Enhancement of CURE Clustering Technique in Data Mining

Seema Maitrey[1], C. K. Jha[2],Rajat Gupta[3] ,Jaiveer Singh[4]
[1] Deptt. of Computer Science & Engg., Banasthali University, Rajasthan, India
[2,3,4] Deptt. Of Computer Science and Engineering,
Krishna Institute of Engg. and Technology, Ghaziabad,(UP), India

ABSTRACT

The precious information is embedded in large databases. To extract them has become an interesting area of Data mining. Clustering, in data mining, is useful for discovering groups and identifying interesting distributions in the underlying data [5]. Among several clustering algorithms, we have considered CURE method from hierarchical clustering. CURE (Clustering usage Representatives) method find clusters from a large database that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE employs a combination of data collection, data reduction by using random sampling and partitioning. With the availability of large data sets in application areas like bioinformatics, medical informatics, scientific data analysis, financial analysis, telecommunications, retailing, and marketing, it is becoming increasingly important to execute data mining tasks in parallel. At the same time, technological advances have made shared memory parallel machines commonly available to organizations and individuals. Although CURE provide high quality clustering, a parallel version was not available. Our new algorithm enabled it to outperform existing algorithms as well as to scale well for large databases without declining clustering quality.

Keywords

Data Mining, KDD, Clustering, Issues, Parallelism.

1. INTRODUCTION

Extremely large amount of data is being captured by today’s organizations and is continue to increase. It becomes computationally inefficient to analyse such huge data. Research in data mining has addressed problem in discovering knowledge from these continuously growing large data sets. The amount of raw data available to researchers, in a variety of scientific fields, has been increasing at an exponential rate. A common method used in data exploration is data clustering [1]. Traditional methods provide reasonable approximations of total solutions for extremely large datasets [17];[16];[18]. However, typically these approaches incrementally update solutions as data becomes available, making them susceptible to noise [17]. In this paper, we explore the problem of approximating clustering solutions for continuously generated data streams, specifically focusing on efficiently providing an accurate solution. We present an approach that aggregates cluster summaries over multiple time periods into one comprehensive solution for the entire duration. Data streams are clustered using the Hierarchical approach by implementing its methods: CURE [17]. The result is an accurate approximation of the clustering solution for the entire data set (or data stream) that requires substantially less computational effort. The scalability of data mining algorithms has become an active area of research with many challenging problems. There are five basic approaches for scalability of data intensive computing, such as: [19]

Manipulating data to fit into memory: There are four basic variants of this approach, namely Sampling, Feature selection, Partitioning, Data summarization.

Reducing the access time: This approach tries to reduce the access time for accessing out of memory data by using specialized data structure to access data on disk such as B+ tree and using some technique such as parallel block reads.

Using multiple processors: There are two basic techniques used to speed up algorithms: Task Parallelism & Data Parallelism.

Pre-computation: It involves Pre-computation of most expensive part of the algorithm, such as sorting, if possible.

Data Reduction: Techniques such as Data Smoothing, Data Compression and Data Transformation can be used.

The parallelism in an algorithm can yield improved performance on many different kinds of computers. For example, on a parallel computer, the operations in a parallel algorithm can be performed simultaneously by different processors.

The rest of our paper is organized as following: In Section 2, we review the related work done by various researchers in the past. We explain about our motivation and goal in section 3, Our contribution of CURE algorithm in section 4, Future scope in section 5 and conclusion in section 6. In section 7 the references that are used for this paper are presented.

2. DATA MINING: REVIEW OF LITERATURE

Data mining is the process of extracting hidden patterns automatically from database by employing one or more computer learning techniques. Data mining is an induction based learning strategy that builds models to identify hidden patterns in data. A model created by a data mining algorithms is a conceptual generalization of data. The generalization may be in the form of tree, a network, an equation or a set of rules. Data mining is a multi-step process that requires accessing and preparing data for a data mining algorithm, mining the data, analyzing results and taking appropriate action. The data to be accessed can be stored in one or more operational databases, a data warehouse or a flat file. Data is mined using supervised learning or unsupervised clustering. The most common type of data mining is supervised. It determine the classification of new, previously unclassified instances. With unsupervised clustering, predefined classes do not exist.

Instead, data instances are grouped together based on a similarity scheme defined by the clustering model. Data mining is different from data query. Database query language and OLAP tools are very good at finding and reporting information within a database when we know exactly what we are looking for. Data mining is a step ahead process that provides potentials useful information even when little or no idea is available what are looking for [9].

In any organization, the amount of data stored in database continues to grow fast.

This large amount of stored data contains valuable hidden knowledge, which is used to improve the decision making process of the organization. Data about previous sales might contain interesting relationships between products and customers. Thediscovery of such relationships can be very useful to increase the sales of the company. Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either shifting a large amount of material or ingeniously probing the material to exactly pinpoint where the value reside. To discover such types of relationship data mining is used on datasets as a four-step process.

Data Mining is a four step – process[4]

i) Assemble data so data mining can be applied on it. These assembled data are kept in database. There is no need to assemble huge amount of data for the data mining algorithm. Most of the data mining algorithms work best on a few datasets.

ii) Apply data mining tools on datasets. At this stage many choices have to be made. Should learning be supervised or unsupervised, which instance of data will be used to build model and which instance will test the model, which attributes will be selected from the list of available attributes.

iii) Interpretation and evaluation of result evolves examining the output of data mining tool to determine whether what has been discovered is useful and interesting. If the result is not satisfactory, data mining step is repeated using new attributes or process is sent back to data warehouse for repeating data extraction process.

iv)Result application is the final step of data mining process. The ultimate goal of data mining is to apply the result in new situation with changed attributes or instances. The result is applied in real life situation to solve the problem.

Data Mining and Knowledge Discovery

With the enormous amount of data stored in files, databases and other repositories, it is increasingly important to develop powerful means for analyzing and interpretation of such data and for the extraction of interesting knowledge that could help in decision making. Data mining, also known as Knowledge Discovery in Databases(KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While Data mining and Knowledge Discovery in Databases are frequently treated as synonyms, Data mining is actually part of the knowledge discovery process [8].

The knowledge discovery in databases process comprises of a few step leading from raw data collections to some form of new knowledge. The knowledge discovery in databases is an iterative process undergoing the following stages:

i)Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

ii)Data integration: at this stage, multiple data sources, often heterogeneous may be combined in a common source.

iii)Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

iv)Data transformation: also known as data consolidation. It is a phase in which the selected data is transformed into forms appropriate for the mining process.

v)Data mining: it is the crucial step in which cleaver techniques are applied to extract patterns potentially useful.

vi)Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on a given measures.

vii)Knowledge representation: this is the final phase in which the discovered knowledge is visually representation to the user. This essential step uses visualization techniques to help users understand and interpret the data mining result.

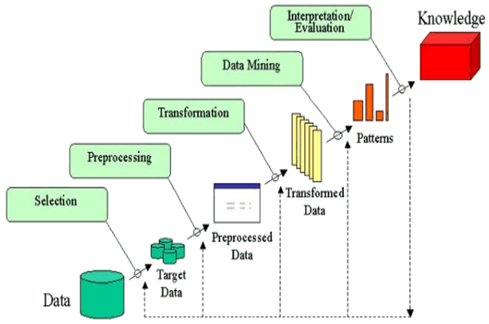

Figure1: Steps of KDD process

It is common to combine some of these steps together. The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation Application of Data Mining measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or a new data sources can be integrated, in order to get different, more appropriate results.

3. THE GOALS OF CLUSTERING

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? [10] It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings [8] ("useful" data classes) or in finding unusual data objects (outlier detection). .

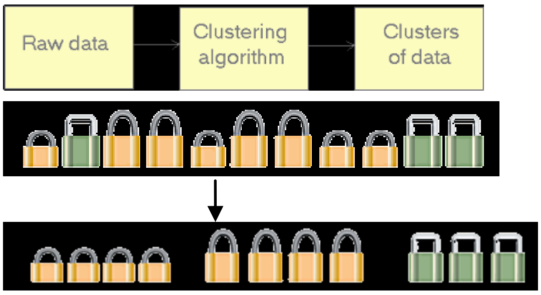

Figure 2: Stages of Clustering

3.1 Components of a Clustering Task

Typical pattern clustering activity involves the following steps

(1) Pattern representation (optionally including feature extraction and/or selection),

(2) Definition of a pattern proximity measure appropriate to the data domain,

(3) Clustering or grouping,

(4) Data abstraction (if needed), and

(5) Assessment of output (if needed).

3.2 Requirements of Clustering

- Scalability;
- Dealing with different types of attributes;
- Discovering clusters with arbitrary shape;
- Minimal requirements for domain knowledge to determine input parameters;
- Ability to deal with noise and outliers;
- Insensitivity to order of input records;
- High dimensionality;
- Interpretability and usability.

3.3 Problems of Clustering

There are a number of problems with clustering. Among them:[15].

- Current clustering techniques do not address all the requirements adequately (and concurrently);
- Dealing with large number of dimensions and large number of data items can be problematic because of time complexity;

3.4Clustering Algorithms

- Hierarchical Methods
- Partitioning Methods
- Density-Based Algorithms
- Grid-Based Methods
- Constraint-Based Clustering
- Support Vector Machine (SVM)

Among these, we have used one of the most popular algorithm called as hierarchical clustering with its CURE clustering approach.

4. MAJOR ISSUES IN DATA MINING

Number of issues are found in data mining which limited it for using in handling large databases. They are as:

1. *Mining methodology and user-interaction issues.* These reflect the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad-hoc mining, and knowledge visualization[6].

2. *Performance issues.* These include efficiency, scalability, and parallelization of data mining algorithms.

3. *Issues relating to the diversity of database types.*

5. OUR CONTRIBUTION

To improve the efficiency and increase the performance of clustering and removing one of the issue in clustering technique, we proposed a generalized parallelization algorithm in CURE clustering.

Design of Parallel algorithm

Basic Techniques of Parallel Processing

The various programming models of parallel processing are characterized by two factors, namely: type of parallelism used and type of inter-processor communication used. Data can be divided into partitions, so also the task into subtasks. Accordingly, based on whether a data is divided into partitions or a task into subtasks, three kinds of parallelism can be identified viz.

Data Parallelism, Task Parallelism and Hybrid Parallelism. Based on how different processors communicate, there are three kinds of parallel processing systems, namely Shared Memory Systems, Message Passing Systems or Remote Memory Operations System.

Data Parallelism: Original data set is divided into number of partitions and the same program runs on each of the partitions. Results obtained by running the program on each data partitions is later on combined to get the final result.

Task Parallelism: Main task is divided into set of smaller subtasks which are identified and assigned to individual processors. The results obtained by these independent processors are used to get the final solution. This speeds up the execution by multifold as the tasks are performed concurrently.

Hybrid Parallelism: Both data and tasks are divided. The subtasks that can be performed on independent data partitions, whose results can be combined together, are identified along with the independent subtasks. These subtasks are assigned to independent processors in proper sequence and using the results of these parallel processing, final result is obtained.

Shared Memory Systems: All the processors share a common global memory. Locking of shared memory is used when two or more processors try to use the shared memory for writing.

Message Passing: Each processor has its own memory. The processors communicate by sending and receiving the messages explicitly using send and receive commands.

Remote Memory Operations: Processors can explicitly access memory of other processors. Separate commands are used for accessing local and remote memory. Parallel computing allows using multiple CPUs where a problem is broken into discrete parts that can be solved concurrently. Each part is further broken down to a series of instructions and instructions from each part execute simultaneously on different CPUs. Before making new algorithm for existing CURE clustering technique, we have a look to common data mining algorithm, such as :

```
{//Outer Sequential Loop//    }
        While()     {
        {    //Reduction Loop//    }
        Foreach (element e)   {
        (i, val)     =   process( e ) ;
        Reduc(i) = Reduc(i) op val ;
        } }
```

Figure 3: Common data mining algorithms

Parallel Architectures

A parallel computer or multi-processor system is a computer utilizing more than one processor. A common way to classify parallel computers is to distinguish them by the way how processors can access the system's main memory because this influences heavily the usage and programming of the system.

Parallel Performance Analysis

Performance analysis is an iterative subtask during program development. The goal is to identify program regions that do not perform well. Performance analysis is structured into three phases:

1. Measurement: Performance analysis is done based on information on runtime events gathered during program execution. The basic events are, for example, cache misses, termination of a floating point operation, start and stop of a subroutine or message passing operation.

2. Analysis: During analysis the collected runtime data are inspected to detect performance problems. Performance problems are based on performance properties, such as the existence of message passing in a program region, the programmer applies a threshold. Only performance properties whose severity exceeds this threshold are considered to be performance problems.

3. Ranking: During program analysis the severest performance problems need to be identified i.e. problems need to be ranked according to the severity. Current techniques for performance data collection are profiling and tracing. Profiling collects summary data only. This can be done via sampling.

Tracing is a technique that collects information for each event.

Step- By – Step Analysis of Our Work

We designed the parallel algorithm for CURE by using data parallelism and shared memory systems in CRCW PRAM. We measured the elapsed time, the speedup and the scaleup of CURE. Speedup gives the efficiency of the parallel algorithm during the change in number of processors. Another interesting measure is scaleup. Scaleup captures how a parallel algorithm handles larger datasets when more processors are available. We observed the elapsed times of CURE on different numbers of processors. We see that the total execution time substantially decreases as the number of used processors increases. In particular, for the largest datasets the time decreases in a more significant way. We can observe that as the dataset size increases the time gain increases as well. For the speedup results obtained for different datasets, we can observe that the CURE algorithm scales well up to 12 processors for the largest datasets, whereas for small datasets the speedup increases until the optimal number of processors are used for the given problem, e.g., 3 processors for 5000 tuples or 7 processors for 20000 tuples. When more processes are used we observe that the algorithm does not scale because the processors are not effectively used and the communication costs increases.

6. FUTURE SCOPE

In order to solve a problem efficiently on a parallel machine it is usually necessary to design an algorithm that specifies multiple operations on each step, i.e., a parallel algorithm. The parallelism in an algorithm can yield improved performance on many different kinds of computers. Some parallel computers cannot efficiently execute all algorithms, even if the algorithms contain a great deal of parallelism. Many experiment and research has shown that it is more difficult to build a general-purpose parallel machine than a general-purpose sequential machine. Our data mining technique contains a collection of tasks having individual algorithms which outperforms parallelism by not allowing any of the components of parallel computer to remain idle. Our technique of one time scan of entire database, then followed by sampling, partitioning, sorting, pattern matching, and merging all done in parallel environment have improved the capabilities of clustering and removed one of the major issue in data mining clustering.

7. CONCLUSION AND FUTUR SCOPE

The above issues are considered major requirements and challenges for the further evolution of data mining technology. Some of the challenges have been addressed in recent data mining research and development, to a certain extent, and are now considered requirements, while others are still at the research stage. The issues, however, continue to stimulate further investigation and improvement. As data is huge and voluminous, we will take parallelism in data mining techniques in order to handle large databases efficiently.

REFERENCES

[1]Anil K. Jain and Richard C. Dubes. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[2] Bernd Mohr Introduction to Parallel Computing. Computational Nanoscience NIC Series, Vol. 31, ISBN 3-00-017350-1, pp. 491-505, 2006.

[3]Clark F. Olson. Parallel algorithms for hierarchical clustering. Technical report, University of California at Berkeley, December 1993.

[4] Devendra Kumar Tiwary ,"Application of Data Mining In Customer Relationship Management (CRM)", Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 3 Number 4 (2010) pp. 527– 540

[5]Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf. Retrieved 2008-12-17.

[6] http:// www. thearling. com/ text/ dmwhite/dmwhite.htm

[7] J. Han and M. Kamber; 2000, "Data Mining: Concepts and Techniques", Morgan Kaufmann.

[8]M.H. Dunham ," http:// engr. smu. edu/~mhd/dmbook/part2. ppt."

[9]Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Pano Vassiliadis. Fundamentals of Data Warehouses. Springer, 1999.

[10] Osmar R. Zaïane: "Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering". http:// www. cs. ualberta. ca/~ zaiane/ courses /cmput690 /slides/ Chapter8 /index.html.

[11] Pavel Berkin , "Survey Of Clustering Data Mining Techniques", 2000

[12] Richard J. Roiger, Michael W. Geatz, 2007, Data Mining A tutorial-based Primer", Pearson Education, New Delhi

[13]Shashikumar G. Totad, Geeta R. B, Chennupati R Prasanna, N Krishna SanthosH , PVGD Prasad Reddy. Scaling Data Mining Algorithms to Large and Distributed Datasets. International Journal of Database Management Systems (IJDMS), Vol.2, No.4, November 2010

[14] U.S. Fayyad, G. Piatetsky Shapiro, P. Smyth, R. Uthurusamy ."Advances in Knowledge Discovery and Data Mining.", AAAI/MIT Press, 1996.

[15]Hinneburg, Keim. Clustering Techniques for Large Data Sets. First publ. in: ACM SIGKDD 1999 Int. Conf. on Knowledge Discovery and Data Mining (KDD'99), San Diego, CA, September, 1999, pp. 141-181

[16] Wang,Aggarwal, C., J. Han, P.S. Yu. 2003. A framework for clustering evolving data streams. In Proc. of the 29th International Conference on Very Large Data Bases, Vol. 29, pp. 81-92.

[17]Guha, S.; Rastogi, R.; Shim, K.; CURE: an efficient clustering algorithm for large databases . 1998 ACM SIGMOD International Conference on Management of Data Seattle, WA, USA 1-4 June 1998 PUBLICATION: SIGMOD Rec. (USA), SIGMOD Record, vol.27, no.2, p. 73-84, 0163-5808 ACM June 1998 .

[18] O'Callaghan, L., N. Mishra, A. Meyerson, S. Guha, R. Motwani. 2002. Streaming-data algorithms for high-quality clustering. In Proc. of the 18th Intl. Conf. on Data Engineering, pp. 685-684.

[19] M. Kaya, R. Alhajj / Fuzzy Sets and Systems 152 (2005) 587–601. Genetic algorithm based framework for mining fuzzy association rules.