Fuzzy Ensemble method is mainly proposed for enhancing the performance and quality of the tumor clustering from bio-molecular dataset. Here fuzzy theory is implemented into the cluster ensemble paradigm in order to accurately denote the samples corresponding to different types of cancer data. Fuzzy theory is mainly used to generate the fuzzy matrices in the ensemble. HFCE-I method [12] uses the Affinity Propagation (AP) algorithm to extract the base clustering results on sample dimension of the dataset. This in turn exemplifies the fuzzy matrices in the ensemble which is based upon the fuzzy membership functions. Initially the base samples are randomized by the AP algorithm. Hybrid fuzzy cluster ensemble method is well suited for performing the tumor clustering from the cancer gene expression datasets. It produces best clustering results for cancer dataset only ,other types of dataset such 20 news group dataset ,mushroom dataset , intrusion detection dataset identification of clusters data points becomes more complex and less clustering results .

Scalability and memory constraint is the challenging problem in clustering large data set. The new incremental algorithm [13] is used to cluster the categorical data. Incremental algorithm finds clusters in less computation time. Categorical data is the one which cannot be ordered and with limited domains. 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 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.

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.

Application of ensembles to co-clustering, the problem of simultaneously clustering the rows and columns of a data matrix into row and column-clusters to achieve homogeneity in the blocks in the induced partition of the data matrix. Co-clustering has emerged as an important technique for mining contingency data matrices. However, almost all existing co-clustering algorithms are hard partitioning, assigning each row and column of the data matrix to one cluster. Recently a Bayesian co-clustering approach has been proposed which allows a probability distribution membership in row and column clusters. The approach uses variational inference for parameter estimation.

Pu Wang et al [15] proposed a nonparametric Bayesian approach to co-clustering ensembles is presented. Similar to clustering ensembles, co-clustering ensembles combine various base co-clustering results to obtain a more robust consensus co-clustering. 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; the actual numbers of clusters can be learned a posteriori from observations. Next, to model non-independence of row- and column-clusters, we employ a Mondrian Process as a prior distribution over partitions of the data matrix.

(1) Dirichlet process-based co-clustering ensemble model (DPCCE), which assumes independent Dirichlet process mixture priors for rows and columns;

(2) A Mondrian process-based co-clustering ensemble model (MPCCE) that places a Mondrian process prior over the matrix partitions. For both the DPCCE and the MPCEE, the number of blocks is not fixed a priori, but is open-ended and inferred from the data.

P. Wang et al [16] extends the BCC model and proposes a collapsed Gibbs sampling and a collapsed variational Bayesian algorithm for it. Smooth the BCC model, by introducing priors for the entry value distributions given row- and column-clusters. Latent Dirichlet Bayesian Co-Clustering (LDCC), since it assumes Dirichlet priors for row- and column-clusters, which are unobserved in the data contingency matrix. The collapsed Gibbs sampling and collapsed variational Bayesian algorithms we propose can learn more accurate likelihood functions than the standard variational Bayesian algorithm.

Gibbs sampling leads to unbiased estimators, it also has some drawbacks: one needs to assess convergence of the Markov chain and to have some idea of mixing times to estimate the number of samples to collect, and to identify coherent topics across multiple samples. In practice, one often ignores these issues and collects as many samples as is computationally feasible, while the question of topic identification is often sidestepped by using just one sample. Hence, there still is a need for more efficient, accurate and deterministic inference procedures. Improvement of the multi-objective cluster ensemble algorithm which is expressed as IMOCLE [17] was proposed. This method mainly shows the superiority of the other techniques and the capability of finding the optimum number of clusters and accuracy. Optimum number of clusters only founds, how weak cluster are converted into best cluster still becomes major issue.

Adjusted Rand Index [18] was proposed between similarity matrix and cluster partition to measure the consistency between the different set of clustering results and their associated consensus matrix in a cluster ensemble. It formulation of two new measures, ARImp and ARImm, which allow the effective comparison between clustering solutions and consensus matrices, and that between consensus matrices respectively Desirable properties of ARI are preserved in the two new measures. This measure is highly meaningful in analysing the cluster performance without the underlying labels rather than with few similarity matrices between the partitions. The results of clustering are not measured accurately.

II. 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 multi-tissue dataset, SRBCT dataset and St. Jude dataset. Other types of categorical dataset are support in this work
- Squared Error Adjacent Matrix ,value of the similarity is assumed then the formation of cluster also less since 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
- Gibbs sampling leads to unbiased estimators, it also has some drawbacks: one needs to assess convergence of the Markov chain and to have some idea of mixing times to estimate the number of samples to collect, and to identify coherent topics across multiple samples.
- Hybrid Fuzzy Ensemble best clustering results for cancer dataset only ,other types of dataset such 20 news group dataset ,mushroom dataset , intrusion detection dataset identification of clusters data points becomes more complex and less clustering results .
- Adjusted Rand Index measure is highly meaningful in analyzing the cluster performance without the underlying labels rather than with 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 lead to different clustering results and falling into less clustering results.

III.SOLUTION TO OVERCOME THESE ISSUES

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.

IV. EXPERIMENTAL RESULTS

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-k), and Type-III (Random- k). The k-modes clustering algorithm 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 .

Table 1: Pairwise Performance Comparison among Examined Clustering Methods

Ensemble type	Methods	Classification accuracy(CA)	Normalized Mutual Information (NMI)	Adjusted Rand (AR)
I	Optimized LCE	187	146	164
	LCE	170	137	149
	CO+SL	34	81	47
	CO+AL	72	114	84
	CSPA	105	132	109
	HGPA	21	19	24
II – Fixed K	Optimized LCE	225	228	212
	LCE	208	204	201
	CO+SL	26	37	28
	CO+AL	131	134	134
	CSPA	86	68	82
	HGPA	64	93	93
II – Random K	Optimized LCE	235	214	219
	LCE	209	203	201
	CO+SL	17	38	28
	CO+AL	97	94	117
	CSPA	76	47	51
	HGPA	67	41	49
III –Fixed K	Optimized LCE	219	203	197
	LCE	197	191	182
	CO+SL	17	37	32
	CO+AL	115	119	139
	CSPA	73	53	61
	HGPA	71	56	65
III –	Optimized	227	206	196

Random K	LCE			
	LCE	203	191	181
	CO+SL	15	34	32
	CO+AL	81	81	103
	CSPA	52	36	45
	HGPA	51	38	63

The parameter analysis aims to provide a practical means by which users can make the best use of the link-based framework. Essentially, the performance of the resulting technique is dependent on the decay factor (i.e., $DC \in [0, 1]$, which is used in estimating the similarity among clusters. The varied the value of this parameter from 0.1 through 0.9, in steps of 0.1, and obtained the results in Fig. 1,2,3,4 and ,5 . Note that the presented results are obtained with the ensemble size (M) of 10. The figure clearly shows that the results of optimized LCE are robust across different ensemble types, and do not depend strongly on any particular value of DC. This makes it easy for users to obtain high-quality, reliable results, with the best outcomes being obtained with values of DC between 0.7 and 0.9. The relations between $DC \in \{0.1,0.2,, \dots 0.9\}$, and the performance of the optimized LCE models (the averages across all validity indices and six data sets), whose values are presented in X-axis and Y-axis.

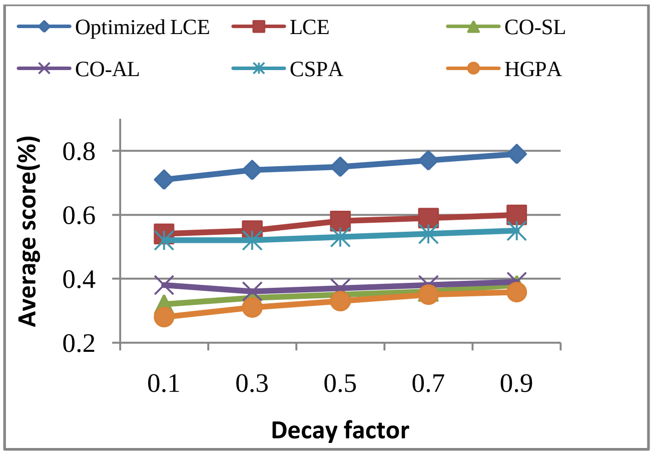

Fig.1. Ensemble Type I

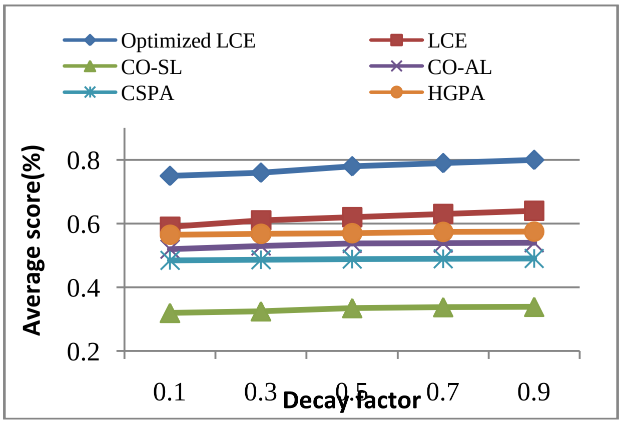

Fig. 2. Ensemble Type II-Fixed-k

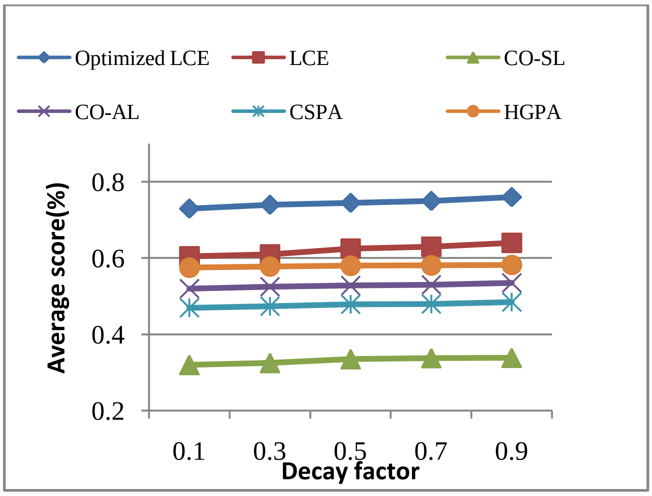

Fig.3. Ensemble Type II-Random –k

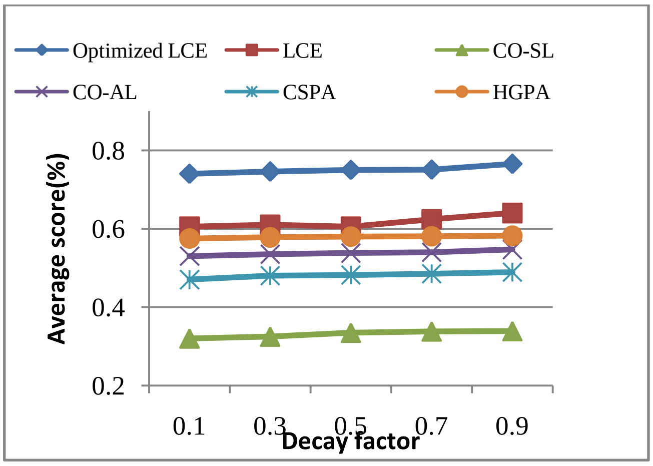

Fig.4. Ensemble Type III-Fixed-k

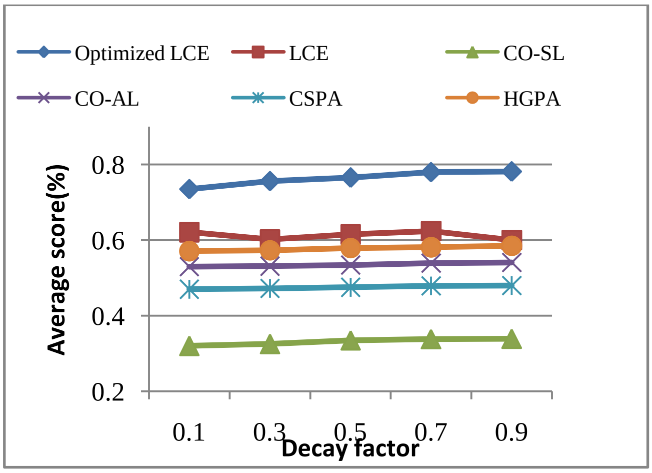

Fig.5. Ensemble Type III-Random-k

Performance of different cluster ensemble methods in accordance with ensemble size ($M \in (10, 20, .., 100)$), as the averages of validity measures (CA, NMI, and AR) shown in figure 6 ,7 ,8 and 9 .It shows that proposed optimized LCE have better results than existing methods .

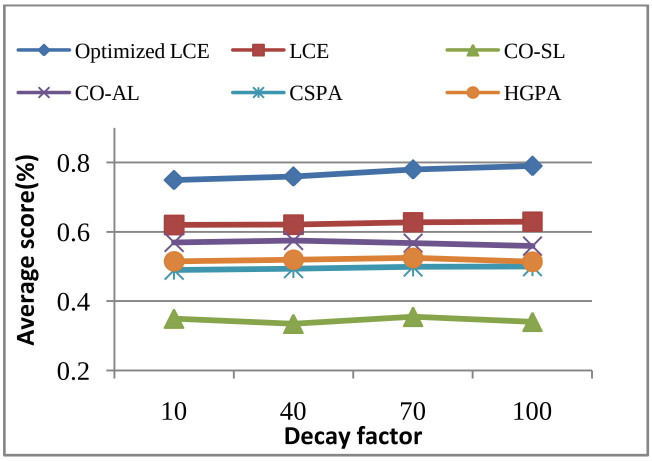

Fig. 6.Ensemble Type II-Fixed-k

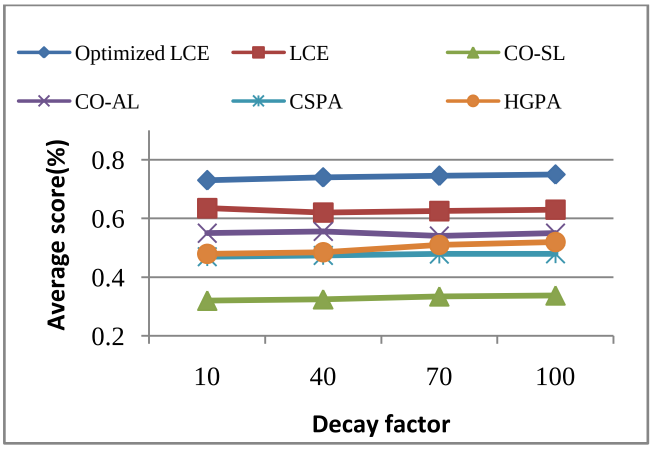

Fig.7. Ensemble Type II-Random –k

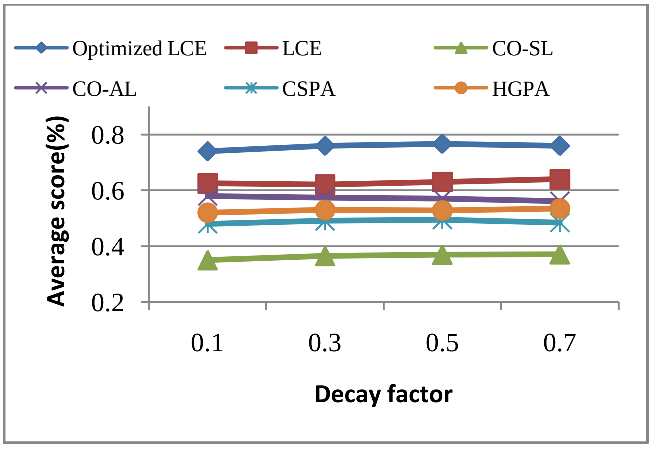

Fig.8. Ensemble Type III-Fixed-k

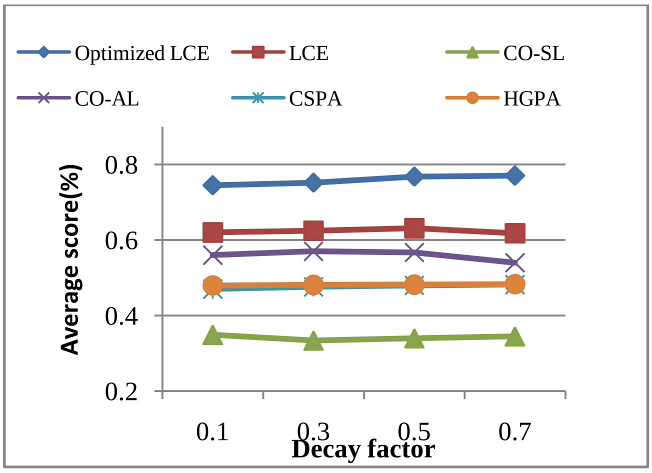

Fig. 9. Ensemble Type III-Random-k

V. CONCLUSION AND FUTURE WORK

Clustering ensembles have emerged as a prominent method for improving robustness, stability and accuracy of unsupervised classification solutions. There are several challenges for clustering ensemble that one of the major problems in clustering ensembles is the consensus function .Consequently this proportional study reveals some of the different categorical cluster ensemble approaches including their similarity measurements functions along with the average accuracy and error rates of each technique with comparative table denotes differential analysis and limitations of the diverse ensemble methods along with the graphical representation of the accuracy levels of different types of cluster ensembles are investigated in this evaluation: Type-I, Type-II (Fixed-k), Type-II (Random-k), Type-III (Fixed-k), and Type-III (Random- k). The comparison result proves that the many of the proposed works in cluster ensemble technique faces accuracy problem on different real world and artificial datasets. Compared accuracy on different datasets in previous techniques LCE, CO+SL, CO+AL, CSPA, Hyper-Graph Partitioning Algorithm HGPA and proposed optimization based LCE. This investigation makes better understanding for the readers and also hopes to be more useful for the society of clustering researchers to perform efficient clustering ensemble methods. Generally, most of the clustering ensembles techniques need to improve their accuracy, therefore improving of accuracy can be an important research in future by using different similarity metrics and clustering methods.

REFERENCES

[1] Nguyen N, Caruana R, “Consensus Clusterings" , In Proceedings of IEEE International Conference on Data Mining, pp. 607-612. IEEE Computer Society, Washington, DC,2007.

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

[3] Iam-on N, Boongoen T, Garrett S, “LCE: A Link-Based Cluster Ensemble Method for Improved Gene Expression Data Analysis", Bioinformatics, Vol.26,No.12, pp.1513-1519,2010.

[4] T. Boongoen, Q. Shen, and C. Price, “Disclosing False Identity through Hybrid Link Analysis,” Artificial Intelligence and Law, Vol. 18, No. 1, pp. 77-102, 2010.

[5] Harun Pirim, Dilip Gautam, Tanmay , Bhowmik, Andy D. Perkins, Burak Ekşioglu, & Ahmet Alkan,. “Performance of an ensemble clustering algorithm on biological datasets”. Mathematical and Computational Applications, Vol. 16, No. 1, pp. 87-96 2011.

[6] 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, pp.76-85, 2011.

[7] 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 ,No.5, pp.1-2, 2012.

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

[9] S. Vega-Pons, J. Ruiz-Shulcloper, Clustering ensemble method for hetero- geneous partitions, in: E. Bayro-Corrochano, J.-O. Eklundh (Eds.), CIARP 2009, 5856, Lecture Notes in Computer Science, pp. 481–488,2009.

[10] Sarumathi S, Shanthi N, Sharmila M, “ A Comparative Analysis of Different Categorical Data Clustering Ensemble Methods in Data Mining” , International Journal of Computer Applications, Vol. 81, No.4 , pp.46-56, 2013.

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

[12] 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.

[13] F. Gullo, C. Domeniconi, and A. Tagarelli, “ Projective clustering ensembles”, In IEEE International Conference on Data Mining, pp. 794-799, 2009.

[14] P. Wang, C. Domeniconi, and K. B. Laskey, "Nonparametric Bayesian Co-clustering Ensembles", Workshop on Nonparametric Bayes, held in conjunction with NIPS, Whistler, BC, Canada, December 11-12,pp.331-342, 2009.

[15] P. Wang, C. Domeniconi, and K. Laskey, “Latent Dirichlet Bayesian co-clustering”, In Proceedings of the European Conference on Machine Learning, Vol.5782, pp. 522-537. Springer Berlin Heidelberg, 2009.

[16] Ruochen Liu, Yong Liu, Yangyang Li， “An Improved Method for Multi-Objective clustering Ensemble Algorithm”, IEEE World Congress on Computational Intelligence ,pp.1-8, 2012.

[17] Shaohong Zhang, Hau-San Wong, “ARImp A Generalized Adjusted Rand Index for Cluster Ensembles”, International Conference on Pattern Recognition, IEEE Computer Society,pp 778 – 781,2010.