A Survey on Parallel Method for Rough Set using MapReduce Technique for Data Mining

Varda C. Dhande[1], B.V. Pawar[2]

[1]M. E (Computer Engg.) II Student, Padmabhushan Vasantdada Patil College of Engineering, Pune University, Pune

[2]Professor, Padmabhushan Vasantdada Patil College of Engineering, Pune University, Pune

Abstract*:* ***In this paper Present survey on Data mining, Data mining using Rough set Theory and Data Mining using parallel method for rough set Approximation with MapReduce Technique. With the development of Information technology data growing at a tremendous rate, so big data mining and knowledge discovery become a new challenge. Rough set theory has been successfully applied in data mining by using MapReduce programming technique. We use the Hadoop MapReduce System as an Implementation platform. The lower and upper approximations are two basic concept of rough set theory. A parallel method is used for the effective computation of approximation and is improving the performance of data mining. With the benefits of MapReduce it makes our approach more ideal for executing large scale data using parallel method.***

Keywords: Data mining, MapReduce, Rough sets, Approximations, Hadoop, HDFS**.**

1. Introduction

Nowadays, with the volume of data growing at unmanageable rate, big data mining and knowledge discovery have become a new challenge. Rough set theory for knowledge discovery has been successfully applied in data mining. Then MapReduce technique has received much attention from both scientific community and industry for its applicability in big data analysis. With the development of information technology, large amount of data are collected from various sensors and devices in different formats. Such data processed by independent or connected applications will usually cross the peta-scale threshold, which would in turn rise the computational requirements. With the fast increase and update of big data in real-life applications, it brings a new challenge to rapidly acquire the useful information with big data mining techniques. For processing the big data, Google developed a software framework called MapReduce to support large distributed data sets on clusters of computers which is effective to analyze large amounts of data. MapReduce has been a popular computing model for cloud computing platforms. Followed by Google's work, many implementations of MapReduce emerged and lots of traditional methods combined with MapReduce have been presented until now.

This paper presents survey on the different fields like data mining, data mining using rough set technique and data mining using MapReduce Technique, for rough set approximation calculation using parallel method. In this Paper section 1 is completely introduction. Section 2 describe about data mining Technique with the related work. Section 3 describe shortly about data mining using rough set theory with its related work. Section 4 is motivation of this paper, which describe parallel method for rough set approximation in data mining with its related work. Section 5 future work on the parallel method for rough set approximation. And lastly section 6 conclude the paper work

2. Data Mining Technique

Data mining is a new developing technology for enterprise data and information integration. It can reduce the operation cost, increase profit, and strengthen market competition of the enterprise. Generally, there are two ways to establish a data mining application tailor to an enterprise: using business intelligence solutions and products available on the market, or developing data mining algorithms all by oneself. However, both of them are impractical in cost and time. The former one costs a lot, while the latter requires developers to be familiar with both enterprise business and data mining technology.

Software reuse is a solution to avoid repeated work in the software development. It is regarded as a approach to solve the software crisis and promote efficiency and quality of software production. As a kennel technique to support software reuse, software component technique gets increasingly wide attention. To fully make use of reusable component, and support mass component's production, classification, search, assembly and maintenance, component library is very important. Applying software component technique to data mining, wrapping individual business modules of data mining in the form of components, and using component technique to achieve the organization, management and retrieval of the components, could greatly increase the efficacy and quality, and decrease the cost and period of data mining application development. The demand of variability of data mining tasks could be met as well, and the application of data mining technology can be broaden. Neelamadhab Padhy, Dr. Pragnyaban Mishra [1] presents a variety of techniques, methods and different areas of the research which are helpful for data mining Technologies. Explain different data mining task, Types of Data Mining System, Data Mining Life Cycle, Data Mining Methods, Data Mining Applications, and The Scope of Data Mining. Nikita Jain, Vishal Srivastava, [2] in this paper the concept of data mining is summarized and its significance towards its methodologies is explained. The data Mining based on

Neural Network and Genetic Algorithm is researched in detail. The different technology to achieve the data mining on Neural Network and Genetic Algorithm are also surveyed. This paper also presents a formal review of the area of rule extraction from ANN and GA .
There are different task in data mining which are
1) Classification
2) Estimation
3) Prediction
4) Grouping or association rule
5) Clustering
6) Description and Visualization.

The first three tasks are all example of directed data mining or supervised learning. The next three tasks are example of undirected data mining.
There different methods for the classification task in data mining and Rough Set Theory is one of the classification method. Rough set theory can be used for classification to discover structural relationships within vague or noisy data.

3. Data Mining Using Rough Set Theory

3.1 Rough set Theory

Rough set theory was developed by Zdzislaw Pawlak in the early 1980's. It is a very powerful mathematical tool for dealing with imprecise information in decision situations. The main goal of the rough set analysis is induction of approximations of concept also it plays an important role in the fields of machine learning, pattern recognition and data mining. Rough set based data analysis uses data tables called a decision table, columns of these tables are labeled by attributes, rows – by objects of interest and entries of the table are attribute values. Attributes of the decision table are divided into two different groups known as condition and decision attributes. Each row of a decision table induces a decision rule, which shows decision or outcome. If some conditions are fulfilled, the decision rule is certain, when a decision rule uniquely identifies decision in terms of conditions. Otherwise the decision rule is uncertain. Decision rules are associated with approximations. Lower approximation refers to certain decision rule of decisions in terms of conditions, whereas boundary region referred by uncertain decision rules of decisions.
A rough set learning algorithm can be used to obtain a set of rules in IF-THEN form, from a decision table. The rough set method provides an effective tool for extracting knowledge from databases. Here creates a knowledge base, classifying objects and attributes within the created decision tables. Then a knowledge discovery process is initiated to remove some undesirable attributes. Finally the data dependency is find out, in the reduced database, to find the minimal subset of attributes called reduct.
Sushmita Mitra, Pabitra Mitra [3] presents a survey on data mining using soft computing. A classification has been provided based on the different soft computing tools and their hybridizations used, the data mining function implemented. The efficiency of the different soft computing methodologies is highlighted. Generally fuzzy sets can use for handling the issues related to understandability of patterns, incomplete or noisy data, mixed media information and human interaction, and can provide estimated solutions faster. Neural networks are nonparametric, and reveal good learning and generalization capabilities in data-rich environments. Rough sets are suitable for handling different types of uncertainty in data.

Silvia Rissino and Germano Lambert-Torres[4] discussed The rough set approach to processing of incomplete data is based on the lower and the upper approximation, and it is defined as a pair of two crisp sets corresponding to approximations. The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information concerning data. Rough set theory has also provided the necessary formalism and ideas for the development of some propositional machine learning systems. Rough set has also been used for knowledge representation; data mining; dealing with imperfect data; reducing knowledge representation and for analyzing attribute dependencies. Rough set Theory has found many applications such as power system security analysis, medical data, finance, voice recognition and image processing; and one of the research areas that has successfully used Rough Set is the knowledge discovery or Data Mining in database.

4. Data mining using parallel method for rough set approximation with MapReduce Technique

To our knowledge, most of the traditional algorithms based on rough sets are the sequential algorithms and existing rough set tools only run on a single computer to deal with large data sets. To expand the application of rough sets in the field of data mining and deal with huge data sets, the parallel computation of the rough set approximations is implemented. And this Parallel approximation can be achieved by using MapReduce Technique.

- **Map-function** takes an input pair and produces a set of key, value pairs. The MapReduce groups together all values associated with the same key I and transforms them to the Reduce function.
- **Reduce-function** accepts key I and a set of values for that key. It merges these values together to form a possibly smaller set of values. By doing sorting and shuffling it produces reduced values get from Map function.

4.1 MapReduce Technique

Hadoop Distributed File System is the storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and allocates them on data nodes, to enable reliable extremely rapid computations. Hadoop consist of two major components which are: File storage and Distributed processing system. The first component of file storage is known as “HDFS (Hadoop distributed file system)”. It provides scalable, dependable, comparatively low cost storage. HDFS stores files across a collection of servers in a cluster. HDFS ensures data availability by continually monitoring the servers in a cluster and the blocks that manage data. The second very important components of Hadoop, is the parallel data processing system called

"MapReduce". The MapReduce framework and the Hadoop distributed file system are running on the same set of nodes. In MapReduce programming, it allows the execution of java code and also uses software written in other languages.

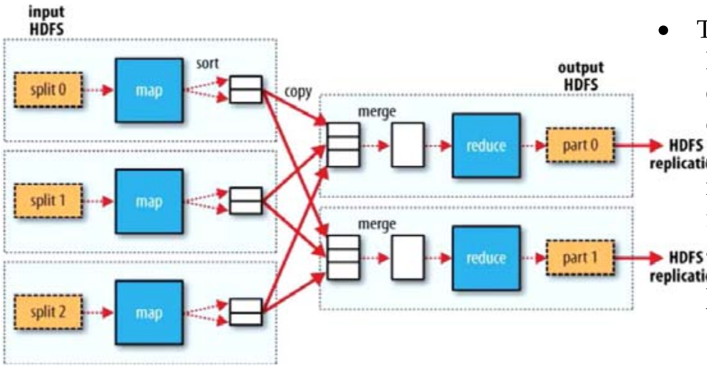

Fig.4.1. The MapReduce programming model

Jeffrey Dean and Sanjay Ghemawat [6] present the brief description about MapReduce programming model, with different programs as example. Also gives execution overview of MapReduce programming model, with Fault Tolerance, Task Granularity and locality. It explains reasons for successful use of MapReduce Programming model at Google. The model hides the details of parallelization, fault-tolerance, and load balancing so it is easy to use. Also a large variety of problems are easily expressible as MapReduce computations. Redundant execution can be used to reduce the impact of slow machines, and to manage machine failures and data loss.

Zdzisław Pawlak and Andrzej Skowron [7] present basic concepts of rough set theory, also listed some research directions and exemplary applications based on the rough set approach. In this paper it mentioned the methodology based on discernibility and Boolean reasoning for efficient computation of different entities including reducts and decision rules. it has been explain that the rough set approach can be used for synthesis and analysis of concept approximations in the distributed environment of intelligent systems.

J Zhang, T Li, Yi pan [8] proposed three rough set based methods for knowledge acquisition using MapReduce technique. To evaluate the performances of the proposed parallel methods used speedup. Comprehensive experimental results on the real and synthetic data sets demonstrated that the proposed methods could effectively process large amount of data sets in data mining. There are three algorithms are used for the knowledge acquisition from big data based on MapRedudce.

Junbo Zhang, Tianrui Li, Da Rusan [9] proposes a parallel method for computing rough set approximations. Accordingly, algorithms corresponding to the parallel method based on the MapReduce technique are put forward to deal with the massive data. Also used speedup, scaleup and sizeup to evaluate the performances of the proposed parallel algorithms. The experimental results on the real and synthetic data sets showed that the proposed parallel algorithms could effectively deal with large data sets in data mining. In this paper there are 4 types of algorithms are using which are Equivalence class computing algorithm, Decision class computing algorithm, association algorithm and Indexes of rough set approximation computing algorithm for parallel method.

5. Future Work

- To implement the Parallel method for rough set using MapReduce technique in data mining there are four types of algorithms are using, which are 1)Equivalence class computing algorithm, 2) Decision class computing algorithm, 3) association algorithm and 4)Indexes of rough set approximation computing algorithm[12]. Here first two algorithms are implementing parallely and third and fourth in series. This is still a time consuming, so to increase time efficiency instead of using four proposed to use three algorithm by merging last two algorithms

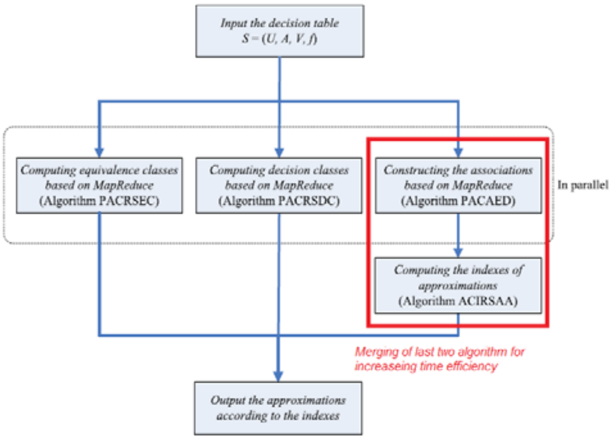
Input the decision table
S = (U, A, V, f)
Computing equivalence classes based on MapReduce (Algorithm PACRSEC)
Computing decision classes based on MapReduce (Algorithm PACRSDC)
Constructing the associations based on MapReduce (Algorithm PACAED)
In parallel
Computing the indexes of approximations (Algorithm ACIRSAA)
Merging of last two algorithm for increaseing time efficiency
Output the approximations according to the indexes

Figure 5.1: The Totally parallel method for computing rough set approximations

- Also to increase time efficiency and speedup the process proposed to used attribute selection option from the massive data, for more accurate result.

6. Conclusion

In this paper the different fields like data mining, data mining using rough set technique and data mining using MapReduce Technique briefly describe, for rough set approximation calculation using parallel method. Also describe the Future work for the current parallel method for calculating rough set approximation.

References

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[2] Nikita Jain, Vishal Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 Nov-2013

[3] Sushmita Mitra, Pabitra Mitra, "Data Mining In Soft Computing Framework: A Survey" IEEE Transactions on Neural Networks, Vol. 13, No. 1, January 2002

[4] Silvia Rissino and Germano Lambert-Torres, "Rough Set Theory – Fundamental Concepts, Principals, Data Extraction, and Applications" Open Access Database www.intechweb.org

[5] Pavel JURKA, "Using Rough Sets In Data Mining**"** Proceedings of the 12th Conference and Competition STUDENT EEICT 2006 Volume 4.

[6] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simple Data Processing on Large Clusters" Proc. To appear in OSDI 2004

[7] Zdzisław Pawlak, Andrzej Skowron "Rudiments of rough sets" Proc. Elsevier accepted in 7 June 2006

[8] J Zhang, T Li, Yi pan "Parallel Rough Set Based Knowledge Acquisition Using MapReduce from Big Data" Proc. ACM Big Mine '12, August 12, 2012 Beijing, China

[9] J Zang, T Li, Da Rausan "A parallel method for rough set approximations" Proc. Elsevier accepted in 11 January 2012

[10] Mert Bal, "Rough Sets Theory as Symbolic Data Mining Method: An Application on Complete Decision Table" Information Science Letters An International Journal Inf. Sci. Lett. 2, 1, 35-47 (2013)

[11] Daniel Zinn, Shawn Bower, Sven Köhler, "Parallelizing XML data-streaming workflows via MapReduce", Journal of Computer and System Sciences 76 (2010) 447–463

[12] Peng Peng, Qianli Ma, Chaoxiong Li, "Research and Implementation of Data Mining Component Library System" http://www.cse.ust.hk/~ppeng/papers/WCSE09 (WCSE2009)