Ensemble based Distributed K-Harmonic Means Clustering

K. Thangavel1 and N. Karthickeyani Visalakshi2
1 Department of Computer Science, Periyar University, Salem, Tamil Nadu, India
Email: drktvelu@yahoo.com
2 Department of Computer Science, Vellalar College for Women, Erode, Tamil Nadu, India
Email: karthicitru@yahoo.co.in

Abstract—Due to the explosion in the number of autonomous data sources, there is a growing need for effective approaches for distributed knowledge discovery and data mining. The distributed clustering algorithm is used to cluster the distributed datasets without necessarily downloading all the data to a single site. K-Means is used as a popular clustering method due to its simplicity and high speed in clustering large datasets. The dependency of the K-Means performance on the initialization of centroids is a major problem. Similarly, distributed clustering algorithm based on K-Means is also sensitive to centroid initialization. It is demonstrated that K-Harmonic Means is essentially insensitive to centroid initialization. In this paper, a novel ensemble based distributed clustering algorithm using K-Harmonic Means is proposed. The simulated experiments described in this paper confirm robust performance of the proposed algorithm.

Index Terms—Distributed clustering, Global centroid, K-Harmonic Means, K-Means, Local Centroid

I. INTRODUCTION

Clustering is a process of grouping a set of data objects into clusters based on the information found in the data objects, in such a way that the objects in the same cluster are similar where as objects in different clusters are different. The clustering plays an important role in various data analysis fields including statistics, pattern recognition, machine learning, data mining, information retrieval and bioinformatics. Generally, clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. An excellent survey of clustering techniques can be found in [5].

Today’s large-scale datasets are usually logically and physically distributed, requiring a distributed approach to clustering. Huge amounts of data are stored in autonomous, geographically distributed data sources over networks with limited bandwidth and large number of computational resources [13]. Traditional centralized clustering methods require all data to be located at the place, where they are analyzed. Due to technical, economical or security reasons, it is not always possible to transmit all data from different local sites to single location and then perform global clustering. It is obvious that alternate distributed clustering algorithms [16] reduce the communication overhead, central storage requirements and computation times by exchanging few data and avoiding synchronization as much as possible.

Some of the distributed clustering algorithms available in the literature [9, 14] aim to generate global clusters among distributed datasets, based on the centroid based partitional clustering algorithms like K-Means (KM) and Expected Maximization (EM). But, these algorithms have the intrinsic problem of depending heavily on the initialization of cluster centroids for achieving good performance. Zhang et al [16] introduced K-Harmonic Means (KHM) to remedy the inconsistencies of clustering results due to random initializations and it is reported that KHM significantly improves the quality of clustering results comparing with KM and EM, in certain cases [2].

In recent years, limited attention has been paid in applying KHM to obtain consistent clusters in centralized environment [2]. Forman and Zhang [16] introduced distributed K-Harmonic Means clustering algorithm to attain distributed data clustering. However, their approach requires multiple rounds of message passing with full synchronization among distributed datasets and somehow increases communication overhead. In this paper, a novel Ensemble based Distributed K-Harmonic Means (EDKHM) algorithm is proposed, which is highly robust to centroid initialization as well as able to perform distributed clustering process in asynchronous manner. The performance of the proposed algorithm is compared with the existing distributed clustering algorithms, Distributed K-Means (DKM) [7] and Centralized Clustering (CC) based on KHM. The rest of this paper is organized as follows: Section 2 provides necessary background discussions on K-Harmonic Means clustering and distributed clustering. Section 3 confers the related works. Section 4 describes the proposed algorithm EDKHM. Section 5 illustrates the experimental analysis performed with benchmark datasets. Finally, Section 6 concludes the paper.

II. BACKGROUND

A. K-Harmonic Means Clustering

K-Harmonic Means [15] is a centroid based partitional clustering algorithm that assigns soft membership to each data object. The formal description of KHM is given in Figure 1. The KHM uses the harmonic averages of the distances from each data object to the centroids as components to its performance function. Unlike KM, KHM assigns dynamic weights to each data objects based on a harmonic average, in every iteration. The harmonic average will assign a large weight to a data object that is not close to any centroids and a small weight to data object that is close to one or more...
The key idea of distributed clustering is to achieve a global clustering that is as good as the best centralized clustering algorithm with limited communication required to collect the local models or local representatives into a single location, regardless of the crucial choice of any clustering technique in local site. Distributed clustering algorithms can be classified along two independent dimensions such as classification based on data distribution and data communication [14].

A common classification based on data distribution in the literature [14] is those, which apply to homogeneously distributed or heterogeneously distributed data. Homogeneous datasets contain the same set of attributes across distributed data sites. Heterogeneous data model supports different data sites with different schemata. For example, a disease emergence detection problem 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.

III. RELATED WORKS

There are various distributed clustering solutions proposed in the literature and their comprehensive survey can be obtained from [10, 14]. This section reviews the recent research works on distributed clustering based on KM and EM.

The Peer-2-Peer (P2P) K-Means algorithm is proposed in [14] for distributed clustering of data streams in a peer-to-peer sensor network environment. In the P2P K-Means algorithm, computation is performed locally, and communication of the local data models, represented by the corresponding centroids and the cluster counts is restricted only within a limited neighborhood.

Jin R. et al. [8] presented distributed version of Fast and Exact K-Means (FEKM) algorithm, which collected sample data from each data source, and communicated it to the central node. The main data structure of FEKM, the cluster abstract table is computed and sent to all data sources to get global clusters. The boundary point and other sufficient statistics are communicated to central node.

Cormode G. et al. [1] have introduced the problem of continuous, distributed clustering, and given a selection of algorithms, based on the paradigms of local vs. global computations, and furthest point or parallel guessing clustering.

In [17], Zhou A. et al. proposed an EM-based (Expectation Maximization) framework to cluster the distributed data streams effectively. In the presence of noisy or incomplete data records, their algorithms learn the distribution of underlying data streams by maximizing the likelihood of the data clusters. A test-and-cluster strategy is proposed to reduce the average processing cost, which is especially effective for online clustering over large data streams.

B. Distributed Clustering

Distributed clustering assumes that the objects to be clustered reside on different sites. This process is carried out in two different levels: local level and global level. In local level, all sites carry out clustering process independently from each other. After having completed the clustering, a local model such as cluster centroids is determined, which should reflect an optimum trade-off between complexity and accuracy. Next, 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 [9].

The key idea of distributed clustering is to achieve a global clustering that is as good as the best centralized clustering algorithm with limited communication required to collect the local models or local representatives into a single location, regardless of the crucial choice of any clustering technique in local site. Distributed clustering algorithms can be classified along two independent dimensions such as classification based on data distribution and data communication [14].

A common classification based on data distribution in the literature [14] is those, which apply to homogeneously distributed or heterogeneously distributed data. Homogeneous datasets contain the same set of attributes across distributed data sites. Heterogeneous data model supports different data sites with different schemata. For example, a disease emergence detection problem 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.

III. RELATED WORKS

There are various distributed clustering solutions proposed in the literature and their comprehensive survey can be obtained from [10, 14]. This section reviews the recent research works on distributed clustering based on KM and EM.

The Peer-2-Peer (P2P) K-Means algorithm is proposed in [14] for distributed clustering of data streams in a peer-to-peer sensor network environment. In the P2P K-Means algorithm, computation is performed locally, and communication of the local data models, represented by the corresponding centroids and the cluster counts is restricted only within a limited neighborhood.

Jin R. et al. [8] presented distributed version of Fast and Exact K-Means (FEKM) algorithm, which collected sample data from each data source, and communicated it to the central node. The main data structure of FEKM, the cluster abstract table is computed and sent to all data sources to get global clusters. The boundary point and other sufficient statistics are communicated to central node.

Cormode G. et al. [1] have introduced the problem of continuous, distributed clustering, and given a selection of algorithms, based on the paradigms of local vs. global computations, and furthest point or parallel guessing clustering.

In [17], Zhou A. et al. proposed an EM-based (Expectation Maximization) framework to cluster the distributed data streams effectively. In the presence of noisy or incomplete data records, their algorithms learn the distribution of underlying data streams by maximizing the likelihood of the data clusters. A test-and-cluster strategy is proposed to reduce the average processing cost, which is especially effective for online clustering over large data streams.
G. Ji and X. Ling [7] 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 the central site; finally, the global mean values of underlying global clustering are obtained by using K-Means again. Jongil Jeong et al. [6] proposed similar type of distributed clustering scenario for clustering huge quantities of biological data. NDKM [10] is an enhanced version of DKM using global normalization and optimal initial cluster centroids.

In [3] Prodip Hore and L. Hall proposed Distributed Combining Algorithm (DCA) to cluster large scale datasets without clustering all the data at a time. Data is randomly divided into almost equal size disjoint subsets. Each subset is clustered using the hard K-Means or fuzzy C-Means algorithm. The centroids of subsets form an ensemble. A centroid correspondence algorithm transitively solves the correspondence problem among the ensemble of centroids. When the number of clusters in each subset is large, the complexity increases in centroid mapping due to collision. Moreover, when the number of clusters in each dataset is different, this type of centroid mapping is found not suitable. Prodip Hore extended this algorithm in [4], to avoid collision and filter bad centroids, but limited to the same number of clusters in each data source.

The Improved Distributed Combining Algorithm (IDCA) [9] is a refined version of Distributed Combining Algorithm [3], designed for distributed hard clustering. The process of centroid mapping is performed effectively, with the support of Hungarian method of unbalanced assignment problem and Euclidean relation, when each dataset produces different number of clusters.


R. Kashef and M. S. Kamel [11] proposed a Distributed Cooperative Hard-Fuzzy Clustering (DCHFC) model for document clustering. This model is based on the intermediate cooperation between the hard distributed K-Means and fuzzy distributed C-Means to enhance the performance of the K-Means reduce the computational time taken by the fuzzy algorithm and produce a better global solution.

IV. PROPOSED ALGORITHM

The proposed algorithm assumes that data to be clustered is 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 desired. It is also assumed that additional computation for distributed clustering can be performed at the central site. The step by step procedure of proposed Ensemble based Distributed K-Harmonic Means algorithm for homogeneously distributed datasets is described in Figure 2.

First, minimum and maximum values of each feature vector are extracted from all local data sources and transmitted to central site, where global minimum and maximum values are identified. These two values are transmitted to local data sources to perform global normalization using (1). Next, normalized objects of local data sources are clustered independently, using K-Harmonic Means algorithm to obtain centroids matrix and cluster index for each data source. Then, all local centroids are merged at central site and clustered using K-Harmonic Means algorithm to group similar centroids and obtain global centroids. The global centroids is now transmitted to local data sources, where the Euclidean distance of each object from the global set of centroids are computed and assigned to the nearest cluster centroid.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Ensemble based Distributed K-Harmonic Means Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Homogeneous $m$ datasets, each with $d$ dimensions and global $k$ value</td>
</tr>
<tr>
<td>Output</td>
<td>Global partitions of $m$ datasets</td>
</tr>
<tr>
<td>Procedure</td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td>Find maximum and minimum values of each feature from each local data source and transmit them into central site.</td>
</tr>
<tr>
<td>Step 2</td>
<td>Compute global maximum and minimum values for all features at central place and transmit them to local data sources.</td>
</tr>
<tr>
<td>Step 3</td>
<td>Normalize real scalar values of local data sources with global maximum and minimum values using (1)</td>
</tr>
<tr>
<td>Step 4</td>
<td>Cluster each local data source by K-Harmonic Means algorithm and obtain centroids matrix along with cluster index for each data source.</td>
</tr>
<tr>
<td>Step 5</td>
<td>Merge cluster centroids of local data source into a single dataset named as ‘centroids dataset’ at central site.</td>
</tr>
<tr>
<td>Step 6</td>
<td>Cluster ‘centroids dataset’ using K-Harmonic Means with global $k$ value to obtain global centroids</td>
</tr>
<tr>
<td>Step 7</td>
<td>Update local cluster indices by assigning each object to nearest cluster centroid, after computing Euclidean distance between the object and global centroids</td>
</tr>
</tbody>
</table>

V. EXPERIMENTAL ANALYSIS

In this section, empirical evidence is provided for EDKHM algorithm, that the robust global cluster models is obtained, with limited communication overhead and high level of privacy. The efficiency of EDKHM is compared against existing distributed clustering algorithm, DKM along with CC, where all local datasets are merged, normalized and clustered using K-Harmonic Means algorithm. All experiments are conducted with the assumption of having non-overlapping objects with same set of features in distributed datasets, for uniform type of data distribution. In uniform type of data distribution, the cardinality of each subset has been kept as almost equal and number of clusters produced by each subset has also been kept as equal.
The algorithms have been implemented in P2P environment and tested with six benchmark numeric datasets available in the UCI machine learning data repository [12]. The information about the datasets is shown in Table 1. For the purpose of experimental setup, the dataset is divided into three disjoint subsets and each subset is considered as distributed data source. The experiment on each dataset runs 50 times to analyze the robustness of the proposed algorithm. The value of parameter $p$ is set as 2.25, while implementing K-Harmonic Means clustering, irrespective of the characteristics of the datasets.

The performance of the proposed algorithm is measured in terms of three external validity measures [10] namely Rand index, F-Measure and Entropy. The external validity measures test the quality of clusters by comparing the results of clustering with the ‘ground truth’ (true class labels). In case of Rand index and F-Measure, the value 1 indicates that the data clusters are exactly same and so the increase in the values of these measures proves the better performance. But, the value 0 signifies that the data clusters are perfect for Entropy measure and so the value of this measure is to be decreased to reach better quality clusters.

The results of EDKHM, in comparison with the results of DKM, in terms of Rand index, F-Measure and Entropy are shown in Table 2, Table 3 and Table 4 respectively. From the tables it is observed that EDKHM algorithm yields consistent and improved results than DKM, almost for all datasets. It is noted that the results of EDKHM is highly appreciable for Breast Cancer dataset, in terms of all three measures. But, the results of both DKM and EDKHM are consistent and same for Australian dataset.

The average results of EDKHM and CC is compared in Table 5. From this table, it is identified that the performance of proposed distributed algorithm is almost same as centralized clustering. Though EDKHM provides equal performance as CC based on KHM, in terms of Rand index, F-Measure and Entropy, it outperforms CC in terms of communication overhead, space complexity and privacy maintenance. In CC, all objects are to be transferred to central place and K-Harmonic Means clustering algorithm is to be executed to find global cluster index. In real application scenario, it needs huge communication cost, since many data sources may contain large number of high dimensional data objects. In distributed approach, only centroids of local clusters and global centroid is to be transmitted between data sources. The centroids are insensitive to a number of objects in each data source and the size of cluster centroid is definitely much less than the size of data objects or even size of label vector. Moreover, centralized clustering needs much memory space at one place, according to the size of objects accumulated for clustering. It remains important to preserve privacy on individual objects for most of the distributed nature of application scenario like financial, banking and medical applications. Since centroids of clusters represent only prototype, the proposed distributed clustering algorithm enables privacy preserving data mining framework.
A novel method of distributed clustering using K-Harmonic Means algorithm is proposed to produce global clusters in asynchronous manner. Comprehensive experiments on six benchmark numerical datasets have been conducted to analyze the benefit of K-Harmonic Means algorithm in distributed clustering. It can be concluded that the proposed algorithm will help in producing superior quality clusters.

ACKNOWLEDGMENT

UGC Grant No. F-NO.34-105/2008 (SR) is gratefully acknowledged.

VI. CONCLUSION

A novel method of distributed clustering using K-Harmonic Means algorithm is proposed to produce global clusters in asynchronous manner. Comprehensive experiments on six benchmark numerical datasets have been conducted to analyze the benefit of K-Harmonic Means algorithm in distributed clustering. It can be concluded that the proposed algorithm leads to obtain robust clusters than the existing distributed clustering algorithm. At the same time, it is also proved that the performance of the proposed algorithm is almost same as the performance of centralized clustering. The proposed algorithm can also be implemented effectively for non-uniform type of data distribution. In future, applying optimization algorithm for tuning of parameter $p$ in K-Harmonic Means algorithm will help in producing superior quality clusters.

REFERENCES