Parallel Rough set approximation using map-reduce technique in Hadoop

Ankita Dubewar[1]

[1]Bharat Ratna Indira Gandhi College of Engineering, Solapur University, Solapur,
India
ankitabachuwar@gmail.com

Abstract: Rough set theory can be considered as a tool to study the uncertain, indescendent data by classifying the set into ternary classification as lower, upper and boundary region. This paper helps explaining the rough set theory using map-reduce in Hadoop. More over using some features we can reduce the number of input dimension, so that the process can be executed in comparative less time. Using this method we can help many organization that are facing problem related to the data handling and data analysis.

Keywords: Hadoop, Hadoop Distributed File System (HDFS), Dynamic Data Placement(DDP)**.**

1. Introduction

Handling the huge amount of data has been the serious problem now days and with this massive data comes the application associated with it such as association rule mining, sequential pattern mining, text mining and temporal mining. But, with the emergence of hadoop implementation into the map-reduce technique we can solve the problem of the many organization that they are facing with it.

2. Introduction to rough set

Rough set theory was proposed by Zdzislaw pawlak in1982. This method is considered as one of the non-static approach in data analysis and deals with the uncertain, indescendent and incomplete data[1]. It is basically a ternary classification by approximating lower, upper and boundary region based on the equivalence relation in the universe.

Finding rough approximations comes in following steps.

- Finding decision set
- Equivalence class
- Association class
- Lower, Upper and Boundary approximations

Consider a set X is a subset of U. We want to characterize the set with respect to Y using rough set theory then,

Lower approximations: It is a set of items which can be definitely classified as X.

Upper approximation: It is define as a set of item which can be possibly classified as item of X**.**

Boundary Approximation: It is a set of items which can be classifies either as items of X or may not be.

Consider the following example for the better understanding of data. In the following example, first column refers to the objects while the next columns such as 'Income', 'Student 'and 'Credit Rating' are the atttributes while the last column refers to the decision of the attributes.

Table 1: An example of data set(Decision table S)

Persons	Income	Student	Credit rating	Buys-computer
X1	High	No	Fair	No
X2	High	No	Excellent	No
X3	High	No	Fair	Yes
X4	Medium	No	Fair	Yes
X5	Low	Yes	Fair	Yes
X6	Low	Yes	Excellent	No
X7	Low	Yes	Excellent	Yes
X8	Medium	No	Fair	No
X9	Low	Yes	Fair	Yes
X10	Medium	Yes	Fair	Yes
X11	Medium	Yes	Excellent	Yes
X12	Medium	No	Excellent	Yes
X13	High	Yes	Fair	Yes
X14	Medium	No	Excellent	No

In the above table we can see person X3 and X3 have same attributes still the results are different.we call this type of data as indescendent or inconsistent data. In such situations rough set proves very useful .

Decision class Yes=[X3,X4,X5,X7,X9,X10,X11,X12,X13]
Decision class No=[X1, X2, X6, X8, X14]
quivalence class: Eq1=[X1, X3] Eq2=[X6, X7] Eq3=[X4, X8]
Eq4=[X5, X9] Eq5=[X6, X7] Eq6=[X10]
Eq7=[X11] Eq8=[X12, X14]

Association class :We compare each of the equivalence class with the decision class.

Approximations:
Lower approximation = [{X5,X9},{X10},{X11}]
Upper Approximation=[{X2}]
Boundary region=[{X2},{X6,X7},{X12,X14},{X1,X3},{X4,X8}]

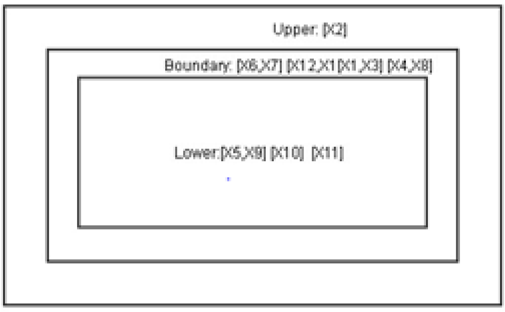
Upper: [X2]
Boundary: [X6,X7] [X12,X1[X1,X3] [X4,X8]
Lower:[X5,X9] [X10] [X11]

Figure 1: Representation of the Approximations

The regularity hidden in the data are generally expressed in terms of the rules. which are the basic tool of data mining.Rule are expressed of the form,If

(attribute1,valur1) and (attribute2,valur2) and …..

Then (decision value)

The number of consistant rule in the table is known as factor of consistence which is denoted by $_{\gamma}$ (C,D)=1. Here, C is the condition and D is the decision, if $_{\gamma}$(C,D)≠1 then the table is not consistent.

3. HDFS and MapReduce

Hadoop mainly consist of two major components as File storage and Distributing processing system. The first component is called 'Hadoop Distributed File Sys-tem'(HDFS) .It is the primary storage use by Hadoop which provides scalable reliable and low cost data storage. Each node in a Hadoop instance generally has a single namenode, and a cluster of datanodes form the HDFS cluster. It stores file across collection of clusters and make it available by continuous monitoring.

The second component is the parallel data processing called 'MapReduce'. It is a linearly scalable programming model. It allows the execution of different languages such as (java, C++, python, Pearl).It works by breaking the processing into two different stage as 'map' and 'reduce', each of the stage has Key-Value pair .These two stages are known as Map function and Reduce function.

Both of the systems HDFS and MapReduce run on the same set of node. These are both open source projects, inspired by technologies created inside Google.

As it can be seen from the following figure that the input is divided into the number of blocks using input split () function. Each of this block is of size 64MB, Data is process in parallel and collected at the reduce function. Before that the data is reduce it is process, using sort () and merge () function.

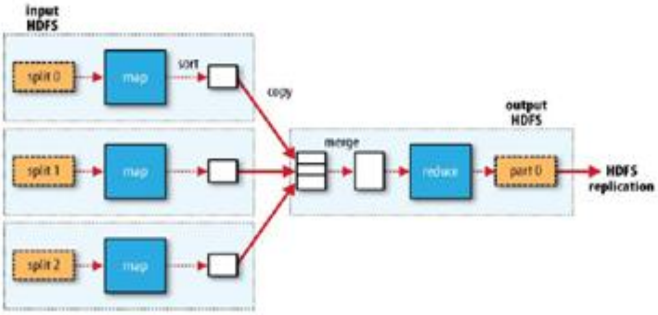

Figure 2: Magnetization MapReduce data flow with single reduce

4. Advantages Of Implementing RoughSet in Hadoop

Design for huge data: Hadoop is mainly design to han-dle and process a large amount of data systematically and using parallel method.

1.Reliability : It achieves reliability by replicating the data across multiple hosts, and hence theoretically does not require RAID storage on hosts .Replication is by default set to 3.
2.Security: Hadoop is implemented in Java while java is known for its security mechanism[4]. In which data is mostly centered around the sand boxing rather than for writing secure services.
3.Streaming pattern: It can handle lots of streaming reads and infrequent writes.
4.Not fully POSIX: HDFS is not completely POSIX (Porta-ble Operating System Interface)-compliant, because the requirements for a POSIX file-system differ from the target goals for a Hadoop application.
5.Increased Bock size: The increase in the block size (64MB) results in handling reduced numbers of blocks and fast processing.
6.Parallel processing: The rough set that we are using is parallel in nature and again parallel processing of it will reduce time processing at much extend

5. Using RoughtSet in Paallel Processing

The general method of rough set is serial i.e. (calculating decision class, equivalence class, association class in serial manner) but calculating equivalence class and decision class in parallel is the intensive calculating using least no of algorithms[10].

For calculating Rough set in MapReduce we divide the decision table into m- decision tables, we call this table as sub decision table of S. As well as, two equivalence class with the same information can be combine to one equivalence class, this way we will have reduce number of equivalence class. The sub decision table can compute with the equivalence class independently. This way, the rough set equivalence class of decision table S can be executed in parallel.

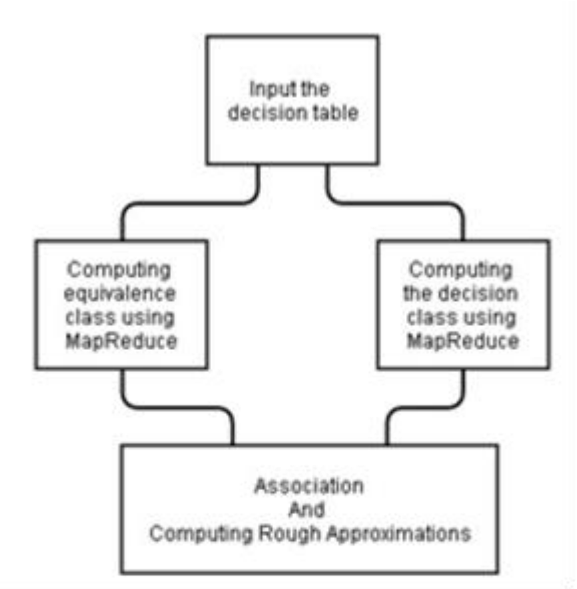
Input the decision table
Computing equivalence class using MapReduce
Computing the decision class using MapReduce
Association And Computing Rough Approximations

Figure 3: Flow Diagram of calculating parallel method of rough set.

5.1 Using Static and Dynamic Data

As the name implies the static data is the one that remains same while the dynamic data represents energetic, capable of action or change.

For data analysis the static data set is very comfortable to use since the values remains the same but the dynamic data which keeps on changing the value make the study complicated. Although there are several strategies that will help the study.

Using any of the type in a homogeneous cluster, the Hadoop strategy can make full use of the resources of each node. However, in a heterogeneous cluster, the computing capacity for each node is not the same. Moreover, for different types of job, the computing capacity ratio of nodes are also not the same. Therefore, a Dynamic Data Placement (DDP) strategy is presented according to the types of jobs for adjusting the distribution of data blocks[7]. The DDP, consists of two main phases:

- When the input data are written into the HDFS.
- When a job is processed

6. Applications of RoughSet

6.1 In Multimedia Data classification

Multimedia is defined as combination of more than one media; they may be two types, static

and dynamic media. Text, graphics and images are categorized as static media, while objects like animation, music, audio, speech, and video are categorized as dynamic media.

Multimedia database management system (MDMS) is developing in purpose to fulfill this requirement. MDMS supports facilities for the indexing, storage, retrieval process and provides a suitable environment for using and managing multimedia data[8].

6.2 Feature selection Approach

Feature selection plays an important role in the data mining. The general goal of feature selection is to reduce the cardinality of the subset as well to increase the information contents in the selected subset[9]. Using such feature can be very benificial for the analysis using specific conditions as well as reduction in the analysis efforts causing less time for the analysis.

7. Limitations

Hadoop is not meant for storing large number of small files.

- HDFS is not suitable to efficiently access small files: it is primarily designed for streaming access of large files. Reading through small files normally causes lots of seeks and lots of hopping from datanode to datanode to retrieve each small file, which is an inefficient data access pattern.
- Every file, directory and block in HDFS is represented as an object in the namenode's memory, each of which occupies 150 bytes. The block size is 64 Mb. So even if the file is of 20 kb, it would be allocated an entire block of 64 Mb. That's a waste disk space.

8. CONCLUSION

The implementation of this theory using Hadoop, for the calculation of Rough set will be very helpful from future perspective since the machine data will increase day by day and need to analyze, on one platform. There are many things that are needed to be focus, specially feature selection techniques as well as using dynamic or heterogeneous data for calculating rough set in hadoop.

References

[1] Ing.Pavel Jurka, Using rough set in data mining.

[2] Z. Pawlak,W.Ziarko, Communication of ACM,38,1995,p.88

[3] Zdzislaw.pawlak, Rough set theory and its applications,Journals of telecommunication and information technology.

[4] Devaraj Das, Owen O'Melly,Sanjay Radia,Kan Zang,Adding security to apache hadoop, Hortonworks Technical Re-port,www.Hortonworks.com

[5] Junbo Zhang,Tianrui Li, Da Ruan, Zizhe Gao, A

Parallel method of rough set Approximations,Elsevier,Information Science 194(2012)209-223

[6] Silvia Rissino,Germano Lambert-Torres,Rough set theo-ry,fundamental concepts, Principal, Data Extraction And Applications

[7] Chia-Wei Lee,Kuang-Yu Hsieh, Sun-Yuan Hsied,A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments, Big Data Research Volume 1, August 2014, Pages 14–22

[8] M.Nordin .A, Rahman, Yuzarimi MLazim, Farham Mo-hamed,Rules Generation for Multimedia Data Classifyingusing Rough Sets Theory International Journal of Hybrid Information TechnologyVol.6, No.5

[9] Prerna mahajan,Rekha khandwal,Rough Set approach in machine learning,volume56-no.10,October. 2012

[10] A.Pradeepa, DR.Antony Selvadoss Thanamani, Hadoop file system and fundamental concepts of MapReduce Interior and closure Roughset Approximations,vol.2, issue 10, October 2013

[11] Tom white, Hadoop: The Definitive Guide,Second edition,2011