IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

An Efficient Parallel Fuzzy Rough set Based Rule Generation Using Map Reduce

N. Sukanya[1], B. Madusudhanan[2]

PG Scholar, Dept of Computer Science and Engineering, United Institute of Technology, Coimbatore, TN, India[1]

Asst. Professor, Dept of Computer Science and Engineering, United Institute of Technology, Coimbatore, TN, India[2]

Abstract: Massive background knowledge detection shows a challenge with the information capacity growing at an unprecented speed. MapReduce has succeeded large computation. The latterly introduced MapReduce method has more consideration from industry for its wide ranging analysis. Rough set concept is a plow which is used to obtain information from completeness data. The proposed system using unrefined sets based knowledge discovery from big data, the parallel approximate set methods for information discovery. The propose fuzzy supported pattern generation. Compare with existing algorithm the proposed system give high accuracy of rule generation based fuzzy based rough set.

Keywords: MapReduce; Bigdata; Roughset theory; fuzzy based rule generation.

I. INTRODUCTION

Rough set theory, a new mathematical model developed by Pawlak in 1980s [1], [2], is an approach for imperfect or vague knowledge. This approach is used in the areas of knowledge acquisition, knowledge discovery, pattern recognition, machine learning and expert systems. Rough set theory provides means of identifying hidden patterns in data, finding minimal set of data, pointing out significant data, generating sets of decision rules from data. Rough set theory assumes that some form of information is associated with every object of universe. Objects are said to be indiscernible, if they are characterized by similar information. The mathematical basis of rough set theory is the indiscernibility relation exhibited by the above information.

Knowledge Discovery process through data mining is divided into four: Selection, Preprocessing, Data Mining and Interpretation [3]. Selection is a process of creating a target data set. It is not that the entire data base is to undergo the data mining process, because of the fact that the data represents a number of different aspects of the unrelated domain. Hence, the very purpose of data mining is to be clearly specified. Pre-processing is nothing but processing or preparing the data set that could be used for analysis by the data mining software. This further involves activities that resolve undesirable data characteristics like missing data, irrelevant non-variant fields and outlying data points. This preprocessing step results in generating a number of subsets of original set. All the data are converted into a format acceptable for data mining software. The above process of collection and manipulation of data in data mining process is called collection and cleaning.

The basic concept of rough set theory is the approximation of spaces. The subset of objects defined by lower approximation is the objects that are definitely part of the interest subset and the subset defined by upper approximation are the objects that will possibly part of the interest subset. The subset defined by the lower and upper approximation [3] is known as Rough Set. Rough set theory has evolved into a valuable tool used for representation of vague knowledge, identification of patterns, knowledge analysis and minimal data set.

II. RELATED WORK

The many existing use roughest for feature selection and knowledge discovery .The MapReduce programming model has simplified the implementations of many data parallel applications. The simplicity of the programming model and the quality of services provided by many implementations of MapReduce attract a lot of enthusiasm among parallel computing communities. J. Y. Liang et al., in this paper propose when a group of objects are added to a decision table, to introduce incremental mechanisms for three representative information entropies and then develop a group incremental rough feature selection algorithm based on information entropy. Q. H. Hu, W. Pedrycz et al., in this paper proposed neighborhood decision error rate (NDER), which is applicable to both categorical and numerical features. In this paper introduce a neighborhood rough-set model to divide the sample set into decision positive regions and decision boundary regions. Q. H. Hu, Z. X. Xie et al., in this paper efficient hybrid attribute reduction algorithm based on a generalized fuzzy-rough model. Q. H. Hu, D. R. Yu et al., in this papers the model the sizes of the neighborhood lower and upper approximations of decisions reflect the discriminating capability of feature subsets. The size of lower approximation is computed as the dependency between decision and condition attributes.

In existing system use theoretic framework based on rough set theory, which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data. In existing have many heuristic attribute based reduction algorithms have been proposed however, quite often, these methods are

computationally time-consuming. The massive data mining and knowledge discovery present a tremendous challenge with the data volume growing at an unprecedented rate. Rough set theory has been successfully applied in data mining. The lower and upper approximations are basic concepts in rough set theory. The effective computation of approximations is vital for improving the performance of data mining or other related tasks. M. Kryszkiewicz et al., in this paper propose Rough Set approach to reasoning in incomplete information systems. The show how to find decision rules directly from such an incomplete decision table, which are as little non-deterministic as possible and have minimal number of conditions. Several existing rough set methods of computing decision rules from incomplete information systems are analyzed and compared. which of these methods are capable of generating all optimal certain rules or a class of optimal certain rules and which methods may lead to generation of false rules.

In existing used on non symmetric similarity relations, while the second one uses valued tolerance relation. Both approaches provide more informative results than the previously known approach employing simple tolerance relation. The attribute-value pair blocks, used for many years in rule induction, may be used as well for computing indiscernibility relations for completely specified decision tables. The compute characteristic sets, a generalization of equivalence classes of the indiscernibility relation, and also characteristic relations, a generalization of the indiscernibility relation. For incompletely specified decision tables there are three different ways lower and upper approximations may be defined: singleton, subset and concept. Set-valued information systems are generalized models of single-valued information systems. The attribute set in the set-valued information system may evolve over time when new information arrives. Approximations of a concept by rough set theory need updating for knowledge discovery or other related tasks. The variation of the relation matrix is discussed while the system varies over time. The incremental approaches for updating the relation matrix are proposed to update rough set approximations.

In existing various approaches to interpreting queries in a database with incomplete information are discussed. A simple model of a database is described, based on attributes which can take values in specified attribute domains. with the corresponding sets of axioms which serve as a basis for equivalent transformations of queries. The technique of equivalent transformations of queries is then extensively exploited for evaluating the interpretation of (i.e. the response to) a query.

In this proposed system using fuzzy based parallel knowledge discovery from the incomplete data. An efficient rule induction in this proposed system. The experimental result compare many dataset based rule generation which compare the existing system. Different problems can be addressed though Rough Set Theory, however during the last few years this formalism has been approached as a tool used with different areas of research. Rough set theory, which has been used successfully in solving problems in pattern recognition, machine learning, and data mining, centers around the idea that a set of distinct objects may be approximated via a lower and upper bound.

III. ROUGHSET THEORY

Rough sets theory provides a mathematical tool that can be used to find out all possible feature subsets. In the feature selection problem the principal idea is to recognize the dispensable and indispensable features, using the discernibility matrix. The purpose of using Rough sets is to find the Core, that is, the set of all indispensable features.

3.1 Rough sets concepts

In this section, we will define some concepts related to Rough sets theory.

Definition1. Let U be a non-empty set and let x, y, and z be elements of U. Consider R such that xRy if and only if (x,y) is in R. R is an equivalence relation if it satisfies the following properties:

i) Reflexive Property: (x, x) is in R for all x in U.

ii) Symmetric Property: if (x, y) is in R, then (y, x) is in R.

iii) Transitive Property: if (x, y) and (y, z) are in R, then (x, z) is in R.

Definition2. A partition P of U is a family of nonempty subsets of U such that each element of U is contained in exactly one element of P.

$$\text{i) } U = \bigcup_{i=1}^{n} U_i \ ,$$

$$\text{ii) } U_i \bigcap U_j = \phi \text{, for all } i \neq j$$

Definition3. The Indiscernibility relation Rough sets theory is based on the Indiscernibility relation. Let T = (U, A,C,D) be a decision system data, where U is a non-empty finite set called the universe, A is a set of features, C and D are subsets of A, named the conditional and decisional attributes subsets respectively. The elements of U are called objects, cases, instances or observations. Attributes are interpreted as features, variables or characteristics conditions. Given a feature a, such that:

$$a : U \rightarrow V_a \text{ for } a \in A, V_a$$

is called the value set of a .

Let $a \in A,\ P \subseteq A$ the indiscernibility relation IND (P), is defined as follows:

$$IND(P) = \{(x, y) \in U \times U : for\ all\ \ a \in P,\ a(x) = a(y)\}$$

In simple words, two objects are indiscernible if we can not discern between them, because they do not differ enough.

The indiscernibility relation defines a partition in U. Let U/IND(P) denote a family of all equivalence classes of the relation IND(P), called elementary sets. Two other equivalence classes U/IND(C) and U/IND(D), called conditional and decisional classes respectively, can also be defined. The decisional attribute D determines the

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

decisional classes U/ IND(D)= {x1,x2….xr(D)} of the universe U, where X {x U : D(x) k} k <1 <(r(d) is called the k-th decisional class of decision system data T. The equivalence classes of the discernibility relation, which are the minimal blocks of the information system, can be used to approximate these concepts, then a set X could be approximate using upper and lower approximation.

Definition4. Lower approximation of a subset Let R C and X U , the R-lower approximation set of X, is the set of all elements of U which can be with certainty classified as elements of X.

$$\underline{R}X = \cup\{Y \in U / R : Y \subseteq X\}$$

According to this definition, we can see that R-Lower approximation is a subset of X, thus RX X.

Definition5. Upper approximation of a subset The R-upper approximation set of X is the set of all element of U that can possibly belong to the subset of interest X.

$$\overline{R}X = \cup\{Y \in U / R : Y \cap X \neq \phi\}$$

Note that X is a subset of the R-upper approximation set, thus Definition 6. Boundary Region: It is the collection of elementary sets defined by:

$$BN(X) = \overline{R}X - \underline{R}X$$

These sets are included in R-Upper but not in R-Lower approximations.

Definition7. A subset defined through its lower and upper approximations is called a Rough set. That is, when the boundary region is a non-empty set.

Definition8. A subset is called Crisp when its boundary region is empty

Definition9. Positive region of a subset

It is the set of all objects from the universe U which can be classified with certainty to classes of U/D employing attributes from C.

$$POS_C(D) = \underset{X \in U/D}{\cup} \underline{C}X$$

where CX denotes the lower approximation of the set X with respect to C. The positive region of the subset X belonging to the partition U/D is also called the lower approximation of the set X. The positive region of a decision attribute with respect to a subset C represents approximately the quality of C. The union of the positive and the boundary regions constitutes the upper approximation. The upper approximation contains all data that can possibly be classified as belonging to the set X.

Definition10. Negative region of a subset the negative region consists of those elementary sets that have no predictive power for a subset X given a concept R. They consist of all classes that have no overlap with the concept. Thus is,

$$NEG_R X = U - \overline{R}X$$

Definition12. Dispensable and Indispensable Features Every dataset contains conditional and decision features. Some of these features are indispensable which are very important in the analysis. The problem of feature selection is searching for indispensable features and eliminating the dispensable features. Let c C, C is the set of conditional features. A feature c is dispensable in the information dataset T if (POS) POS D C- c = POS (D) C; otherwise feature c is indispensable in T and should be considered in the final best subset of feature. The main purpose in the feature selection process is to retain all indispensable features that cause the decision system data T to be consistent. Thus, if c is an indispensable feature, deleting it from T will cause T to be inconsistent. In the other hand, if a feature is dispensable, it could be eliminated from the dataset and in this way the dimensionality of the dataset will be reduced.

Definition 13. Reduct A system T = (U, A,C,D) is independent if all c in C are indispensable. A set of features R Í C is called the reduct of C if T'= (U, A, R,D) is independent and POS (D) POS (D) R C = . Furthermore, there is not T Ì R such that

$$POS_T(D) = POS_C(D)$$

A Reduct is a minimal set of features that preserves the indiscernibility relation produced by a partition of C. There could be several subsets of attributes like R. Similar or indiscernible objects may be represented several times on an information table, some of the attributes may be superfluous or irrelevant, and they could be removed without loss of classification performance.

Definition14. The Core: The set of all the features indispensable in C is denoted by CORE(C). The Core is the set of all single element entries of the discernibility matrix, that is,

$$CORE(C) = \{a \in C : m_{ij} = \{a\} \text{ for some } i, j\}$$

We have

$$CORE(C) = \cap RED(C)$$

where RED(C) is the set of all reducts of C. Thus, the Core is the intersection of all reducts of an information system. The Core does not consider the dispensable features and it can be expanded using Reducts. The feature subset obtained is good enough to make information induction.

Definition 15. The Dependency coefficient: Let T = (U, A, C,D) be a decision table. The Dependency Coefficient between the conditional attribute C, and the decision attribute D is given by

$$\gamma(C, D) = \frac{card(POS(C, D))}{card(U)}$$

where, card indicates cardinality of a set. The dependency coefficient varies between 0 and 1, since it expresses the proportion of the objects correctly classified with respect to the total, considering the conditional features set. If g=1, D depend totally on C, if 0<g<1, the D depends partially on C, and if g=0, then D does not depend on C. A decisional attribute depends on the set of conditional features if all values of decisional feature D are uniquely determined by values of conditional attributes. i.e. there exists a dependency between values of decisional and conditional

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

features. An algorithm to calculate de Dependency coefficient is given below
i. Create the partition of the Dataset D without considering the class feature.
ii. Set Positive equal to zero, where Positive represents the cardinality of the Positive region.
iii. Search for Elementary sets that only belong to a unique class.
iv. For i=1 to the number of elementary sets
If card(class(elementarySet[i])) =1 then
P = Card(elementarySet[i])
Positive = positive + P
iv. Finally calculate dependency as follows:

$$Dependency = \frac{card(Positive)}{card(data)}$$

In the worst case the order of the algorithm is O(n2×p), where n is the number of instances and p is the number of attributes. Since the creation of the partition is of order O(n2p) and the computation of the positive is of order O(n) in the worst case.
Definition16. Accuracy of the approximation: The accuracy of the approximation to the set X from the elementary subsets is measured as the ratio of the lower and the upper approximation size. The ratio is equal to 1, if no boundary region exists, which indicates a perfect classification. In this case, deterministic rules for the data classification can be generated.

$$\alpha(X) = \frac{Lower(X)}{Upper(X)}$$

Thus, a set X with accuracy equal to 1 is crisp. Otherwise X is rough.
Definition17. Dependency relation matrix: Given the information table, we can calculate the Dependency Matrix for each couple of feature i a and j a according to the class feature.

$$D(a_i, a_j \mid C) = \sum_{a_i, a_j, Y_c} \frac{|\, Pos_{a_i}^{Y_c}(a_j) \,|}{card(Y_c)}$$

represents the positive region of attribute j a relative to attribute i a within the class value c.

IV. FUZZY BASED KNOWLEDGE DISCOVERY

Fuzzy sets use the membership function to give a degree of membership. A fuzzy set on a classical set X is defined as follows:

$$\tilde{A} = \{(x, \mu_A(x)) \mid x \in X\}$$

m denotes the fuzzy membership function. A subset A is a fuzzy set when its membership in X is not crisp, but it is subject to gradation; formally this is expressed in the interval [0,1] by the fuzzy membership function. The membership function (x) A m quantifies the grade of membership of the elements x to the fundamental set X. An element mapping to the value 0 means that the member is not included in the given set, 1 describes a fully included member. Values strictly between 0 and 1 characterize the fuzzy members. Sometimes, a more general definition is used, where the membership function takes on values in an arbitrary fixed algebra or structure L.
The parallel methods for knowledge acquisition based on rough set theory using Map Reduce. Each sub-decision table can compute numbers of elements in equivalence classes, decision classes and union classes independently. At the same time, the classes of different sub-decision tables can combine together if their information sets are the same. Hence, it could be changed to a Map Reduce problem and we design three parallel methods based on rough set theory for knowledge acquisition our parallel models for three kinds of knowledge acquisition methods, which all contain two steps.
Step 1: By the above analysis, the computation of the cardinality of the equivalence classes |E|, the cardinality of decision classes |D| and the cardinality of
union classes |E ∩ D| can be executed in parallel. In detail, we present the corresponding algorithms based on MapReduce/Combine, which are outlined in Algorithms 1, 2 and 3, respectively
Step 2: After computing the cardinalities of the equivalence classes, decision classes and union classes, by Definition 4, the accuracy Acc(D|E) and the coverage Cov(D|E) are computed. Then, three kinds of rule sets are generated.

ALGORITHM 1: MAP (KEY, VALUE)
Input:
//key: document name
//value: Si = {Ui, C ∪ D, V, f}
//Global variable: B ⊆ C
Output:
//key': the information set of the object with respect to the sets B, D and B ∪ D
//value': the count

1 begin
2 for each x ∈ Ui **do**
3 let key'= 'E' +x B; //Here, 'E' is flag, which means the equivalence class
4 output.collect(key', 1);
5 let key'= 'D' +x D; //Here, 'D' is flag, which means the decision class
6 output.collect(key', 1);
7 let key'= 'F' +x B∪D; //Here, 'F' is flag, which means the association between the equivalence class and decision class
8 output.collect(key' , 1);
9 end
10 end

ALGORITHM 2: COMBINE(KEY, V)
Input: //key: the information set of the object with respect to the sets B, D and B ∪ D
//V: a list of counts
Output:
//key': the information set of the object with respect

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

to the sets B, D and B ∪ D
//value': the count.
1 begin
2 let value'= 0 and key'= key;
3 for each v ∈ V **do**
4 value' = value'+ v;
5 end
6 output.collect(key', value');
7 end
REDUCE(KEY, V)
Input:
//key: the information set of the object with respect to the sets B, D and B ∪ D
//V: a list of counts
Output:
//key': the information set of the object with respect to the sets B, D and B ∪ D
//value': the count.
1 begin
2 let value'= 0 and key'= key;
3 for each v ∈ V **do**
4 value'= value'+ v;
5 end
6 output.collect(key', value');
7 end

Combine and Reduce phases in Step 1 of three parallel methods. Then, the accuracy and the coverage of rules are computed and rule sets are generated, where √ and × means it is a rule or not, respectively. Obviously, the numbers of rules in sets are 4, 4, 5 for these methods.

VI. EXPERIMENTAL RESULTS

Utilize the large data set KDD99 from the machine learning data repository, University of California at Irvine, which consists of approximately five million records. Each record consists of 1 decision attribute and 41 condition attributes, where 6 are categorical and 35 are numeric.

Since our method can only deal with categorical attributes, discretize the 35 numeric attributes firstly. In addition, three synthetic data sets have been generated by means of the WEKA data generator.

Where HDFS means Hadoop distributed file system. The data sets KDD99, Weka-1.8G, Weka-3.2G and Weka-6.4G are split into 64, 64, 128 and 256 blocks, respectively when we upload these data to HDFS.

Table 1 Accuracy, Coverage, Rules

Data sets	Samples	Features	Classes	Size	Number of Blocks in HDFS
KDD99	4,898,421	41	23	0.48 GB	64
Weka-1.8GB	32,000,000	10	35	1.80 GB	64
Weka-3.2GB	40,000,000	15	45	3.20 GB	128
Weka-6.4GB	80,000,000	15	58	6.40 GB	256

V. CONCLUSION

In this proposed system knowledge acquisition based on rough fuzzy sets, which combines features of rough sets and fuzzy sets. The continuous attributes in the decision table are fuzzified with fuzzy membership functions. The domain partition is accomplished after establishing fuzzy similarity matrix. Attributes reduction can be obtained using rough-fuzzy dependency, and then decision rules can be acquired. At last, an example is illustrated and proves the approach is effective and practical.

REFERENCES

1. Ahrens, J. and Dieter, U. [1974]. Computer methods for sampling from Gamma, Beta, Poisson and Binomial distributions. Computing, (12), 223–246.
2. Ahrens, J., Kohrt, K. and Dieter, U. [1983]. Algorithm 599: sampling from Gamma and Poisson distributions. ACM Transactions on Mathematical Software, 9(2), 255–257.
3. Artin, E. [1964]. The Gamma Function. New York, NY: Holt, Rinehart and Winston. Translated by M. Butler.
4. Auer, P. [1997]. On learning from multiple instance examples: empirical evaluation of a theoretical approach. In Proceedings of the Fourteenth International Conference on Machine Learning (pp. 21–29). San Francisco, CA: Morgan Kaufmann.
5. Bilmes, J. [1997]. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussia nmixture and hidden Markov models. Technical Report ICSI-TR-97-021, University of Berkeley.
6. Chevaleyre, Y. and Zucker, J.-D. [2000]. Solving multiple-instance and multiplepart learning problems with decision trees and decision rules. Application to the mutagenesis problem. Internal Report, University of Paris.
7. Chevaleyre, Y. and Zucker, J.-D. [2001]. A framework for learning rules from multiple instance data. In Proceedings of the Twelveth European Conference on Machine Learning (pp. 49–60). Berlin: Springer-Verlag.
8. Cohen, W. [1995]. Fast effective rule induction. In Proceedings of the Twelveth International Conference on Machine Learning (pp. 115–123). San Francisco, CA: Morgan Kaufmann.
9. Dempster, A., Laird, N. and Rubin, D. [1977]. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistics Society, Series B, 39(1), 1–38.
10. Dietterich, T., Lathrop, R. and Lozano-P´erez, T. [1997]. Solving the multiple instance problems with the axis-parallel rectangles. Artificial Intelligence, 89(1- 2), 31–71.
11. Frank, E. and Witten, I. [1998]. Generating accurate rule sets without global optimization. In Proceedings of the Fifteenth International Conference on Machine Learning (pp. 144–151). San Francisco, CA: Morgan Kaufmann.
12. Frank, E. and Witten, I. [1999]. Making better use of global discretization. In Proceedings of the Sixteenth International Conference on Machine Learning (pp. 115–123). San Francisco, CA: Morgan Kaufmann.
13. Frank, E. and Xu, X. [2003]. Applying propositional learning algorithms to multiinstance data. Working Paper 06/03, Department of Computer Science, University of Waikato, New Zealand.