International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue14
November 2017

Distributed K-Modes Clustering in P2P Networks

Golla Saidulu & Masku Naveen Kumar
[1]Assistant Professor, Dept of CSE, CMR College of Engineering And Technology
[2]Assistant Professor, Dept of CSE, SCIENT Institute of Technology

ABSTRACT:*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.*

KEYWORDS-

I. INTRODUCTION

As computing and conversation over wired and wireless networks superior, many pervasivedispensed computing environments such as net, intranets, LANs, ad-hoc wi-finetworks and P2P networks have emerged. These environments often cope with specialdistributed assets of voluminous information, multiple computing nodes, and dispensed personcommunity. Apposite usage of those distributed assets must be assured in miningsuch environments. Also, local data resources can be of restrained availability due to privateness and asan end result data sets at different websites ought to be processed in a dispensed style without amassingthe whole thing to ansingle important site. Traditional data mining technique is to download the relevantinformation to a centralized region after which perform the records mining operations. Many of theallotted, privateness-touchy records mining packages cannot make use of this centralizedtechnique.Distributed Data Mining (DDM) explores strategies of how to apply records mining in a noncentralized way. DDM requires an architecture that is definitely various from the one used incentralized approach. In a distributed environment, the structure has to facilitate to paycautious attention to distributed sources of statistics, computing, and verbal exchange and human-pc interaction [1]. P2P networks are gaining developing popularity in many dispensedprograms which includes record-sharing, internet caching, community garage, searching and indexing ofrelevant documents and P2P community-danger evaluation. It permits a collection of nodes (peers) topercentage computer sources in a decentralized manner. Collectively the friends already save amassive quantity of extensively various records accrued from specific resources.

Distributed Data Mining is a framework to mine distributed data which operates on anarchitecture that is totally different from centralized approach. It cares the distributed sourcesof data, computing and communication.DDM architecture includes multiple sites each havingindependent computing power and storage capability. Each site performs local computation onits own and finally either a central site communicates with each distributed site to compute theglobal models or a peer-to-peer architecture is used. In the latter case, individual nodesperform most of the tasks by communicating with neighboring nodes by message passing overan asynchronous network.

The architecture for DDM is as shown in figure 1. From the figure it is clear that, in a distributedsetting several local models are generated on different nodes and finally aggregated to form aglobal model which represents the mining result of the entire dataset.

Distributed data mining comprises two types of architectures - centrally coordinate architectureand P2P [2]. In the first architecture, the entire data mining work is split into multiple workersand a central process

coordinates the workers. However this approach suffers from theproblem of single point of failure. Privacy issues and communications concerns are alsoassociated with this. The second architecture, P2P data mining overcomes these problemswhere a large number of nodes are connected in an ad-hoc way. As communication is only withits neighbors, overhead is low and elegant handling of failure of single nodes.

Distributed Clustering

Database records are partitioned into clusters via clustering where elements of a cluster share aset of common properties that distinguish them from other clusters. The goal of clustering is tominimize inter-cluster similarity and maximize intra-cluster similarity. A clustering algorithminvolves 3 steps in general. First step is to compute local models using local clusteringalgorithms. Next aggregate local models by a central node and finally either compute the globalmodel or aggregated models are sent back to all the nodes to produce locally optimized clusters[2]. Some of the distributed clustering algorithms are K-means, model based, density grid andhierarchical.

K-means: Initially the k cluster centers among all the random points are randomly chosen.These k cluster centers are then sent to every local representative and local K-means clusteringis performed. Each local machine then gathers the statistics about membership within its ownclusters. Each of the statistics is then transmitted to a central controller to aggregate themodels. As we are transferring the statistics rather than the entire data, data privacy ismaintained. But this statistics need to be sent over and over again until convergence. Thisalgorithm does not scale well and not assured to be a very quick process [4].

Model Based: This algorithm uses expectation maximization clustering [5] on the local levelwhich is similar to K-means, except that decision on final clustering based on additionalfunctions like Gaussian function. It is described in [6].Initially the local system processes its ownindividual pieces, by local EM clustering and each cluster is modeled as a sum of Gaussianfunctions. These functions are then transferred to a central coordinator which combines thefunctions to give global information about the probability density of the global picture. Thisinformation is then sent to each local source and they can make use of the new information,reevaluate the data if needed. This algorithm employs good measures for privacy and accuracy.This method suffers from a standard problem such that two large clusters with a small denselyconnected component can end up being in the same cluster even if they should not be so [4].

II. RELATED WORK

However, in many real applications today, like sensor monitoring and location-based services [10], data mostly containsinherent uncertainty due to the random nature of the data generation, measurement inaccuracy, sampling discrepancy, datastaling, and other errors. Generally, with uncertainty, the dataobject is no longer a single point in space but is representedby a probability density function (pdf) [11]. The traditionalclustering algorithms are limited to considering geometricdistance-based similarity measures between certain datapoints, and cannot efficiently evaluate the difference betweenuncertain data objects. Lots of new clustering algorithms foruncertain data have been proposed to tackle this issue [12].Early studies on uncertain data clustering are mainlyvarious extensions of traditional clustering algorithms forcertain data, by defining new similarity measurements betweenuncertain data objects, including the ED-based similarity [13],the density-based similarity [14], and the distribution-basedsimilarity [15]. Chau et al. [13] propose the first ED-basedclustering algorithm for uncertain data named the uncertainK-means (UK-means) algorithm. It enhances the traditionalk-means algorithm with the use of a new distance-basedsimilarity, i.e., the expected distance (ED), to handle the datauncertainty.

III. PROPOSED WORK

K-Mode Clustering: The K-Means algorithm is well known for its efficiency in clustering large data sets2. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. Haung20, 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

$$d(x,y) = \sum_{j=1}^{m} \delta(x_i, y_j)$$

where $\delta(x_i, y_j) = \begin{cases} 0 & (x_i = y_j) \\ 1 & (x_i \neq y_j) \end{cases}$ and d(X,Y)gives equalimportance 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 , . . . ,M_m. When the distance function defined in Eq. (1) is used as the dissimilarity measure for categorical data objects, the cost function becomes

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

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* clustersProcedure

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

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 site5. 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, 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.

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

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 indexfor each data source.
Step-2: Merge cluster centers of local data sources into a single dataset named as 'center-dataset' at centralsite.
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 computinghamming distance between the object and global

Figure-2Ensemble based Distributed K-Modes Clustering Algorithm

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

IV. SIMULATION RESULTS

In this section, empirical evidence is provided for D-K-Md algorithm that the high quality globalcluster models is obtained with limited communication overhead and high level of privacy. The efficiency of DK-Md is compared with existing distributed clustering algorithm, DKM along with CC, where all local datasetsare merged and clustered using K-Modes algorithm. The existing DKM algorithm is not directlyendurable 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, thevalues 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 tothree sequential numbers such as 1, 2, and 3.

V. CONCLUSION

This paper emphases on uncertain data clustering problem andproposes a distributed clustering algorithm in P2P networks. Most of the prevailing distributed partitional clustering algorithms have been developed for groupingnumerical datasets. The distributed K-Modes clustering algorithm is proposed based on cluster ensemble tocluster categorical datasets in distributed environment.

REFERENCES

[1] Byung-Hoon Park and HillolKargupta. "Distributed data mining: Algorithms, systems, and applications"Data Mining Handbook, 2002.

[2] www.horatiumocian.com/s/ Distributed Clustering Survey.pdf

[3] G. Forman and B. Zhang. "Distributed Data Clustering Can Be Efficient and Exact" SIGKDD Explorations,2(2):34–38, 2000.

[4] Michael Byrd ,ConnyFranke "The State of Distributed Data Mining" *ECS265 Project Report+,http://wwwcsif.cs.ucdavis.edu/~franke/ecs265-.pdf

[5] TK Moon. "The expectation-maximization algorithm." Signal Processing Magazine, IEEE, 13(6):47–60,1996.

[6] H.P. Kriegel, P. Kroger, A. Pryakhin, and M. Schubert. "Effective and Efficient Distributed Model-basedClustering" In The Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, TX, November 2005

[10] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Querying imprecisedata in moving object environments," IEEE Trans. Knowl. Data Eng.,vol. 16, no. 9, pp. 1112–1127, Sep. 2004.

[11] A. D. Sarma, O. Benjelloun, A. Halevy, S. Nabar, and J. Widom, "Representing uncertain data: Models, properties, and algorithms," VLDB J.,vol. 18, no. 5, pp. 989–1019, May 2009.

[12] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms andApplications. Boca Raton, FL, USA: CRC Press, Sep. 2013.

[13] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: Anexample in clustering location data," in Proc. 10th Pacific–Asia Conf.

Knowl.Discovery Data Mining, vol. 3918. Apr. 2006, pp. 199–204.

[14] H. P. Kriegel and M. Pfeifle, "Density-based clustering of uncertaindata," in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery DataMining, Aug. 2005, pp. 672–677.

[15] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data basedon probability distribution similarity," IEEE Trans. Knowl. Data Eng.,vol. 25, no. 4, pp. 751–763, Apr. 2013.

BIODATA

Golla Saidulu working as Assistant Professor, Dept of CSE, in CMR College of Engineering And Technology with Experience of 3.6 years.

Masku Naveen Kumar working as Assistant Professor, Dept of CSE, in SCIENT Institute of Technology with Experience of 2 years.