International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 03 (March 2015), PP.79-89

Ensemble based Distributed K-Modes Clustering

N. Karthikeyani Visalakshi[1], K. Arunprabha[2]
[1]*Department of Computer Applications, Kongu Engineering College, Perundurai, Tamil Nadu, India*
[2]*Department Computer Science, Vellalar College for Women, Erode, Tamil Nadu, India*

Abstract:- Clustering has been recognized as the unsupervised classification of data items into groups. Due to the explosion in the number of autonomous data sources, there is an emergent need for effective approaches in distributed clustering. The distributed clustering algorithm is used to cluster the distributed datasets without gathering all the data in a single site. The K-Means is a popular clustering method owing to its simplicity and speed in clustering large datasets. But it fails to handle directly the datasets with categorical attributes which are generally occurred in real life datasets. Huang proposed the K-Modes clustering algorithm by introducing a new dissimilarity measure to cluster categorical data. This algorithm replaces means of clusters with a frequency based method which updates modes in the clustering process to minimize the cost function. Most of the distributed clustering algorithms found in the literature seek to cluster numerical data. In this paper, a novel Ensemble based Distributed K-Modes clustering algorithm is proposed, which is well suited to handle categorical data sets as well as to perform distributed clustering process in an asynchronous manner. The performance of the proposed algorithm is compared with the existing distributed K-Means clustering algorithms, and K-Modes based Centralized Clustering algorithm. The experiments are carried out for various datasets of UCI machine learning data repository.

Keywords: - Distributed Clustering, K-Modes, Categorical data, Local Model and Global Model

I. INTRODUCTION

Data clustering is a prominent approach adopted to bring into play the partitioning operation. It bestows an intelligent way to find interesting groups when a problem becomes intransigent for human analysis. Clustering groups data objects based on the information that describes objects and their relationships. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters[22]. Clustering has been studied in the field of machine learning and pattern recognition and it plays an important role in data mining applications such as scientific data exploration, information retrieval, and text mining. It also has a significant function in spatial database applications, web analysis, customer relationship management, marketing, bio-medical analysis and many other related areas [43].

Traditionally, the clustering algorithms require full access to the data which are to be analyzed. These days, with the rapid growth and advancements, the ever-increasing computing power and computer storage capacity, the invasion of the internet into human routine and the increasingly automated business, manufacturing, and scientific processes, dataset sizes have grown briskly[27]. Furthermore, most of these datasets are distributed across multiple sites geographically. For instance, the huge number of sales records from hundreds of chain stores is stored at different locations. In order to administer the classical knowledge discovery process in distributed environments, collection of distributed data in a data warehouse is essential. However, this is normally unproductive or infeasible for the following issues: central storage requirements, communication overhead, computational cost, security and privacy, whereas, the alternate distributed clustering algorithms these issues by exchanging a few data and avoiding synchronization as much as possible[38].

Cluster ensemble[41] is a problem of combining multiple partitioning of a set of objects into a single consolidated clustering without accessing the features or algorithms that determined these partitioning. It has come up as an effective method for enabling data clustering and improving the stability as well as the robustness of unsupervised learning solutions. Several research have been proposed on ensemble based distributed clustering[5,14,31,40] based on K-Means, Expected Maximization (EM), density estimation and hierarchical clustering methods. In most of these cluster ensemble based algorithms, each clustering solution is represented by high resolution representations such as label vector, cluster sub-samples, dendogram, etc. which, in turn, suffer from time and memory complexity. However, only very few ensemble based distributed clustering algorithms[18,24,28] use low resolution representation, specifically, centroid or cluster center as local cluster model.
Clustering of datasets containing many categorical attributes is one of the most noticeable challenges in data mining. In recent years, limited attention has been paid in handling categorical data to obtain partitional clusters in centralized environment[1]. The K-Modes clustering algorithm[19] is an extension to the standard K-Means[15] clustering algorithm for clustering categorical data, with distance function, cluster center representation and the iterative clustering process being the major modifications to K-Means. Most existing distributed clustering

research [18, 24, 25, 27, 28] focuses on numerical datasets, and cannot directly apply to categorical sets where there is little sense in calculating distances among data points. Bin Wang et al.[3] proposed a distributed hierarchical clustering algorithm for categorical data called Coercion, which is based on Squeezer. However, their approach requires multiple rounds of message passing with full synchronization among distributed datasets and somehow increases communication overhead.

In this paper, by modifying Genlin's Distributed K-Means (DKM) algorithm[24], a new Ensemble based Distributed K-Modes (D-K-Md) clustering algorithm is proposed to handle categorical data and to perform distributed clustering process in asynchronous manner. The performance of the proposed algorithm is compared with the existing distributed clustering algorithms, DKM and Centralized Clustering (CC) based on K-Modes.

II. MATERIALS AND METHODS

K-Mode Clustering: The *K-Means* algorithm is well known for its efficiency in clustering large data sets[2]. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. Haung[20, 21] proposed K-Modes algorithm which extends the *K-Means* algorithm to categorical domains. In this algorithm three major modifications has made to the K-Means algorithm, i.e., using different dissimilarity measure, replacing K-Means with K-Modes, and using a frequency based method to update modes. These modifications guarantee that the clustering process converges to a local minimal result. Since the K-Means clustering process is essentially not changed, the efficiency of the clustering process is maintained.

The simple matching dissimilarity measure (Hamming distance) can be defined as following. Let *X* and *Y* be two categorical data objects described by m categorical attributes. The dissimilarity measure *d(X, Y)* between *X* and *Y* can be defined by the total mismatches of the corresponding attribute categories of two objects. Smaller the number of mismatches, more similar the two objects are. Mathematically, it can be defined as

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_i, y_j) \qquad (1)$$

where $\delta(x_i, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$, and $d(X,Y)$ gives equal importance to each category of an attribute. Let *N* be a set of *n* categorical data objects described by *m* categorical attributes, $M_1, M_2, \ldots, M_m$. When the distance function defined in Eq. (1) is used as the dissimilarity measure for categorical data objects, the cost function becomes

$$C(Q) = \sum_{i=1}^{n} d(N_i, Z_i) \qquad (2)$$

where N_i is the ith element and Z_i is the nearest cluster center of N_i. The K-Modes algorithm minimizes the cost function defined in Eq. (2). The K-Modes algorithm consists of the following steps [20]:

```
Algorithm.  K-Modes
Input  : Dataset X of n objects with d categorical attributes and number of clusters K, (K < n)
Output: Partitions of the input data into K clusters
Procedure
Step-1:  Randomly select K unique objects as initial modes, one for each cluster.
Step-2:  Calculate the distances between each object and cluster mode. Allocate the object to one of the k clusters whose mode is the nearest to it according to distance function (1).
Step-3:  Update the mode of each cluster based on the frequencies of the data objects in the same cluster.
Step-4:  Repeat step-2 and step-3 until convergence.
```

Figure-1 K-Modes Clustering Algorithm

Chaturvedi et al.[1] reported an equivalent approach to Huang's K-Modes algorithm. The relationship of the two K-Modes methods is described in[46]. Though few other methods for clustering categorical data were also proposed in last two decades, including the dynamic system approach[10], Cactus[8], ROCK[11], and Coolcat[4], these methods are largely stayed in research stage and not been widely applied to real world applications. However, different variants of K-Modes are proposed by Huang[21,44,45] and other researchers[7,16,32] to attain optimum performance in diverse application scenario. Hence an effort on K-Modes algorithm is taken here to make it suitable for clustering distributed categorical datasets.

Distributed Clustering: The development of distributed clustering algorithms is driven by factors like the huge size of many databases, the wide distribution of data, and the computational complexity of centralized clustering algorithms. Distributed clustering is based on the presumption that the data to be clustered are in different sites. This process is carried out in two different levels - the local level and the global level. In the local level, all the sites carry out clustering process independently, after which a local model such as cluster center or cluster index is determined, which should reflect an optimum trade-off between complexity and accuracy. Further, the local model is transferred to a central site, where the local models are merged in order to form a global model. The resultant global model is again transmitted to local sites to update the local models. Instead of local model, local representative samples may also be transmitted to reach global clusters[23].

The main intent of distributed data clustering algorithms is to cluster the distributed datasets without gathering all the data to a single site. The pivotal idea of distributed data clustering is to achieve a global clustering that is as good as the best centralized clustering algorithm with limited communication to collect the local models or local representatives into a single location, regardless of the crucial choice of any clustering techniques in local site[5]. Most of the applications that deal with time-critical distributed data are likely to benefit by paying careful attention to the distributed resources for computation, storage, and communication cost. Moreover, there exist a growing number of clustering applications, where the data have to be physically distributed, either owing to their huge volumes or privacy concern. Distributed data clustering is a promising approach for applications like weather analysis, financial data segmentation, distributed medical diagnosis, intrusion detection, data fusion in sensor networks, customer record segmentation, distributed gene expression clustering, click stream data analysis and census data analysis [24].

A common classification based on data distribution is, those which apply to homogeneously distributed or heterogeneously distributed data[39]. Homogeneous datasets contain the same set of attributes across distributed data sites. Heterogeneous data model supports different data sites with different schemata. For instance, a disease emergence detection may require collective information from a disease database, a demographic database and biological surveillance databases. According to the type of data communication, distributed clustering algorithms are classified into two categories: multiple communications round algorithms and centralized ensemble-based algorithms. The first group consists of methods requiring multiple rounds of message passing. These methods require a significant amount of synchronization, whereas the second group works asynchronously.

Related Works: The comprehensive survey of various distributed clustering solutions are available the literature[14,28,39]. This section reviews the recent research works on distributed clustering based on center based partitional clustering algorithms and hierarchical clustering algorithms Ji and Ling[24] derived the distributed clustering model through ensemble learning and proposed Distributed K-Means (DKM) algorithm This algorithm first performs local clustering using K-Means, and then sends all mean values to central site; finally global mean values of underlying global clustering are obtained by using K-Means again. Since, global set of centroids is computed using simple K-Means, the common issues of K-Means algorithm may lead to inconsistent global centroids.

Kashef and Kamel[31] presented the distributed Cooperative Partitional-Divisive Clustering (CPDC) algorithm which is based on intermediate cooperation between the distributed K-Means and the distributed bisecting K-Means clustering algorithms. This distributed approach exhibits better improvements in the global clustering results and maintains reasonable efficiency in a framework with larger number of nodes.

Hore et al.[18] proposed two heuristic algorithms, Bipartite Merger and Metis Merger, to partition the ensemble of centroids. The authors address the merging of multiple partitions formed from the widely used fuzzy K-Means and hard K-Means algorithms in a scalable framework. The Bipartite Merger algorithm works same as Distributed Combining(DC) algorithm[17] and restricts that number of clusters in the base clustering must be same for all partitions. On the other hand, the second algorithm, Metis Merger, does not use the assumption that the number of clusters in each base clustering is the same. But, it tends to partition ensembles of centroids, using the graph partitioning package METIS.

Karthikeyani et al.[25] proposed Improved Distributed Combining (IDC) algorithm to avoid the collision and improve the efficiency of the mapping process in DC algorithm[17], by introducing hungarian method of centroid mapping. It also supports variable number of clusters in each datasource. The same authors proposed modified distributed combining algorithm[26] to cluster disparate data sources having diverse, possibly overlapping set of features and also need not share objects. Both K-Means and Fuzzy C-Means algorithm is used for local clustering.

Karthikeyani and Thangavel[29] extended DKM algorithm by applying global normalization before performing the clustering on distributed datasets, without necessarily downloading all the data into a single site. The performance of proposed normalization based distributed K-Means clustering algorithm is compared against DKM algorithm and normalization based centralized K-Means clustering algorithm.

In[28], the performance of two distributed clustering algorithms, IDC[25] and DKM[24] are compared against traditional centralized clustering algorithm. Thangavel and Karthikeyani[42] proposed a novel ensemble based distributed K-Harmonic means algorithm, which is highly robust to centroid initialization as well as able to perform distributed clustering process in asynchronous manner.

In[13] Hammouda and Kamel introduced an approach for distributed data clustering, based on a structured P2P network architecture. This Hierarchically Distributed P2P Clustering (HP2PC) model involves a hierarchy of P2P neighbourhoods, in which the peers in each neighbourhood is responsible for building a clustering solution, using P2P communication, based on the data they have access to. As the hierarchy is moved up, clusters are merged from lower levels in the hierarchy. At the root of the hierarchy, one global clustering can be derived.

Datta et al.[5] proposed two approximate K-Means clustering algorithms that work by sampling a part of the network. The first is a locally synchronous in that each peer synchronizes at each iteration only with its topologically immediate neighbours in the network. The algorithm works in dynamic P2P networks and is observed empirically to produce accurate clustering results with respect to centralized K-Means clustering. Pakhira[29] presented a distributed K-Means algorithm which requires multiple rounds of message passing in terms centroid.

Intuitionistic fuzzy based distributed fuzzy clustering algorithm is proposed by Karthikeyani et al.[30] to incorporate intuitionistic fuzzy approach with distributed fuzzy clustering, in order to deal with uncertainty among the dispersed data objects and obtain effective and efficient fuzzy clusters in distributed environment.

D. Pedro, A. Forero, Alfonso Cano and Georgios B. Giannakis[39] developed two distributed algorithms for both deterministic and probabilistic approaches based on K-Means and EM respectively. The centralized problem of clustering spatially distributed data is solved by recasting it to a set of smaller local clustering problems with consensus constraints on the cluster parameter. The resulting iterative schemes do not exchange local data among nodes, and rely only on single-hop communications.

A distributed clustering algorithm for categorical data called Coercion is proposed by Bin Wang et al[3]. In this algorithm, the hierarchical clustering technique Squeezer is executed in different distributed server independendly and concurrently. All clustering results are transferred to one server, where similarity between every pair of cluster is measured using newly defined similarity measure called SimC. If the value of SimC is more than a threshold, these two clusters are combined and the new cluster structure is formed.

Hammouda and Kamel[14] derived two abstract models for distributed clustering in peer-to-peer environment with different goals. The first is the Locally optimized Distributed Clustering (LDC) model, which aims for better local clustering quality through collaboration. The second is the Globally optimized Distributed Clustering (GDC) model, whose objective is approximation of centralized clustering with the benefit of scalability through hierarchical distribution.

Distributed K-Modes Algorithm: The proposed algorithm is based on the assumption that data to be clustered are available at two or more nodes, which are referred to as local data sources. In addition, there is a node denoted as central site, where the results of clustering are attained and the additional computation for distributed clustering can be performed. The step by step procedure of proposed Ensemble based Distributed K-Modes (D-K-Md) algorithm for homogeneously distributed datasets is described in Figure-2.

First, data objects of local data sources are clustered independently, using K-Modes algorithm to obtain center matrix and cluster index for each data source. Then, all local centers are merged at central site and clustered using K-Modes algorithm to group similar centers and obtain global centers. The global centers are now transmitted to local data sources, where the hamming distance of each object from the global set of centers are computed and assigned to the nearest cluster center.

```
Algorithm. D-K-Md
Input :  Homogeneous p  datasets, each with d  categorical attributes and global K value
Output:  Global partitions of p datasets
Procedure:
Step-1: Cluster each local data source by K-Modes algorithm and obtain center matrix along with cluster index for each data source.
Step-2:  Merge cluster centers of local data sources into a single dataset named as 'center-dataset' at central site.
Step-3:  Cluster  'center-dataset' using K-Modes with global K value to obtain global centers
Step-4:  Update local cluster indices by assigning each object to nearest cluster center, after computing hamming distance between the object and global centers
```

Figure-2
Ensemble based Distributed K-Modes Clustering Algorithm

III. RESULTS AND DISCUSSION

In this section, empirical evidence is provided for D-K-Md algorithm that the high quality global cluster models is obtained with limited communication overhead and high level of privacy. The efficiency of D-K-Md is compared with existing distributed clustering algorithm, DKM along with CC, where all local datasets are merged and clustered using K-Modes algorithm. The existing DKM algorithm is not directly endurable for categorical datasets, because it uses the local clustering algorithm as K-Means and euclidean distance for the computation of local and global centroid. To execute this algorithm for categorical datasets, the values of each attribute are converted into number format by assigning sequential numbers for each category. For example, if an attribute 'color' contains three values such as 'blue', 'green', and 'red', they are mapped to three sequential numbers such as 1, 2, and 3.

Experimental Setup: The algorithms have been implemented and tested using five bench mark categorical datasets available in the UCI machine learning data repository[33]. The information about the datasets is shown in Table-1. For the experimental setup, the dataset is divided into different disjoint subsets and each subset is considered as a distributed data source. The experiment on each dataset runs 50 times and the average result is considered for analysis. All experiments are conducted with the assumption of having non-overlapping objects with same set of features in distributed datasets, for both uniform and non-uniform data distribution.

Evaluation Methodology: The performance of the proposed algorithm is measured in terms of four external validity measures[12,37] namely Rand index, Jaccard coefficient, F-Measure and Entropy. The external validity measures investigate the quality of clusters by comparing the results of clustering with the 'ground truth' (true class labels). The Rand index measures the agreement between true class labels and cluster results. The Jaccard coefficient measures the proportion of pairs that are in the same cluster as well as in the same class from those that are either in the same cluster or in the same class. The F-Measure measures the extent to which a cluster contains only objects of a particular class and all objects of the class. The Entropy is used to measure the degree to which each cluster consists of objects of a single class. In case of Rand index, Jaccard coefficient and F-Measure, value 1 indicates that the data clusters are exactly the same. Therefore increase in the values of these measures substantiates better performance, whereas, value 0 signifies that the data clusters are perfect for Entropy measure and hence the value of this measure has to be decreased to reach better quality clusters.

Uniform Type Data Distribution: In uniform type data distribution, the cardinality of each subset and number of clusters produced by each subset has been made equal. In first experiment, all datasets are divided into three subsets under uniform type data distribution and the algorithms are evaluated. The results of D-K-Md, in comparison with the results of DKM and CC in terms of Rand index, Jaccard coefficient, F-Measure and Entropy are shown in Table-2, Table-3, Table-4 and Table-5 respectively. From the Tables, it is observed that D-K-Md algorithm yields better results than DKM for all datasets, in terms of Jaccard coefficient and Entropy. With regard to Rand index and F-Measure, the performance of D-K-Md algorithm overshadows the performance of DKM algorithm except for tic-tac-toe dataset. It is observed that the performance is more commendable for balance scale and car evaluation datasets with D-K-Md algorithm. As the quality of clusters produced by D-K-Md is as good as CC with regard to all four validity measures for almost all datasets, the objective of distributed clustering is achieved effectively. The average performance of these three algorithms in terms of Rand index, Jaccard coefficient, F-Measure and Entropy are depicted in Figure-3, Figure-4, Figure-5 and Figure-6 respectively.

In next experiment, the chess dataset is divided into different number of subsets, in order to assess the scalability of the proposed distributed clustering algorithm. Table-7 shows the performance of D-K-Md on chess dataset, for different number of subsets. First row of the table (Chess -1S) depicts the results of Rand index, Jaccard Coefficient, F-Measure and Entropy, when all objects are kept in a single place, second row (Chess – 3S) potrays similar results, when the objects are divided into 3 subsets, and so on. Thus it is established that the proposed algorithm is consistent and is independent of the number of subsets. It is also noted that the performance in terms of F-Measure and Entropy increases, when the number of subsets are increased.

Table-1 Details of datasets

S. No.	Dataset	No. of Attributes	No. of Classes	No. of Instances
1	Balance Scale	4	3	**625**
2	Car Evaluation	6	4	**1728**
3	Chess	36	2	**3196**
4	Congressional Voting	16	2	**435**
5	**Tic-Tac-Toe Endgame**	**9**	**2**	**958**

Table-2 Performance analysis based on Rand index

S. No.	**Dataset**	**DKM**	**CC**	**D-K-Md**
1	Balance Scale	0.534	0.622	**0.631**
2	Car Evaluation	0.489	0.578	**0.568**
3	Chess	0.504	0.534	**0.543**
4	Congressional Voting	0.572	0.585	**0.575**
5	**Tic-Tac-Toe Endgame**	**0.569**	**0.539**	**0.541**

Table-3 Performance analysis based on Jaccard coefficient

S. No.	**Dataset**	**DKM**	**CC**	**D-K-Md**
1	Balance Scale	0.312	0.379	**0.362**
2	Car Evaluation	0.218	0.338	**0.347**
3	Chess	0.371	0.412	**0.409**
4	Congressional Voting Records	0.392	0.395	**0.399**
5	**Tic-Tac-Toe Endgame**	**0.322**	**0.336**	**0.333**

Table-4 Performance analysis based on F-Measure

S. No.	**Dataset**	**DKM**	**CC**	**D-K-Md**
1	Balance Scale	0.568	0.683	**0.691**
2	Car Evaluation	0.401	0.600	**0.621**
3	Chess	0.580	0.618	**0.628**
4	Congressiona l Voting	0.668	0.672	**0.678**
5	**Tic-Tac-Toe Endgame**	**0.618**	**0.575**	**0.569**

Table-5 Performance analysis based on Entropy

S. No.	**Dataset**	**DKM**	**CC**	**D-K-Md**
1	Balance Scale	0.634	0.445	**0.447**
2	Car Evaluation	0.693	0.684	**0.673**
3	Chess	0.688	0.665	**0.658**
4	Congressiona l Voting	0.616	0.613	**0.602**
5	**Tic-Tac-Toe Endgame**	**0.576**	**0.570**	**0.572**

Table-6 Comparative analysis for different no. of subsets

S. No.	**Dataset**	**Rand index**	**Jaccard coefficient**	**F-Measure**	**Entropy**
1	Chess - 1S	0.534	0.412	0.618	**0.665**
2	Chess - 3S	0.543	0.409	0.628	**0.658**
3	Chess - 5S	0.541	0.411	0.635	**0.655**
4	Chess - 7S	0.529	0.421	0.646	**0.639**
5	**Chess - 10S**	**0.536**	**0.415**	**0.658**	**0.621**

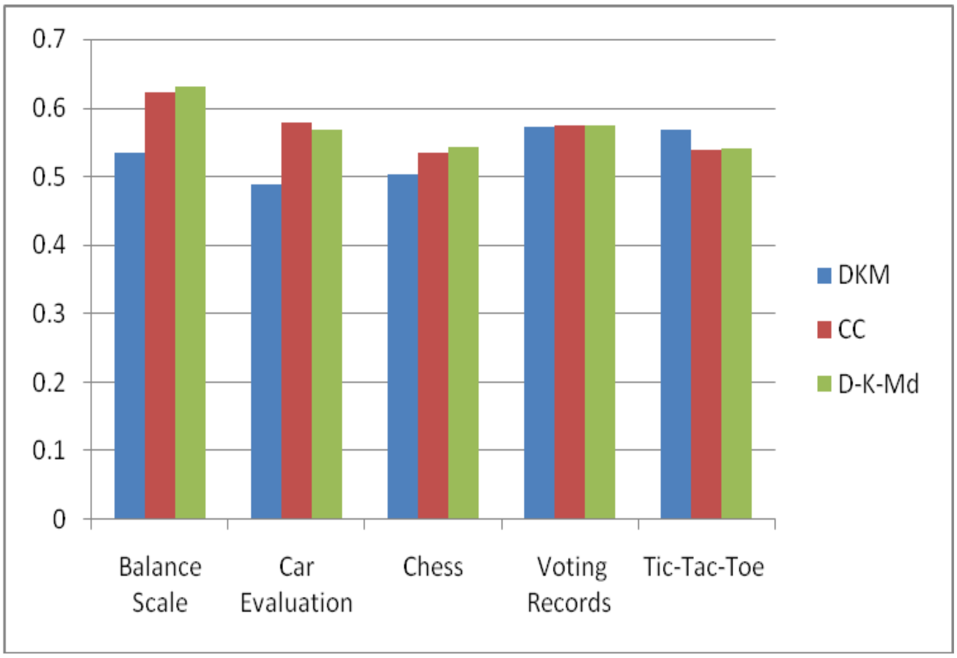

Figure-3 Comparative analysis based on Rand index

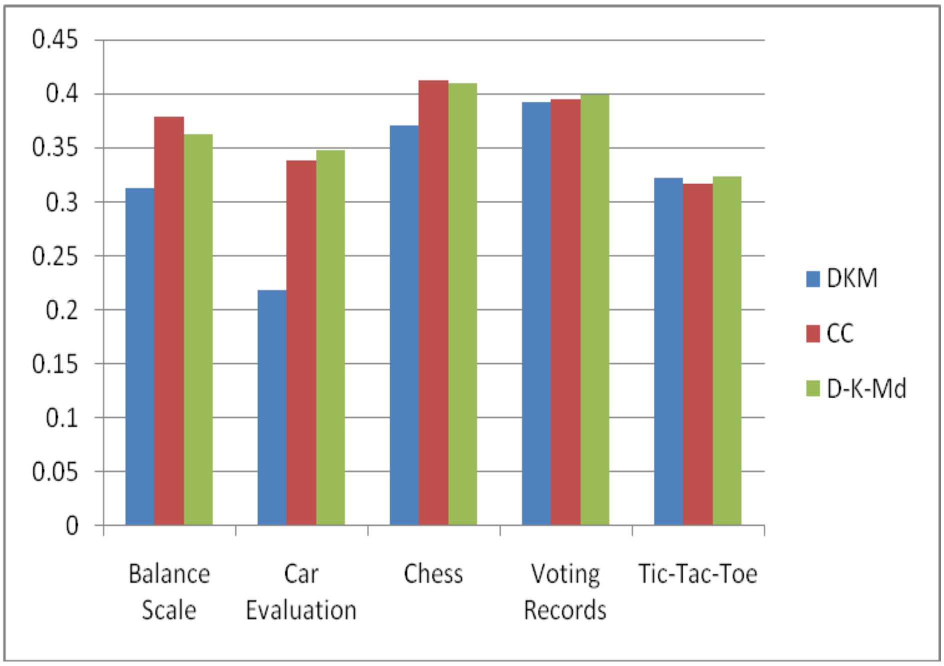

Figure-4 Comparative analysis based on Jaccard coefficient

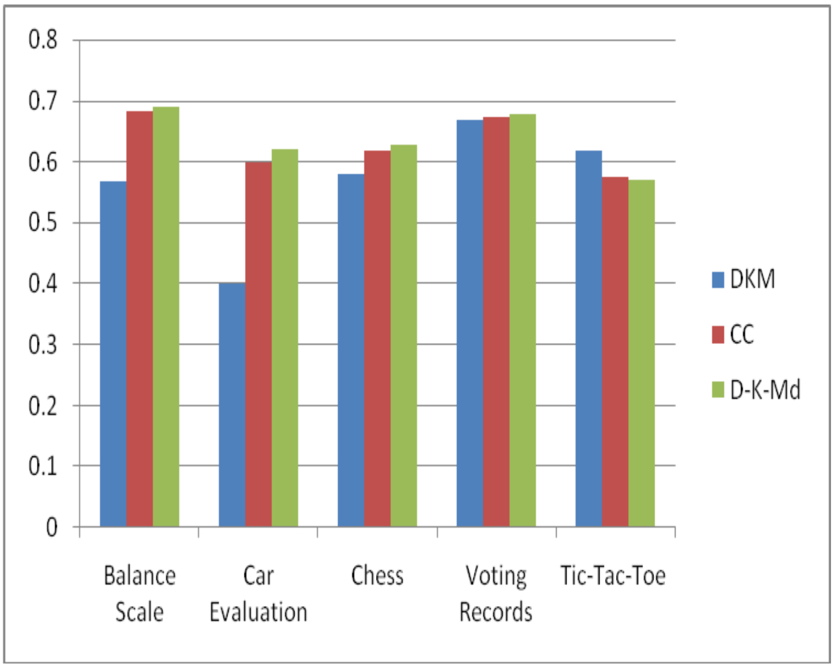

Figure-5 Comparative analysis based on F-Measure

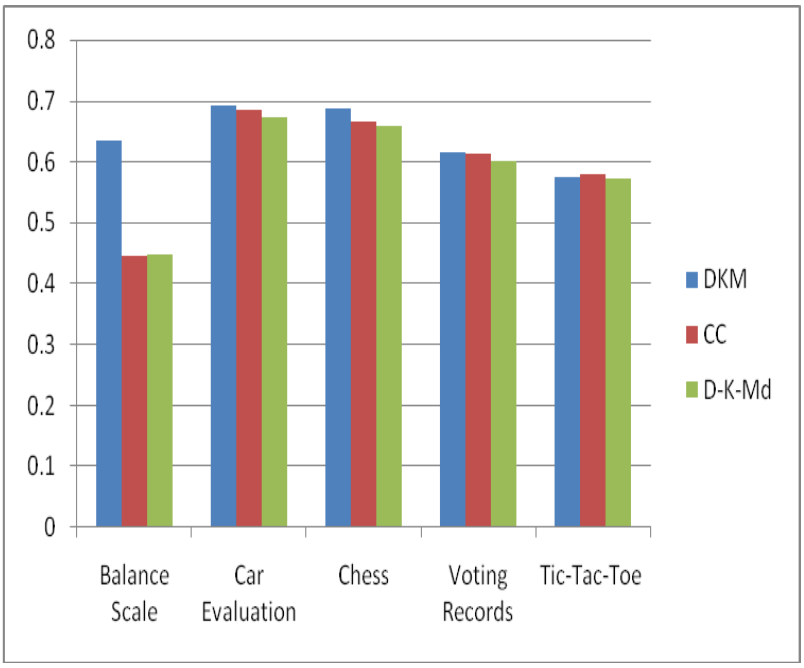

Figure-6 Comparative analysis based on Entropy

Non-Uniform Type Data Distribution: Non-uniform type data distribution is based on the assumption that individual data sources are having varying numbers and varied types of clusters. In some applications, one or more classes may be missing in the data sources. Sometimes, data sources may have entirely different type of clusters. For instance, data marts of individual stores of a retail company may deal with different types of customers or products, according to the demands that prevail there. Similar situation is also efficiently handled in the proposed algorithm simply by providing required number of global clusters, independent of the number of local clusters.

The results of car evaluation dataset in terms of Rand index, Jaccard coeffient, F-Measure and Entropy obtained for this assumption is represented in Table-7, along with the number of objects in three data sources namely S1, S2 and S3. The numbers provided in braces indicate the cluster labels in the corresponding data sources.

Though D-K-Md provides equal performance as CC with respect to all four validity measures, it outperforms CC in terms of communication overhead, space complexity and privacy maintenance. In CC, all objects should be transferred to a central place and K-Modes algorithm is to be executed to find global cluster indices. In real application scenario, it needs huge communication cost, since data sources may contain large number of high dimensional data objects. In distributed approach, only centers of local clusters and global clusters are to be transmitted between data sources. The cluster centers are insensitive to a number of objects in each data source and the size of cluster centers is definitely much less than the size of data objects. Moreover, centralized clustering needs more memory space at one place depending on the size of objects accumulated for clustering. It is important to preserve privacy on individual objects for most of the distributed applications like financial, banking and medical applications. Since cluster centers represent only prototype, the proposed method of distributed clustering enables privacy preserving data mining framework.

IV. CONCLUSION

Most of the existing distributed partitional clustering algorithms have been developed for grouping numerical datasets. The distributed K-Modes clustering algorithm is proposed based on cluster ensemble to cluster categorical datasets in distributed environment. The experimental results on five benchmark datasets demonstrate the suitability of the proposed algorithm. Ongoing research focuses on optimizing the proposed algorithm by modifying the similarity measure, applying suitable cluster initialization procedure and automatic selection of K value to make it suitable for real life distributed scenario.

Table-7 Results of non-uniformly distributed Car Evaluation dataset (No. of global clusters – 4)

S. No.	**Size of datasets (cluster labels)**			**Rand index**	**Jaccard coefficient**	**F-Measure**	**Entropy**
	S1	S2	S3				
1	340 (1, 2, 3)	809 (1, 2, 3,	679 (1, 2, 4)	0.572	0.347	0.647	**0.682**
2	784 (1, 2)	500 (1, 2)	444 (1, 3, 4)	0.582	0.392	0.664	**0.621**
3	669 (1, 3)	375 (1, 2, 4)	484 (1, 2)	0.586	0.365	0.678	**0.645**
4	629 (1, 2, 3)	665 (1, 4)	535 (1, 2)	0.574	0.357	0.638	**0.624**
5	**744 (1, 2, 3,**	**484 (1, 2)**	**500 (1, 2)**	**0.594**	**0.421**	**0.682**	**0.618**

REFERENCES

[1]. Anil Chaturvedi, Paul E., Green J., Douglas Caroll., K-Modes Clustering, Journal of Classification , **18(1),** 35-55 **(2001)**

[2]. Anil K. Jain, Data Clustering: 50 Years Beyond K-Means, Pattern Recognition Letters, **(2009)**

[3]. Bin Wang , Yang Zhou , Xinhong Hei, Coercion: A Distributed Clustering Algorithm for Categorical Data, 9th International Conference on Computational Intelligence and Security, 683 – 687 **(2013)**

[4]. Daniel Barbara, Julia Couto, Yi Li, COOLCAT: an entropy-based algorithm for categorical clustering, Proceedings of the eleventh international conference on Information and knowledge management, **(2002)**

[5]. Datta S., Giannella C.R., Kargupta H., Approximate distributed K-Means clustering over a peer-to-peer network, IEEE Transactions on Knowledge and Data Engineering, **21(10)**, 1372-1388 **(2009)**

[6]. Fan-rong Meng, Bin Liu, Chu-jiao, Wang, Privacy preserving clustering over distributed data, 3rd International Conference on Advanced Computer Theory and Engineering , **5**, 544 - 548 **(2010)**

[7]. Fuyuan Cao, Jiye Liang, Deyu Li, Xingwang Zhao, A weighting K-Modes algorithm for subspace clustering of categorical data, Neurocomputing, **108**, 23–30 **(2013)**

[8]. Ganti V., Gehrke J. and Ramakrishnan R., CATUS - Clustering categorical data using summaries, Proc. Int. Conf. Knowledge Discovery and Data Mining, San Diego, USA, 73–83 **(1999)**

[9]. Ghosh J., Merugu S., Distributed Clustering with Limited Knowledge Sharing, Proceedings of the 5th International Conference on Advances in Pattern Recognition, 48-53 **(2003)**

[10]. Gibson D., Kleinberg J., and Raghavan P., Clustering categoricaldata: An approach based on dynamic systems, Proc. 24-th Int. Conf. Very Large Databases, New York, 311–323 (**1998)**

[11]. Guha S., Rastogi R. and Shim K., ROCK: A robust clustering algorithm for categorical attributes, Information Systems, **25(5),** 345–366 (2000)

[12]. Halkidi M., Batistakis Y., Vazirgiannis M., Cluster Validity Methods: Part II. ACM SIGMOD Record, **31(3)**, 19-27 **(2002)**

[13]. Hammouda K.M. and Kamel M.S., Hierarchically distributed peer-to-peer document clustering and cluster summarization, IEEE Transactions on Knowledge and Data Engineering, **21(5),** 681-698 **(2009)**

[14]. Hammouda K.M and Kamel M.S., Models of distributed data clustering in peer-to-peer environments, Knowledge and Information Systems, **38 (2)**, 303-329 **(2014)**

[15]. Han J. and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, **(2006)**

[16]. He Z., Deng S., Xu X., Improving K-Modes algorithm considering frequencies of attribute values in mode, International Conference on Computational Intelligence and Security, LNAI 3801, 157-162 **(2005)**

[17]. Hore P. and Lawrence O. Hall, Scalable Clustering: A Distributed Approach, IEEE International Conference on Fuzzy Systems, 25-29 **(2004)**

[18]. Hore P., Lawrence O. Hall, Dimitry B. Goldgofz, A Scalable Framework for Cluster Ensembles, Pattern Recognition, **42(3),** 676-688 **(2009)**

[19]. Huang Z., A fast clustering algorithm to cluster very large categorical data sets in data mining, Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Know ledge Discovery, Tucson, Arizona, USA,1-8 **(1997a)**

[20]. Huang Z., Clustering large data sets with mixed numeric and categorical Values, Proceedings of the FirstAsia Confernce on Knowledge Discovery and Data Mining, 21-34 **(1997b)**

[21]. Huang Z., Extensions to the K-Means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery , **2(3)**, 283-304 **(1998)**

[22]. Jain A., Murthy K., Flynn M.N., Data Clustering: A Review, ACM Computing Surveys, **31(3)**, 265-323 **(1999)**

[23]. Januzaj E., Kriegel Hans P., Pfeifle M., DBDC: Density Based Distributed Clustering, E. Bertino, S. Christodoulakis, D. Plexousakis (eds.), Advances in Databases Technology – EDBT, 2992, 529-530 **(2004)**

[24]. Ji Genlin and Ling Xiaohan, Ensemble Learning based Distributed Clustering, T. Washio et al., (eds.), Emerging Technology and Knowledge Discovery and Data Mining, 4819, 312-321 **(2007).**

[25]. Karthikeyani Visalakshi N., Thangavel K., Alagambigai P., Ensemble Approach to Distributed Clustering, Natarajan et al.(eds.), Mathematical and Computational Model, Narosa Publishing House, New Delhi, 252-261 **(2007)**

[26]. Karthikeyani Visalakshi N., Thangavel K., Alagambigai P., Distributed clustering for data sources with diverse schema, Third International Conference on Convergence and Hybrid Information Technology, IEEE Computer Society, 1058-1063 **(2008).**

[27]. Karthikeyani Visalakshi N. and Thangavel K., Ensemble based Distributed Soft Clusteing, International Conference on Computing , Communication and Networking, IEEE, **(2008)**

[28]. Karthikeyani Visalakshi N. and Thangavel K., Distributed data clustering: a comparative analysis, Ajith Abraham, Aboul-Ella Hassanien, Andr´e Ponce de Leon F. de Carvalho, and Václav Snášel (Eds.), Foundations of Computational Intelligence, Studies in Computational Intelligence, Springer Berlin / Heidelberg, 371-398 **(2009)**

[29]. Karthikeyani Visalakshi N., and Thangavel K., Impact of Normalization in Distributed K-Means Clustering, International Conference on Soft Computing, **4(4**), 165-172 **(2009)**

[30]. Karthikeyani Visalakshi N. and Thangavel K., An intuitionistic fuzzy approach to distributed fuzzy clustering, International Journal of Computer Theory and Engineering, **2(2),** 1793-8201 **(2010)**

[31]. Kashef R. and Kamel M., Distributed Cooperative Hard-Fuzzy Document Clustering, Proceedings of the Annual Scientific Conference of the LORNET Reseach Network, Montreal, 8-10 **(2006)**

[32]. Liang Bai and Jiye Liang, The K-Modes type clustering plus between cluster information for categorical data, Neurocomputing,**133(10),** 111–121 **(2014)**

[33]. Merz C. J. and Murphy P. M., UCI Repository of Machine Learning Databases. Irvine, University of California, http://www.ics.uci.eedu/~mlearn/ **(1998)**

[34]. Ng M.K., Li M.J., Huang J.Z., He Z., On the impact of dissimilarity measure in K-Modes clustering algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence , **29(3)**, 503–507 **(2007)**

[35]. Omar S. Soliman, Doaa A. Saleh, and Samaa Rashwan, A Bio Inspired Fuzzy K-Modes Clustring Algorithm, ICONIP, Part III, LNCS 7665, 663–669 **(2012)**

[36]. Pakhira M.K., Clustering large databases in distributed environment, IEEE International Advanced Computing Conference, 351-358 **(2009)**

[37]. Pang-Ning T., Steinbach M., Kumar V., Cluster Analysis: Basic Concepts and Algorithms, Introduction to Data Mining, 491-501, Pearson Addison Wesley, Boston, **(2006)**.

[38]. Park B. and Kargupta H., Distributed Data Mining, Nong Ye, (ed.), The Hand Book of Data Mining, Lawrence Erlabum Associates, Publishers, Mahwah, New Jersey **(2003)**

[39]. Pedro A. Forero, Alfonso Cano, Georgios B. Giannakis, Distributed Clustering Using Wireless Sensor Networks, IEEE Journal of Signal processing, **5(4),** 707-721 **(2011)**

[40]. Sanghamitra B., Giannella C., Maulik U., Clustering Distributed Data Streams in Peer-to-Peer Environments, Information Science, **176 (4),** 1952-1985 **(2006)**

[41]. Strehl A., Ghosh J., Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions, Journal of Machine Learning Research, 3, 583-617 **(2002)**

[42]. Thangavel K. and Karthikeyani Visalakshi N., Ensemble based Distributed K-Harmonic Means Clustering , International Journal of Recent Trends in Engineering, **2(1)**, 125-129 **(2009)**

[43]. Xu R., Wunsch II D., Survey of Clustering Algorithms, IEEE Transaction on Neural Networks, **16 (3)**, 645-678 **(2005)**

[44]. Zhexue Huang and Michael K. Ng, A Fuzzy K-Modes Algorithm for Clustering Categorical Data, IEEE Transactions on Fuzzy Systems, **7(4),** 446-452 **(1999)**

[45]. Zhexue Huang and Michael K.Ng, A note on K-Modes Clustering, Journal of Classification, Springer-Verlag New York, Inc. Secaucus, NJ, USA, **20(2)** , 257 – 261 **(2003)**