Obtaining Rough Set Approximation using MapReduce Technique in Data Mining

Varda Dhande[1], Dr. B. K. Sarkar[2]

[1]*M.E II yr student, Dept of Computer Engg, P.V.P.I.T Collage of Engineering Pune, Maharashtra, India*
[2]*Professor & HOD, Dept of Computer Engg, P.V.P.I.T Collage of Engineering Pune, Maharashtra, India*

Abstract—Nowadays data growing at tremendous rate, which creates a many challenges in immense data mining and knowledge discovery. Rough set theory is one of the tool for data mining in which lower and upper approximations are basic concepts. By using Rough Set theory with Map-Reduce technique parallelization of approximation is possible. This paper proposes a parallel method for rough set approximation for handling unstructured data, with the help of Hadoop technology. Related algorithms for parallel method based on map-reduce technique are put forward to deal with unstructured data. The time efficiency in calculating Rough Set approximation using parallel method can be increased by using attribute selection option, for avoiding redundant attributes which are responsible for time consummation and for degrading quality of decision rules, which will discovered from lower and upper rough set approximation.

***IndexTerms*—Data Mining, Map-Reduce, Rough Set, Approximations, Hadoop, HDFS**

I. INTRODUCTION

In recent years data is growing at tremendous rate day by day and therefore big data mining is becoming a new challenge. Data mining is a process or technology which uses a variety of data analysis tools to retrieve knowledgeable information or data, also to discover patterns and relationships in data that may be used to make valid predictions [6]. There are many different types of tools are available for data mining. Rough set theory is one of the data mining tool use for data analysis [9]. Rough set theory was developed in 1980s by ZdzislawPawlak [5]. It is a very powerful mathematical tool, which deals with incomplete information while decision taking situation. This rough set theory has applications in many fields, such as decision support, engineering, banking, medicine, machine learning, pattern recognition and data mining [4]. The rough set theory is associated with lower and upper approximations of rough set [7]. The lower approximation contains of all objects which are surely belongs to the precise set and upper approximation consist of all objects which possibly belong to the precise set. Difference between upper and lower approximation shows boundary region. These upper and lower approximation sets are useful for discovering quality rules or decisions.

The traditional way of calculating rough set approximation is serial processing, but now a parallel method for rough set calculation is available, which is time efficient. To implement this parallel method Map-Reduce technique is used. Map-Reduce is framework or a programming model which used to handle a HDFS (Hadoop distributed file system) [10]. It is a second major component of the Hadoop system and it is a parallel data processing system. The model is consists of mainly two functions: Map and Reduce. Here Hadoop is used to build the parallel rough set approximation system. Hadoop is an open source framework used for storing and processing big data in a distributed fashion on multicluster. This is also useful to handle unstructured data.

II. RELATED WORK

ZdzisawPawlak and AndrzejSkowron [1] presented 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.

J Zhang, T Li, Yi pan [2] 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.

Jeffrey Dean and Sanjay Ghemawat [3] shows and explains reasons for successful use of Map-reduce Programming model at Google. Such that the model is easy to use because it hides the details of parallelization, fault-tolerance, locality optimization, and load balancing also a large variety of problems are easily expressible as MapReduce computations.

This paper has presented novel approaches for exploiting data parallelism for efficient execution of XML-based processing pipelines [8]. MAPREDUCE is a parallel distributed programming framework introduced in [11], which can process huge amounts of data in a massively parallel way using simple commodity machines.

III. SYSTEM ARCHITECTURE

Architecture of the system is as shown in the following figure.1 This system uses and process structured, semi-structured and unstructured data for calculating rough set approximation. These structured and unstructured data can be load in the system from data source like websites for example UCI machine learning repository, which is a source of data sets.

After loading high dimensional data for example any medical data set, data passes through preprocessing. Preprocessing is nothing but simply removing the missing values or performing some operation to remove noise and outlier. After preprocessing data passes through tool but before that system providing a attribute selection option. This option facilitate a user with facility of attribute

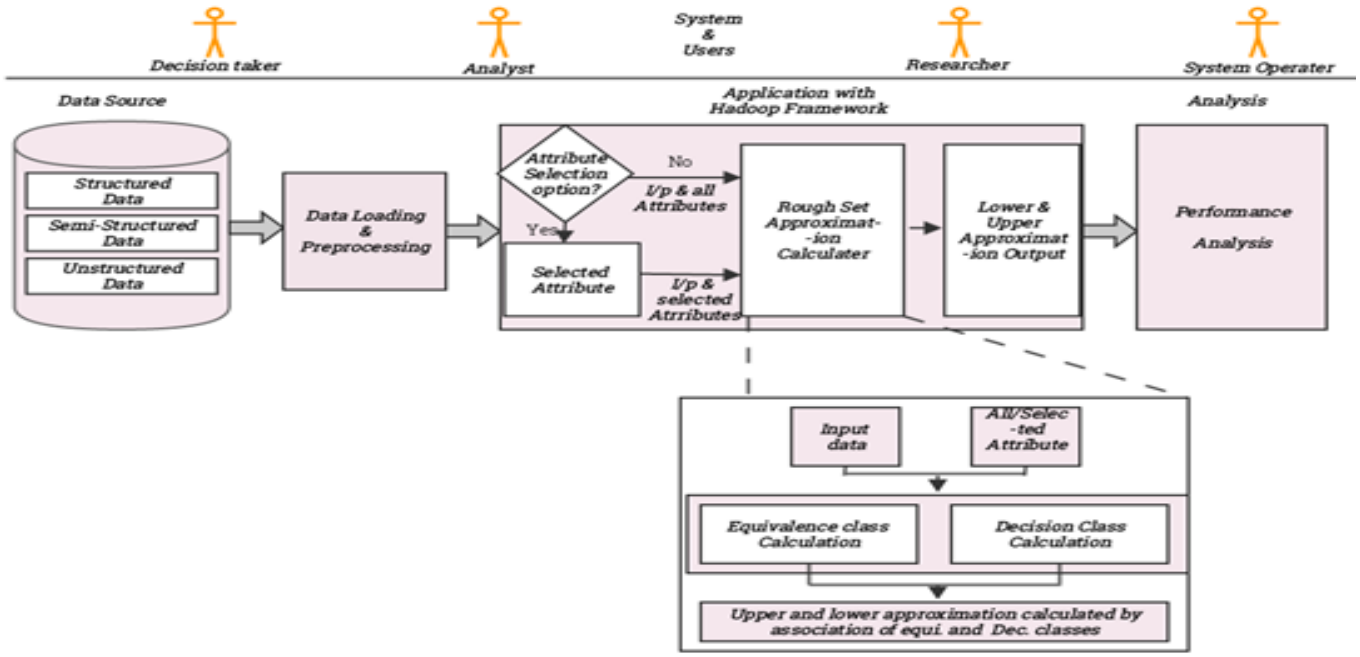

Fig.1- System Architecture

selection from a large data sets, due to this option redundant attributes are get avoided during calculation which are not necessary for calculating rough set approximation. Because of this not only time complexity for calculating approximation increases of large data sets but also the quality of discover decision will get increase. If user select attribute selection option rough set approximation calculation done on selected attributes otherwise calculation will perform on all attributes in data set.

In rough set approximation calculator following algorithms are used for lower and upper approximation calculation.

1) Algorithm for Rough Set Equivalence Classes computation, this is done by using Map and Reduce algorithms
2) Algorithm for Rough Set Decision Classes computation, this done by using Map and Reduce algorithms
3) Lastly Algorithm AI-Algorithm which is the merging of two algorithms from existing system, which are Association algorithm and indexing algorithm, in this by doing association of equivalence class and decision class with indexing process approximations are get calculate.

Here performance of system will increase by decreasing time consumption due to merging of two algorithms in new AI-algorithm. So the prediction graph of comparison of this is as follows which takes place using following example.1

Example: Consider the example as shown in table 1.

Table-1: Decision Table

	Age	LEMS	Walk
X1	16-30	50	Yes
X2	16-30	0	No
X3	31-45	1-25	No
X4	31-45	1-25	Yes
X5	46-60	26-49	No
X6	16-30	26-49	Yes

X7	**46-60**	**26-49**	**No**

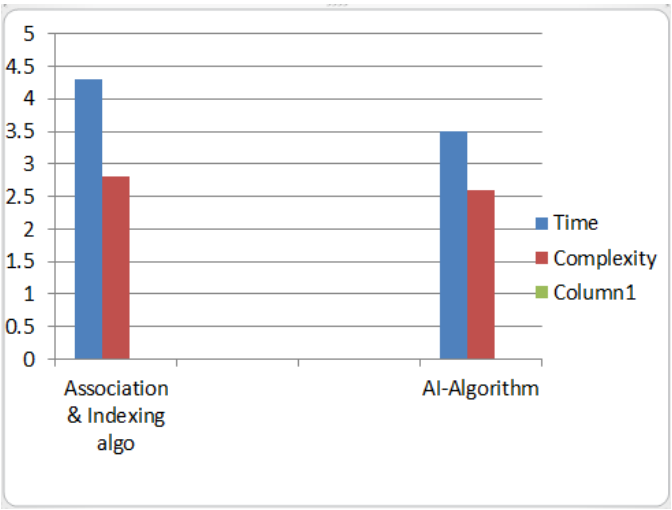

Chart-1 Comparison of Time and Complexity

There are different types of users which can use this system which are Analyst, Decision Maker, Researchers and System Operator.

IV. INPUT-OUTPUT PROCESS

In our proposed system we are going to calculate the rough approximation using structure/ unstructured data, with a attribute selection option for more accurate and better performance. The existing system uses four algorithms to generate equivalent class, decision class, association and indexes to compute the rough approximation. Existing method requires more computation. we are going to attempt merging of association and indexes algorithm to get better performance. doing this we get a new algorithm called AI-algorithm.

Hence algorithm for our proposed system

Input: Decision table (data set)

Output: Upper approximation, Lower Approximation, Boundary region

Method:

Step 1:
Choose the data set (Structure/ unstructured)

Step 2:
If data set is unstructured and without objects, it get converted in structure and objects get allotted.

Step 3:
By using attribute selection option select the conditional attribute if needed.

Step 4:
Put selected input to Hdfs

Step 5:
Computing of rough set approximation

i) Compare the objects having same condition but different decision attribute value. This calculate equivalence class

ii) Every comparison, compares objects equal to decision domain. This calculate decision class.

iii) In comparison if we found for more than one position contains certain value/s, add that objects into boundary region of corresponding decision.

iv) If we found only one value and other positions are null, then add that object into lower approximation of corresponding decision.

v) Upper approximation is union of lower approximation and boundary region.

.

V. SYSTEM RESULT

Our system includes different phases initially the input which can structured/ unstructured is converted in proper input form, from the given dataset. As we have taken preliminary dataset of some record consisting some entries with decision column with referring table 1 so that such data can be used to check the efficiency of our system.

So, the result of given example in table.1 is

- Boundary Region contains (x3,x4)
- Upper approximation contains (x2,x5,x7)
- Lower approximation contains (x1,x6)

Similarly preliminary results of our system has been generated as shown above. The datasets will be changed for the further work so that more efficient result can be obtained. With the completion of our work the output for the following system is as shown in Fig-2. It shows over all outcome from the reducer of MapReduce will be generated as for the upper approximation, boundary region, lower approximation [4].

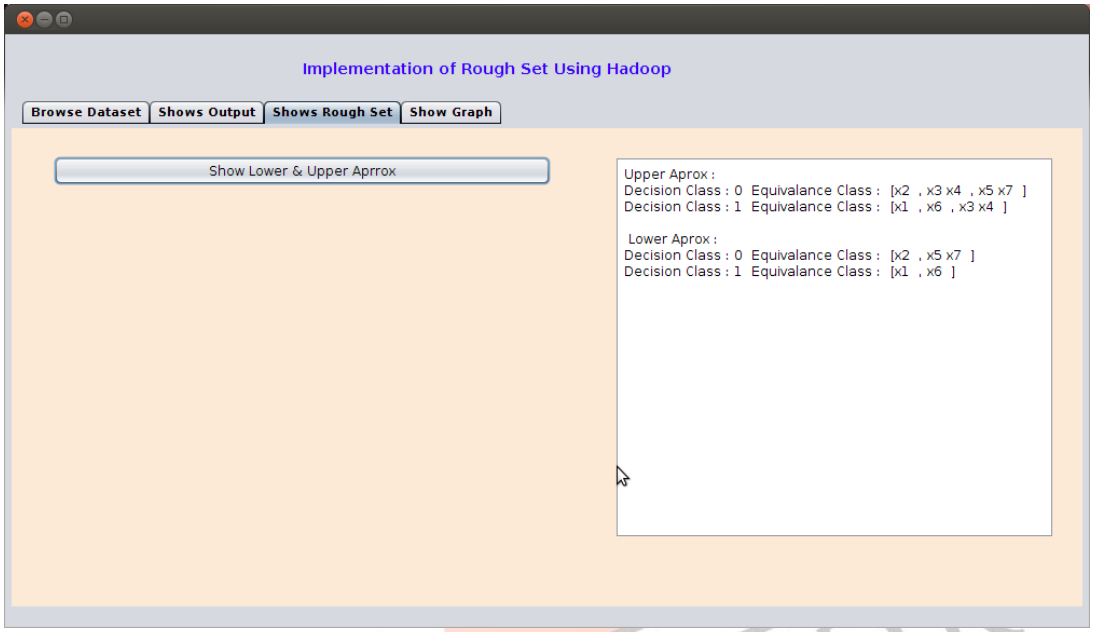

Fig-2: Final Result of approximation

VI. CONCLUSION

In this paper the basic concept of roughest and data mining is been discussed. With the previous system there are some of approaches for roughest approximation. With the proposed system we have focus on the roughest with MapReduce using structured/unstructured data with attribute selection option for more efficient result. Also time is get save by merging two algorithm association and indexing in AI-algorithm So with the proposed system it will be time saving to obtain the lower and the upper approximation.

REFERENCES

[1] ZdzisławPawlak, AndrzejSkowron "Rudiments of rough sets" Proc. Elsevier accepted in 7 June 2006

[2] 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

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

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

[5] MertBal, "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)

[6] 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

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

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

[9] NeelamadhabPadhy, 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

[10] A.Pradeepa1, Dr. Antony SelvadossThanamani "Hadoop File System And Fundamental Concept Of Mapreduce Interior And Closure Rough Set Approximations" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013

[11] PrachiPatil, "Data Mining with Rough Set Using MapReduce" International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) ISSN(Online) : 2320-9801 Vol. 2, Issue 11, November 2014.

[12] M.Young,TheTechnicalWriter'sHandbook.MillValley,CA:UniversityScience,1989.