UNCERTAIN DATA CLUSTERING IN DISTRIBUTED PEER-TO-PEER NETWORKS

[1] K.RATNA MOUNIKA, [2] A.CHIRANJEEVI
[1]M.Tech student, Dept of CSE, Ramachandra College Of Engineering, Eluru, India
[2]Assistant Professor, Dept of CSE, Ramachandra College of engineering, Eluru, India

ABSTRACT: Uncertain data clustering has been recognized as an essential task in the research of data mining. Many centralized clustering algorithms are extended by defining new distance or similarity measurements to tackle this issue. With the fast development of network applications, these centralized methods show their limitations in conducting data clustering in a large dynamic distributed peer-to-peer network due to the privacy and security concerns or the technical constraints brought by distributive environments. In this paper, we propose a novel distributed uncertain data clustering algorithm, in which the centralized global clustering solution is approximated by performing distributed clustering. To shorten the execution time, the reduction technique is then applied to transform the proposed method into its deterministic form by replacing each uncertain data object with its expected centroid. Finally, the attribute-weight-entropy regularization technique enhances the proposed distributed clustering method to achieve better results in data clustering and extract the essential features for cluster identification. The experiments on both synthetic and real-world data have shown the efficiency and superiority of the presented algorithm.

KEY WORDS: clustering, data mining, very large databases, parallel algorithms, distributed computing.

I.INTRODUCTION

Clustering has emerged as an essential data mining technique for statistical analysis, pattern recognition, and image segmentation. It partitions the data into clusters according to the similarities between objects and helps in extraction of new information or discovering new patterns. In the past few decades, a large number of clustering algorithms have been proposed, in which the K-means algorithm is one well-known clustering method. Then the variants of this algorithm are further discussed, and the strong consistency of this method has been proved. However, in many Real applications today, like sensor monitoring and location-based services, data mostly contains inherent uncertainty due to the random nature of the data generation, measurement inaccuracy, sampling discrepancy, data staling, and other errors. Generally, with uncertainty, the data object is no longer a single point in space but is represented by a probability density function (pdf). The traditional clustering algorithms are limited to considering geometric distance-based similarity measures between certain data points, and cannot efficiently evaluate the difference between uncertain data objects. Lots of new clustering algorithms for uncertain data have been proposed to tackle this issue.

It enhances the traditional k-means algorithm with the use of a new distance-based similarity, i.e., the expected distance (ED), to handle the data uncertainty. Then, some improved algorithms are put forward to reduce the complexity of ED calculations by using some pruning tricks or by speeding up the ED calculation itself. The work reduces the UK-means algorithm to the certain K-means (CK-means) algorithm by replacing each uncertain data object with its expected centroid, thereby tremendously decreasing the computational complexity for ED calculation. For the density-based clustering, Kriegel and Pfeifle define two fuzzy distance functions, i.e., the distance density function and the distance distribution function, to express the similarity between uncertain data objects, and they also integrate these new distance functions into the hierarchical clustering method. Different from these two kinds of similarities above, the clustering algorithms with distribution-based similarity consider using divergences to measure the similarity between data objects. Most early researches usually utilize Kullback–Libeler (KL) divergence or Bergman divergence to cluster the object with known distribution. A

recent work on uncertain data clustering is based on probability distribution similarity, in which the uncertain data object is modeled as a random variable following a probability distribution and then the KL divergence is used to directly compute the probability distribution similarity between uncertain data objects.

Here we extend the applicability of this class of parallel clustering algorithms to show that it works very well for clustering data that is inherently distributed, for example, clickstream log files at Web sites mirrored across the world. By applying the parallel version of the clustering algorithms, the data can be clustered in-place with the exact same computational result as if the data set had been assembled at a central site for clustering and the communication costs and delays for transmitting large volumes of data to the central site. The extra storage space and computing resources that would be needed at the central site. The administration complexity needed to manage copies of the remote portions of the data set, r. many algorithms for data clustering developed in recent decades all face a major challenge in scaling up to very large database sizes. Clustering algorithms with quadratic (or higher order) computational complexity, such as agglomerative algorithms, do not scale up. Even for more efficient algorithms, such as K-Means and Expectation-Maximization, which have linear cost per iteration, research is needed to improve their ability to handle ever-growing data sets.

Performing clustering on the weather condition data can reveal interesting insights on the weather correlation between different regions of the city in different months. In these new applications, data sources are distributed over a large network containing no special central control. The traditional centralized clustering approaches for uncertain data have shown the weaknesses: 1) raw information sharing is discouraged due to the confidentiality and security requirements in distributed P2P networks; 2) effective data collection from all peers to the central site is not guaranteed due to the energy or bandwidth limitations; and 3) high-computational complexity with large data sets. These motivate seeking a new clustering algorithm in distributed network environments for uncertain data, i.e., the distributed uncertain data clustering.

II. BACKGROUND

Fundamentally challenging to perform data clustering on very large databases. There have been several recent publications in scaling up K-Means and EM by approximation. For example, in BIRCH a single scan of the data and subsequent aggregation of each local cluster into a single representative "point" containing sufficient statistics for the points it represents enables a data set to be pared down to fit the available memory. Such algorithms provide an approximation to the original algorithm and have been successfully applied to scale up to very large datasets. However, the higher the aggregation ratio, the less accurate the results are in general. It is also reported in the BIRCH paper that the quality of the clustering depends on the original scanning order.

There is also recent work on non-approximated, parallel versions of K-Means. The Kantabutra and Couch algorithm [KC99] rebroadcasts the data set to all computers each iteration, which leads to heavy network loading and significant communication protocol processing overhead. Their analytical and empirical analysis estimates 50% utilization of the processors; such an algorithm becomes completely impractical in a distributed wide-area networking (WAN) environment. Even in a local-area networking (LAN) environment, the technology trend is for processors to improve faster than networks are improving, making the network a greater bottleneck in the future. Finally, their algorithm limits the number of computing

units to the number of clusters to be found. The parallel algorithm by Dhillon & Modha was discovered independently and is an instance of the class of parallel algorithm. In our previous paper, we described a parallel decomposition for center-based clustering algorithms that limits inter-processor communication to sufficient statistics only, reducing the network bottleneck. The data set is partitioned randomly across the memory of the processors and does not need to be transferred between iterations. Load balancing can be achieved trivially by migrating data points selected arbitrarily. The number of computing units is not limited in any way by the number of clusters sought.

The results are exactly as if the original algorithm were run on a single computer, i.e. no approximation. The method can be used in conjunction with sampling or aggregation techniques by combining with our approach, even larger data sets can be handled or better accuracy can be achieved by less aggregation. Because the amount of communication is small, the parallel decomposition achieves excellent speed-up efficiency on networks of workstations without any special low-latency or high-bandwidth interconnect--note that the computing resources in a collection of PCs or desktop workstations can easily exceed the total computing resources available in a supercomputer. Further, they can be considered a free resource off they can be utilized when they would otherwise be idle. Because of the small amount of communication per iteration, in this paper commend this algorithm for the geographically distributed case where the interconnect is a high-latency, low bandwidth WAN, instead of a common LAN.

III. LITERATURE SURVEY

Uncertain data clustering has been recognized as an essential task in the research of data mining. Many centralized clustering algorithms are extended by defining new distance or similarity measurements to tackle this issue. With the fast development of network applications, these centralized methods show their limitations in conducting data clustering in a large dynamic distributed peer-to-peer network due to the privacy and security concerns or the technical constraints brought by distributive environments. In this paper, we propose a novel distributed uncertain data clustering algorithm, in which the centralized global clustering solution is approximated by performing distributed clustering. To shorten the execution time, the reduction technique is then applied to transform the proposed method into its deterministic form by replacing each uncertain data object with its expected centroid. Finally, the attribute-weight-entropy regularization technique enhances the proposed distributed clustering method to achieve better results in data clustering and extract the essential features for cluster identification.

In this paper, we have proposed a new prototype-based classifier, based on adjusted self-organizing incremental neural network (SOINN). We call this method the adjusted SOINN classifier (ASC). Using an adaptive similarity threshold, the system can grow incrementally and accommodate input patterns of incremental data distribution. By deleting the within-class insertion, the system requires fewer parameters than SOINN. The ASC can reduce the prototypes caused by noise and make it robust to noise and possible to achieve a low classification error. The deletion of unnecessary prototypes during the classification process makes ASC much faster than some other classifiers. In the experiment, ASC is compared with some other classifiers in terms of the classification error, compression ratio, and speed up ratio. ASC achieves the best performance and shows that it is a very efficient method.

In spite of its computational efficiency and wide spread popularity, the FCM algorithm does not take the spatial information of pixels

into consideration. In this paper, a multiple kernel fuzzy c-means clustering (MKFCM) algorithm is presented for fuzzy segmentation of magnetic resonance (MR) images. By introducing a novel adaptive method to compute the weights of local spatial values in the objective function, the new multiple kernel fuzzy clustering algorithm is capable of utilizing local contextual information to impose local spatial continuity, thus improving the classification accuracy and reduces the number of iterations. To estimate the intensity in homogeneity, the global intensity is introduced into the coherent local intensity clustering algorithm. Our results show that the proposed MKFCM algorithm can effectively segment the test images and MR images. Comparisons with other FCM approaches based on number of iterations and time complexity demonstrate the superior performance of the proposed algorithm.

Data analysis plays an indispensable role for understanding various phenomena. Cluster analysis, primitive exploration with little or no prior knowledge, consists of research developed across a wide variety of communities. The diversity, on one hand, equips us with many tools. On the other hand, the profusion of options causes confusion. We survey clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. Factorial k-means (FKM) clustering is a method for clustering objects in a low-dimensional subspace. The advantage of this method is that the partition of objects and the low-dimensional subspace reflecting the cluster structure are obtained, simultaneously. In some cases that the reduced k-means clustering (RKM) does not work well; FKM clustering can discover the cluster structure underlying a lower dimensional subspace. Conditions that ensure the almost sure convergence of the estimator of FKM clustering as the sample size increases unboundedly are derived. The result is proved for a more general model including FKM clustering.

IV. PROPOSED ALGORITHM

To find K cluster centers, a center-based data clustering problem is formulated as an optimization (minimization) of a performance function, Peo((S, M), depending on both the N points in the data set S and the K center location estimates M. A popular performance function for measuring the goodness of a clustering is the total within-cluster variance, or the sum of the mean-square error (MSE) of each data point to its center. The K-Means algorithm attempts to find a local optimum for this performance function, and is one of the most popular, used widely across many disciplines for its easily interpretable result. The K-Harmonic Means (KHM) algorithm optimizes the harmonic average of these distances. The advantages of KHM are that its convergence rate can be adjusted with a parameter, and it is highly insensitive to the initialization of the center locations, a major problem for KMeans that many authors have tried to address with clever initializations. The Expectation-Maximization (EM) algorithm is also widely used and, in addition to the centers, optimizes a covariance matrix and a set of mixing probabilities.

This algorithms fit a class of iterative center-based clustering algorithms that parallelizes as follows:

1. Arbitrarily distribute the N elements of the data set S to the local memories of a set of P computers.
2. Pick the K initial center location estimates M by any scheme, such as a random sample. A coherent copy is kept on each computer throughout the computation.
3. Iterate:

3.1. Each computer independently computes its contribution to a set of global sufficient statistics SS, which includes information for computing the performance function.

3.2. Global reduction (summation across processors) of the sufficient statistics, followed by broadcasting the global results back to all computers.

3.3. Independent local computation to adjust the center location estimates. The results are identical on each computer and are exactly the same as the uniprocessor sequential algorithm would produce.

4. Stop when the performance function converges, or after a fixed number of iterations.

5. Output the K centers M. regarding the distribution of the data set: the partitioning may be arbitrary and has nothing to do with the clustering structure in the data.

It has no effect on the computed results and is static, unless one wishes to migrate some data points for load balancing to enhance speedup efficiency. The sizes of the partitions, besides being constrained by the storage of the individual units, are ideally set to be proportional to the speed of the computing units. Partitioned thus, it will take about the same amount of time for each unit to finish its computation in each iteration, optimizing overall efficiency. Unlike K-Means and K-Harmonic Means in which only the centers are to be estimated, the EM algorithm also estimates the co-variance matrices and the mixing probabilities. We found that the communication latency varies little. The steady transfer rate varies more. Since the amount of sufficient statistics that have to be transferred over the network is typically small (e.g. a few kilobytes), the communication delay is dominated by latency. Hence, the variation on the steady transfer rate of the network has little impact on the performance of the distributed clustering algorithm. The below figure (1) shows the comparison of existed and proposed system.

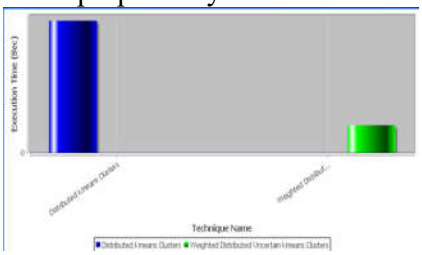

FIG. 1: TIME COMPARISON OF EXISTED AND PROPOSED SYSTEM

V. CONCLUSION

This paper focuses on uncertain data clustering problem and proposes a distributed clustering algorithm in P2P networks. The centralized clustering solution is obtained in a distributive mode at each peer by collaborating with the neighboring peers only. Based on the reduction technique, the distributed uncertain data clustering algorithm actually turns out to be equivalent to the deterministic clustering, which greatly shortens the execution time of the algorithm. The attribute-weight-entropy regularization technique is applied in the distributed clustering method to achieve ideal distribution of attribute weights, which ensures the good clustering results. Experiments on several synthetic and real-world data sets have demonstrated the good performance of the proposed algorithms. The results of this paper provide some valuable directions for future work. The proposed algorithm is of great generality and could be further applied in uncertain data clustering research in distributed environments. Currently, most of the study on clustering set the number of clusters as a user-defined parameter, which is difficult to specify.

VI. REFERENCES

[1] V. N. Vapnik, Statistical Learning Theory. New York, NY, USA: Wiley, 1998.

[2] L. Wang, B. Yang, Y. Chen, X. Zhang, and J. Orchard, "Improving neural-network classifiers using nearest neighbor partitioning," IEEE Trans. Neural Netw. Learn. Syst., to be published, doi: 10.1109/TNNLS.2016.2580570, 2016.

[3] L. Chen, C. L. P. Chen, and M. Lu, "A multiple-kernel fuzzy C-means algorithm for image segmentation," IEEE Trans. Syst., Man B, Cybern., vol. 41, no. 5, pp. 1263–1274, Oct. 2011.

[4] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, May 2005.

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley

Symp. Math. Statist. Probab., vol. 1. 1967, pp. 281–297.

[6] T. Zhang, L. Chen, and C. L. P. Chen, "Clustering algorithm based on spatial shadowed fuzzy C-means and I-ching operators," Int. J. Fuzzy Syst., vol. 18, no. 4, pp. 609–617, Aug. 2016.

[7] L. Chen, J. Zou, and C. L. P. Chen, "Kernel spatial shadowed C-Means for image segmentation," Int. J. Fuzzy Syst., vol. 16, no. 1, pp. 46–56, Mar. 2014.

[8] D. Pollard, "Strong consistency of K-means clustering," Ann. Statist., vol. 9, no. 1, pp. 135–140, Jan. 1981.

[9] V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, and C. Priebe, "Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding," Electron. J. Statist., vol. 8, no. 2, pp. 2905–2922, Mar. 2014.

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