International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

DMMI Methods for Theoretic Clustering using Data Mining

Ankita G. Joshi[1], R. R. Shelke[2]

[1]HVPM COET, Amravati, Maharashtra, India

[2]Professor, HVPM COET, Amravati, Maharashtra, India

Abstract: ***The main objective of the data mining process is to extract information from a large data set and transform it into an understandable structure for further use. Clustering is a main task of exploratory data analysis and data mining applications. Theoretic Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). The objective of clustering is typically exploratory in nature; we desire clustering algorithms that make as few assumptions about the data as possible. Distributed clustering is to explore the hidden structure of the data collected/stored in geographically distributed nodes. Information theoretic measures take the whole distribution of cluster data into account for better clustering results. For this, we incorporate an information theoretic measure into the cost function of the distributed clustering. We interpret the motivation for choosing the MMI (Maximum Mutual Information) criterion to develop distributed clustering algorithms. The proposed Linear and Kernel DMMI algorithms can achieve almost as good clustering results as the corresponding centralized information theoretic clustering algorithms on both synthetic and real data.***

Keywords: Data mining, Theoretic clustering, Information theory, Mutual information.

1. Introduction

Data Mining has been gaining popularity in knowledge discovery field, particularity with the increasing availability of digital documents in various languages from all around the world. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [1]. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. Figure a show the data mining process for any data input to get valid output. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly [2].

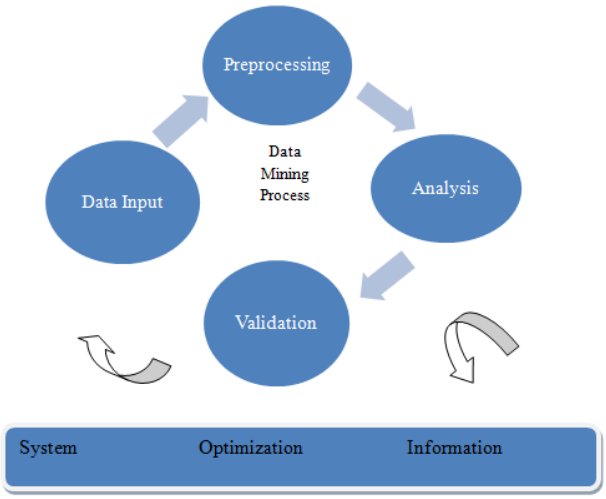

Figure a: Data Mining Process

Clustering provides a common means of identifying structure in complex data, and there is renewed interest in clustering as a tool for the analysis of large data sets in many fields. Data clustering is to explore the hidden structure of data and group data items into a few clusters in an unsupervised way [10].

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept.

Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning [11].

A good clustering method will produce high quality clusters with high intra-cluster similarity and low intercluster

similarity. The quality of a result produced by clustering depends on both the similarity measure used by the method and its implementation [17]. The quality of a clusters produced by clustering method is also measured by its ability to discover some or all of the hidden patterns. Other requirements of clustering algorithms are scalability, ability to deal with insensitivity to the order of input records and with noisy data [18].

Several distributed data clustering techniques have been developed based on the K-means algorithm or the Gaussian mixture model. In these methods, data structures are captured by measures only based on the first and the second order statistics. When the structure of cluster data is complicated, these statistics are insufficient and may lead to unsatisfactory clustering results. To get good clustering performance, we use information theoretic measure.

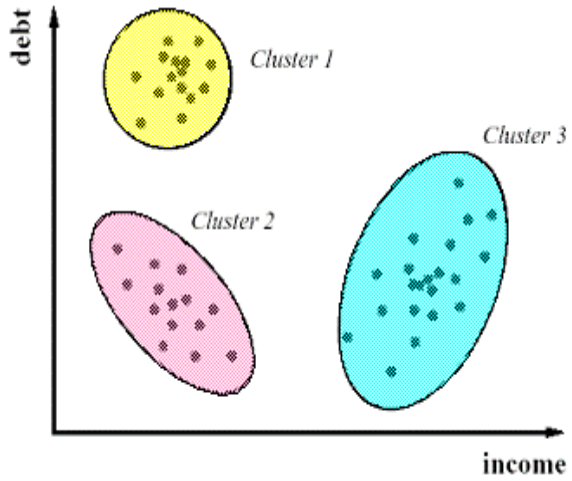

Figure b: An example of cluster

Most of the existing distributed data clustering algorithms are based on the K-means method or the Gaussian mixture model (GMM). In K-means based clustering algorithms, the cost functions usually measure the sum of distances (squared differences) between data items and estimated centroids of clusters. In GMM based clustering algorithms, the distribution of an individual cluster is assumed to be Gaussian, which is fully specified by its mean/centroid and variance. In these kinds of methods, data structures are captured by measures only based on the first and the second order statistics.

Centralizing the whole distributed data to one fusion node to perform centralized clustering may be impractical. Thus, there is a great demand for distributed data clustering algorithms in which the global clustering problem can be solved at each individual node based on local data and limited information exchanges among node [15]. In addition, compared with the centralized clustering, distributed clustering is more flexible and robust to node and/or link failure. We present distributed clustering algorithms based on an information theoretic measure. We incorporate the maximum mutual information (MMI) criterion into the cost function in distributed clustering, to present distributed MMI-based (DMMI) clustering algorithms.

In our method, each node solves a local clustering problem through cooperation with its single-hop neighboring nodes. Each node can utilize global information to help clustering its local data, at a low communication cost. Besides, in the cooperation, nodes do not transmit original data but merely exchange a few parameters of clusters [16]. Hence, clustering task is performed under privacy preservation, which is important in some practical distributed applications [3], [4].

2. Literature Review

Nowadays we have many applications with massive amount of data which are caused limitation in data storage capacity and processing time. We can identify two main groups of techniques for huge data bases mining. One group refers to streaming data and applies mining techniques whereas second group attempts to solve this problem directly with efficient algorithms. Data considered as a stream of data which come in from one side and exit from another side so we aren't able to visit data for second time just like river. This data stream's main property arise some difficulties. Two main problems which are related to this property include:

1) One scan is possible for processing data.
2) Data is included evolutionary stream and concepts are changed during the time. Data stream mining is categorized in three main techniques: classification, clustering and association rules extraction.

Many researchers interest is to apply some techniques for increasing compactness of representation, fast and incremental processing of new data points, clear and fast identification of outliers. Primitive Clustering Methods and Data Stream Clustering Methods are used in data stream clustering. Most data stream applications are high dimensions and new methods need to be developed. Concept drift is nature of data stream and should be managed by new methods. On the other hand data stream is similar to a river: data come in and come out. Evolving data, visiting data once and space limitations are major issues in data stream clustering. For the divergence-based clustering, there are roughly two types of algorithms, which are the parametric type and the nonparametric type, respectively. The Bregman soft clustering algorithm is a representative and typical sample for the former [5]. In [5], the authors model the data source with a mixture of exponential family distributions (one component for one cluster), and pose the clustering problem as a parameter estimation problem for the mixture model. They find the correspondence between exponential families and regular Bregman divergences, and thereby bring up a Bregman divergence viewpoint for learning the maximum likelihood parameters of the mixture model. The algorithm provides a framework for clustering different datasets by using different Bregman divergences (or equivalently, parametric models of different exponential distributions). For a given application (dataset), to obtain good clustering performance, it is expected to artificially choose a specific Bregman divergence (or equivalently, parametric model of a specific exponential distribution) which matches the generative model of current data. For a clustering result, large divergence means there are obvious differences or boundaries between data items belonging to

different clusters. Hence, their goal is to maximize the divergence, by adjusting the assignment of cluster/class label on each data item. In this kind of method, calculating divergence relies on unknown *conditional pdfs of cluster data*, which need to be estimated during clustering. In order to make clustering adaptable to datasets of different data structures, the authors choose to directly estimate the conditional pdfs from labeled data in a *nonparametric* manner [6], [7], rather than to model them by predefined parametric models, e.g. exponential family distributions. Accordingly, the optimization of corresponding cost functions are directly related to the cluster label of each data item. Note that, when the algorithms are extended to the distributed clustering field, this characteristic would lead to request for transmission of original data (it may be not necessary under some kind of modification, however, we have not yet found an efficient modification scheme to avoid the transmission of data while maintaining good clustering performances).

Distributed Mmi-Based Clustering

We considered a network composed of nodes distributed over a geographic region. Every node collects/stores a set of data items denoted by, where are -dimensional data items, or named by feature vectors, with components. For each node, the data items are considered to be samples of a random variable with probability measure, and the random variables are supposed to follow the same probability measure.

In other words, the data items can also be viewed as part of samples of a global random variable with probability measure. The total number of data samples for over the whole network is. Without loss of generality, we model the network by a connected graph, where denotes the node set and denotes the edge set [8].

If two nodes are connected by an edge, then they are the one-hop-communication neighbor for each other. All the one-hop neighbors of node and itself constitute its neighbor set. Node is supposed to cluster its local data into different classes based on cooperation with nodes belonging to. In other words, each data item stored in node needs to be attached with a class label. The class label is also considered to be a random variable with probability measure.

Note that though direct cooperation is limited within one-hop neighbors, in a connected graph, information shared by one node can still be diffused over the whole network in the following steps. Thus each node actually can utilize global information in its local clustering, which makes it possible that distributed clustering algorithms achieve as good clustering results as the corresponding centralized clustering algorithms [13].

3. Analysis of Problem

- In centralized clustering algorithm, due to unsupervised way the large amounts of data are not centrally collected/stored in one source but dispersedly collected/stored in geographically distributed nodes over networks.
- Centralizing the whole distributed data to one fusion node to perform centralized clustering may be impractical.
- Privacy preservation and communication resource saving (original data can be large) are usually important in real distributed applications. So, directly transmitting original data is not preferred.
- As real datasets are not always linearly separable, the linear DMMI algorithm presented may not work well, when the boundaries between different clusters are complicated.
- Drawbacks of theoretic clustering are complexity and inability to recover from database corruption.
- Testing on large datasets with hadoop not possible and Working on structural graph data not possible.
- Processing data streams regarding architectural aspects have received considerable attention, but most effort is concentrated on the mining and clustering aspects of the problem.
- Algorithms suffer from the ability to handle difficult clustering tasks without supervision.

- The algorithms required expert assistant in the form of the number of partitions expected or the expected density of clusters.
- Required to re-learn any recurrently occurring patterns, Compactness and separateness of data.
- Efficiency in terms of speed is a vital problem in data mining clustering.
- Some difficulties in realizing exact clusters of data in many applications is due to arbitrary shape causes.
- The most popular challenges in data stream clustering is outliers detecting and High dimensionality is one of the major causes in data complexity.
- Data type treatment, Cluster validity, Space limitation, High dimensional data stream and uncertain data.

4. Proposed Methodology

In this paper, various datasets will be collected which will contain documents for theoretic clustering. This datasets will be used for the evaluation of the project. Afterwards, the diffusion kernel DMMI algorithm will be implemented, which will use the distributed maximum mutual information and used this for the cost function in distributed clustering [14]. Then, the accuracy of algorithm will be evaluated on all the datasets. Various data mining techniques will be studied and results will be checked from the base papers of the techniques.

The best technique for data mining will be implemented for the desired clustering algorithm and result will be checked for future evaluation. Finally, the Data mining technique will be combined with diffusion kernel DMMI technique, in order to improve the clustering output.

Several distributed data clustering algorithms have been proposed based on the:

- K-means method
- The Gaussian mixture model (GMM)

We present distributed clustering algorithms based on an information theoretic measure [12]. We incorporate the maximum mutual information (MMI) criterion into the cost function in distributed clustering, to present distributed MMI-based (DMMI) clustering algorithms named as:

- LINEAR DMMI
- KERNEL DMMI

Since real datasets are not always linearly separable, the linear DMMI algorithm may not work well when the boundaries between different clusters are complicated. In order to handle linearly non-separable problems, the kernel DMMI algorithm is proposed by using the modified kernel discriminative clustering function

5. Conclusion

The overall goal of the data mining process is to separate the information from a large data set and transform it into an understandable form for further use. Clustering is an important task in data analysis and data mining applications. An optimal method for Document/ theoretic clustering has been proposed, which is both memory and time efficient, while maintaining good level of accuracy. Through this paper, we have compared the K-means-based and the GMM-based clustering algorithms with the MMI-based algorithms [9], and the MMI based algorithms can capture the data structures beyond the first and the second order statistics, thus leading to more satisfactory clustering results for datasets with complicated data structures. The linear DMMI algorithm is appreciated for linearly separable problems. But in order to handle linearly non-separable problems, we have proposed the kernel DMMI algorithm by using the modified kernel discriminative clustering function. The kernel DMMI shows excellent ability in exploring the overall data structure for the real atmosphere dataset, which indicates that the proposed information theoretic clustering algorithms are applicable in real distributed applications like environmental monitoring. The performances of the proposed algorithms are evaluated on three synthetic datasets and one real dataset. To reflect the flexibility and applicability of the algorithms for practical cases, the proposed two DMMI algorithms maintain good clustering performance in the cases of unbalanced data distribution over nodes.

References

[1] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
[2] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
[3] S. Merugu and J. Ghosh, "A privacy-sensitive approach to distributed clustering," *Pattern Recognit. Lett.* vol. 26, no. 4, pp. 399–410, 2005.
[4] A. M. Elmisery and H. Fu, "Privacy preserving distributed learning clustering of healthcare data using cryptography protocols," in *Proc. IEEE 34th Ann. Computer. Software and Appl. Conf. Workshops*, 2010, pp. 140–145.
[5] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 158–170, April 2002.
[6] R. Jenssen, T. Eltoft, and J. C. Principe, "Information theoretic clustering: A unifying review of three recent algorithms," in *Proc. 6th Nordic Signal Process. Symp.* Espoo, Finland, 2004, pp. 292–295.
[7] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.
[8] P. A. Forero, A. Cano, and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE J. Sel. Topics Signal Process.* vol. 5, no. 4, pp. 707–724, Aug. 2011.
[9] R. Xu and D. Wunsch-II, "Survey of clustering algorithms," *IEEE Trans. Neural Network*, vol. 16, no. 3, pp. 645–678, May 2005.
[10] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computer. Surveys*, vol. 31, no. 3, pp. 265–323, 1999.
[11] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 158–170, April 2002.
[12] J.Grabmeier and A. Rudolph, "Techniques of cluster algorithms in data mining," *Data Mining and Knowledge. Discovery*, vol. 6, no. 4, pp. 303–360, 2002.
[13] P. A. Forero, A. Cano, and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE J. Sel. Topics Signal Process*, vol. 5, no. 4, pp. 707–724, Aug. 2011.
[14] M. Klusch, S. Lodi, and G. Moro, "Distributed clustering based on sampling local density estimates," in *Proc. Int. Joint Conf. Artificial. Intell.* 2003, pp. 485–490.
[15] C. Li, P. Shen, Y. Liu, and Z. Zhang, "Diffusion information theoretic learning for distributed estimation over network," *IEEE Trans. Signal Process*, vol. 61, no. 16, pp. 4011–4024, Aug 15, 2013.
[16] A. M. Elmisery and H. Fu, "Privacy preserving distributed learning clustering of healthcare data using cryptography protocols," in *Proc. IEEE 34th Ann. Computer Software and Appl. Conf. Workshops*, 2010 pp. 140–145.
[17] L. O Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-data algorithms for high-quality clustering," 2002.
[18] D. Barbara, "Requirements for clustering data streams," *ACM SIGKDD Explorations Newsletter*, vol. 3, pp. 23-27, 2002.