Rough Approximation Methods: Inter-Table and Data Cube

[1] Prachi M. Patil, [2] Anilkumar Kadam, [3] Rohitkumar Kawhale
[1] Student ME(Computer), [2] Prof. ME(Computer), [3] Student ME(Computer)
[1] Department of Computer Engineering
[1] AISSMS College of Engineering, Pune, India.

***Abstract* – Data mining as an important contribution to data analysis, data discovery and autonomous decision making. Rough set theory (RST) is the technique in data mining is an approach for decision rule extraction from data. Lower approximation, Upper approximation and boundary region are the principle parts of RST. There are two different methods to obtain rough approximation based on data cube and inter-table comparison. In data cube method, data has put in multidimensional way and accessed via map reduce. Another technique also based on map reduce, but it divides the given dataset into number of sub tables. In this paper, we are going to analyze these two different method of computing rough approximation.**

***Index Terms* – Rough Set, Lower approximation, Upper Approximation, Data Cube**

I. INTRODUCTION

Data mining technique is becoming important day by day as the size of digital data in growing. This field is used to extract knowledge from large amount of data. It discovers required knowledge from large amount of data [1]. The data is either stored in data warehouse, database, or other information repositories. Massive data mining is a big challenge, where number of techniques are used to achieve knowledge from raw data. Some of them are fuzzy set, neural network, bays theorem and rough set. In case of rough set theory, the terms lower approximation, upper approximation and the boundary region are very basic and most vital. There are number of ways to calculate these rough approximations [3]. There are number of fields where rough set is used in wide way like medical, engineering, banking, intrusion detection, pattern recognition, quality analysis, artificial intelligence etc. Hadoop is a java framework which is used to store and process large amount of data on commodity hardware. We are able to deal with massive data.

Data mining is a nontrivial process to determine valid, easily understandable dependencies in data. As the information technology field is developing, data volumes processed by many applications crossed the threshold in peta-scale [2], as a result it will in turn increase the computational requirements. Data processing and knowledge discovery [2] for colossal data is ever a burning topic in data mining. The big problem in data mining is the deficiency and indeterminateness. Such type of problems solved by using new procedures and theories, e.g. genetic algorithms, fuzzy sets, or rough sets etc.

This paper describes the two parallel of computing rough approximation i.e. rough approximation using data cube and rough approximation based on inter- table comparison. And also compares these two parallel methods. By representing given dataset using data cube, rough approximation obtained in easier way. Only locations of data cube need to be compared for different decisions. It reduces our task, as number of comparison are not more than domain of decision attribute. In another method, we can calculate rough approximation independent decision class and equivalence class. This can be achieved by dividing input dataset, so that number of comparisons get reduced. The division of dataset based on decision attribute in the dataset. These are new methods for computing rough set approximation. Using map-reduce it can deal with massive data and able to compute rough approximation for massive dataset.

II. RELATED WORK

Rough set theory is the most popular method in data mining. It is used to achieve rule generation from different datasets. Z.pawalak [9] [8] invented rough set technique, describes about the importance of rule generation. Map Reduce terminology has been included in our research work so with referral to that many papers had describe about functionality of Hadoop file system [5]. So we have focused on map reduce terminology in below section of our paper. Initially some serial methods have been designed to achieve rough set approximation [6]. A massive data mining and knowledge discovery introduce a huge dispute with the growing data at an unpredicted rate [1]. Map Reduce is used to process big data at commodity hardware. It manages many large-scale computation [1] [12].

Algorithms corresponding to the parallel method based on the Map-Reduce technique are put forward to deal with the massive data. A. Pradeepa [8], explained the purpose of data mining for big data, computing modes in parallel and algorithms are typical methods in research fields. Then the comprehensive result to evaluate the performances on the massive data sets show that demonstrated technology can effectively process of big data [3]. The mathematical principles of rough sets theory are explained and a sample application about rule discovery from a decision table by using different algorithms in rough sets theory is presented. In the document author described basic concepts of rough set and its advantage [4].

Rough set is a classifier [5] which has great importance in cognitive science and artificial intelligence, especially in machine learning, decision analysis, expert systems and inductive reasoning. It is also used to predict the malignancy degree of brain glioma [11]. The effective computation of approximation is essential improving the performance of data mining and other related task [9]. MapReduce has been implemented in manage many large-scale computation. The recently introduced MapReduce technique has received much consideration from both scientific community and industry for its applicability in big data analysis [3] [2] [10]. The research works have been carried on performing the cube computation, cube aggregation using the MR framework. Nandi et al. [7] [5] developed a scheme to handle special holistic measures.

.

III. ROUGH SET THEORY

Z. Pawlak has created a mathematical tool, rough set theory in the beginning of the 1980s. It has been applied widely to extract knowledge from database [1]. It discovers hidden patterns in data. In decision making, rough set methods have a powerful essence in dealing with uncertainties. Rough sets can be used separately or combined with other methods such as statistic methods, fuzzy sets, genetic algorithms etc. The Rough set theory (RST) has been applied in several fields including data mining, pattern recognition, knowledge discovery, medical informatics, image processing etc.

In RST [7] [9], inconsistencies are not corrected or aggregated. In spite of this lower and upper approximations of all decision concepts are computed and using those rules get generated. Approximation perform vital role in performance of rough set theory. Because the rules are categorized into certain and approximate (possible) rules depending on the lower and upper approximations. Basically rough set is depend on approximation i.e. upper approximation and lower approximation and boundary region as mentioned below which is calculated later in this paper. Approximations are fundamental concepts of rough set theory.

 Lower approximation – The lower approximation consists of all the objects without any ambiguity based on attributes. These are surely belong to the set. In another way, we can say that, objects in lower approximation have only one decision for corresponding condition attribute value.

 Upper approximation – The objects are probably belong to the set, cannot be described as not belonging to the set based on the knowledge of the attributes. It contains all objects which possibly belong to the set. In another way, we can say that, objects in upper approximation are union of lower approximation and boundary region of corresponding decision value.

 Boundary region – Boundary region consist of the all objects having same condition attribute value but different decision value. In this set, we can get one than one decision value for same condition attribute value.

Pawlak suggested two numerical measures of imprecision of rough set approximations as,
_ Accuracy measure
_ Roughness measure
Accuracy measure is the ratio of the lower approximation of decision to the upper approximation of corresponding decision. Mathematically is can be defined as,
Where,

$$\alpha_D = \text{lower}_D(X) / \text{upper}_D(X). \quad (1)$$

X is dataset,
$X \neq \emptyset$; and
|.| denotes the cardinality of set.
$0 <= \alpha_D <= 1$

Based on the accuracy measure, the roughness measure is defined as,

$$\alpha D(X) = 1 - \alpha D(X) \quad (2)$$

If boundary region of any set is empty, then the set is called crisp set [16]. It means that all the objects in set has unique decision value for every condition. If the boundary region of set is nonempty, then it is called as rough set. In this set, for same condition value we can have more than one decision value. Rough set deals with vagueness and uncertainty which is most important in decision making. Data mining is a field that has an important contribution to data analysis, discovery of new meaningful knowledge, and independent decision making [15]. The rough set theory offers a feasible approach for decision rule extraction from data. Rough set theory (RST) employed mathematical modeling to deal with class data classification problems, and then proved to be a very useful tool for decision support systems, particularly when hybrid data, vague concepts and uncertain data were involved in the decision process [6].

IV. APPROXIMATION USING DATA CUBE

In this method, rough approximation get calculated using data cube. Basically data cube will be used to represent the dataset. Data cube are an easy way to look at the data. It is used to represent data along some measure of interest. Author has merged two concepts i.e. rough set and data cube to get better performance. In this method, we need to determine the dimensions of the dataset. No of dimension of data cube is always equal to no of attributes. There are two types of attributes in dataset i.e. conditions attribute and decision attribute. There can be more than one condition attribute but number of decision attribute is always equal to 1.

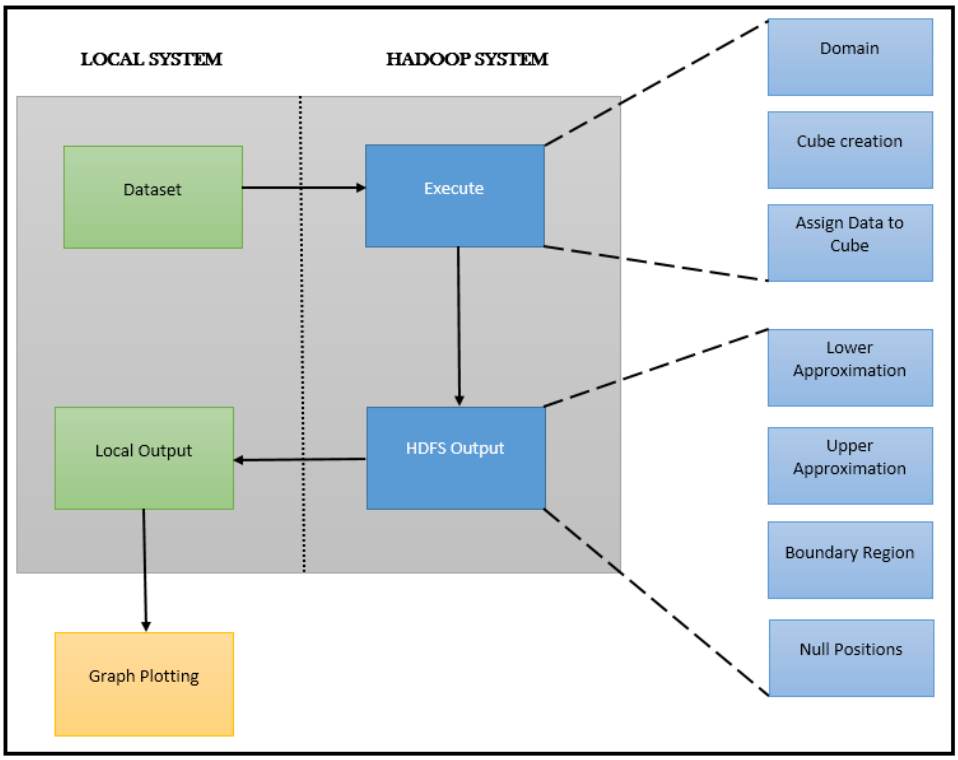

Figure 1: System architecture of data cube method

As shown in figure 1, dataset need to choose from local system. As this system uses map reduce for parallel computing, the dataset need to move from local system to Hadoop system. In Hadoop approximation get calculated and gives output in the form of lower approximation, boundary region and null values.

Initially, the data cube and assign it to NULL i.e. Values of all positions (indexes) of data cube are initialize to NULL. Number of positions of data cube are same as multiplication of value of each dimension of data cube or multiplication of domain of condition and decision attribute.

To compute rough approximation we need to follow the steps as:

i) Compare the objects having same condition but different decision attribute value.

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

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.

Basically data cube will be used to represent the dataset. Data cube are an easy way to look at the data. It is used to represent data along some measure of interest. Here, rough set and data cube merged to get better performance.

V. APPROXIMATION USING INTER-TABLE COMPARISON

In this system we can compute rough set approximation without computing equivalence classes, decision classes and associations between them using the Map-Reduce technique as existing system. Lower and upper approximations are computed by comparing sub tables generated by input table. The system architecture is as shown in figure 2. The most important step in the system is to divide dataset on the basis of decisions. Hence number of sub tables is equal to the domain of decision attribute from the dataset. Here, number comparisons get reduced. So this is the key step in the system. One more advantage is we can get boundary region independent on lower approximation and upper approximation, dataset (decision table) need to divide based on decision attribute domain. Number of sub tables generated are always equal to the domain of decision attribute in dataset.

To compute approximation based on inter table comparison method, we need follow steps as (compare inter table objects) :

i) If value of object in one table matches with value of objects in another table, add both objects in boundary region of corresponding decision.

ii) If value of object in one table does not matches with any value of objects in another tables, add that object in lower approximation of corresponding decision.

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

In the system user has choice to select desired condition attributes. The selected input dataset need to move from local machine to HDFS. So there program execution takes place.

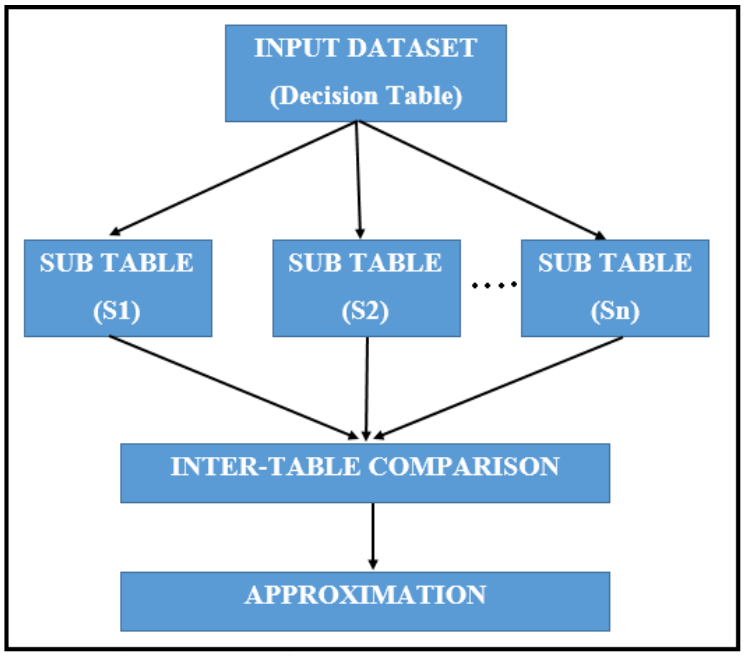

Figure: System architecture of inter-table comparison method

VI. Conclusion

In this paper, we have described two different methods to compute rough set approximation. The basic concept of roughest and data mining is been discussed. In the data cube method, system can retrieve data easily from the data cube. So with this method it will be better to obtain the lower and the upper approximation. Such that, the new algorithm is been designed to work with approximation in one of different way. This algorithm enhances the knowledge from the decision table to the data cube for finding the approximation for the roughest.

Many rough sets based algorithms have been developed for data mining. But enlarged data in applications made these algorithms based on RST a challenging task. Computation of rough set approximation is very important step. We can improve the quality and speed of calculating approximation. This is one way where we have lots of opportunities to achieve speed and accuracy. Another method i.e. based on inter-table comparison, is a parallel method for computing rough set approximation in more efficient way.

References

[1] Rohitkumar Kawhale, Sarita Patil, "Obtaining Approximation with Data Cube using Map-Reduce," International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 3 Issue: 7.

[2] Prachi Patil, "Data Mining with Rough Set Using Map-Reduce", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2014.

[3] Rohitkumar Kawhale, Sarita Patil, "Data Cube Materialization Using Map Reduce", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2014.

[4] Prachi Patil, Anilkumar Kadam, "Obtaining Approximation using Map Reduce by Comparing Inter-Tables", International Journal on Recent and Innovation Trends in Computing and Communication.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] Arnab Nandi, Cong Yu, Philip Bohannon, and Raghu Ramakrishnan," Data Cube Materialization and Mining over MapReduce", IEEE transactions on knowledge and data engineering, vol. 24, no. 10, October 2012.

[7] Junbo Zhang, Tianrui Li, Da Ruan, Zizhe Gao, Chengbing Zhao "A parallel method for computing rough set approximations" Information Sciences 194 (2012) 209–223. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] A.Pradeepa, Dr. Antony SelvadossThanamaniLee, "hadoop file system and Fundamental concept of Mapreduce interior and closure Rough set approximations", International Journal of Advanced Research in Computer and Communication EngineeringVol. 2, Issue 10, October 2013.

[9] Zdzisaw Pawlak, Andrzej Skowron "Accelerator for attribute reduction in rough set theory Rudiments of rough sets", Information Sciences 177 (2007) 327.

[10] Zdzisaw Pawlak, Andrzej Skowron,"Rough sets: Some extensions", Information Sciences 177 (2007) 2840.

[11] Zdzisaw Pawlak, Andrzej Skowron, "Rough sets and Boolean Reasoning", Information Sciences 177 (2007) 4173.

[12] Mehran Riki , Hassan Rezaei "Introduction of Rough Sets Theory and Application in Data Analysis" Journal of Mathematics and Computer Science 9 (2014), 25-32