Rough Sets Base Associative Classification Rules Extraction from Big Data

Hanumanthu Bhukya, M.Sadanandam

Abstract: ***Big Data is a current burning challenge for the data analytics research community. Many conventional data analytics techniques have been extended to the MapReduce framework to process Big Data. But in our literature review, we find that for the MapReduce system there is an absolute lack of rough set-based technique. To facilitate this and recognize the importance of the rule-based classification techniques, we suggest a rough-set associative classification rules extraction process for the MapReduce framework. The implementation and evaluation of the Big Data Standard data set demonstrated the efficiency of our suggested approach.***

I. INTRODUCTION

In the recent era of development, the magnitude of data increasing at uncontrollable speed, the mining of Big Data and the discovery of knowledge is conceivably a new challenge. Rough set theory was applied effectively in data mining to detect knowledge. MapReduce's approach subsequently noted a great deal of interest from both groups, such as scientific and industry, for its importance throughout big data research. With the escalation of information and communication technology, vast amounts of data are collected in various forms through various sources such as gadgets and sensors. In order to process such data, autonomous or coupled applications usually go beyond the zeta scale, the need for computation is successively required. With rapid growth and updating of big data in real-life applications, the latest challenge is to quickly obtain helpful information with big-data mining methods. Google built a system called MapReduce to process big data, to effectively evaluate huge amounts of data, to carry huge disseminated data sets on clusters of computers. MapReduce turns out to be a popular cloud computing environment model. Go ahead with Google's work, there are numerous developments in MapReduce and a lot of conventional methods linked to the MapReduce framework.

To determine formerly unknown patterns efficiently from huge databases Data mining has turned out to be a burning issue for decision-maker. RST was invented by Pawlak in 1982 as the usual mathematical theory, the representation of knowledge in the area of relevance in terms of the equivalence relationship group [1].

Rough sets don't need any overture or supplementary information concerning data similar to probability in probability theory, grade of membership in fuzzy set theory this is the primary advantage of it. Several rough-set methods to data mining have been applied successfully at present [2].

Revised Manuscript Received on October 15, 2019

Hanumanthu Bhukya*, Full-Time Research Scholar, Department of CSE, UCE & T, Kakatiya University, Warangal, India. Email: bhcsekits@gmail.com

Dr.M.Sadanandam, Department of CSE, Kakatiya University, Warangal, India. Email: msadanandam@kakatiya.ac.in

Since it is a useful tool for dealing with vagueness, incompleteness and elusiveness, the roughest theory has recognized both theoretical and practical awareness as it was launched in 1982[3]. Several extensions were introduced to address the requirements of actual applications in order for tolerance [4], similarity [5], relations of governance[6], two-fold relationships[7], wrappers[8], and environs[9] to take care of indiscernibility[2]. Various hybrid systems, for example decision-making rough sets [10], game theoretical rough sets [11], floating rough sets and raw fluctuation sets [12] have evolved as a combination of new theories. These theoretical frameworks are useful in a broad range of areas, for instance knowledge on the detection of credit fraud, medical diagnosis, databases, fault analysis, selection of features, outlier detection, etc.[13–24].

Because of the plenty of noisy, immaterial or deceptive features, vague and conflicting information in real life issues has become a major requirement for the selection of features. Rough sets [25-27] are capable of removing ambiguity and vagueness, of finding patterns in conflicting information.

Rough set

- Rough set theory is a method for working out the incidence of imprecision. RST expansion is a classical set theory, for use corresponding indistinctness or elusiveness. A rough set involves working on the edge regions of the set [28]. The idea of space approximation is a fundamental RST scheme that can be an ordered pair P = (N, S), where
- *P* system for information,
- *N* group of objects are nonempty, also known as the *universe*, and
- *S uniformity* relation on *N*, nothing but the *indiscernibility relation*. If a, bϵ*N*, & a*Rb*, a and *b* be identical in *P*.

Each class of uniformity stimulated by S is known as a basic set in N, a group of finite traits can be symbolized as N / S. Any finite union of basic sets in P is a definable set P. Pro aϵ*U*, let $[A]_S$ refer the uniformity class of S includes a. Pro every $A \subseteq N$, *A* characterized within *P* via a couple of sets, its *min* and *max approximations* in *P* be defined as correspondingly:

$$\underline{S}A = \{A \in N | [A]_S \subseteq A\},$$
$$\bar{S}A = \{A \in N | [A]_S \cap A \neq \{\}\}.$$

Rough set P includes every subsets of N by the identical min and max approximations.

Rule generation is the fundamental job in any system of learning [29]. At this point, we illustrate how generate decision rules based on the reduct system acquired as of the relation with the equivalence relations related by attribute sets may be used to produce decision rules. Presume that a set P = {p_i, p_i, . . . , p_n} of autonomous attributes and a only dependent attribute k. Here no constraint of generalization, since we are passing through merely the division information of θ_k, and therefore k can be a compound attribute achieved from Q $\subseteq \Omega$.

Presume that the division induced by means of θp is {*A*1, *A2*, . . . , *Au*}, and the lone induced via θ_k is {*B*1, *B*2, . . . , B_v}. Through every A*i* link the set N*i* = {*Bj* : A*i* ∩ *Bj* = {}}. Because the sets B1, *B* 2, . . . , B_v partition N, we get: If a $\in$ *Ai*, then a $\in$ *Bj*1 or . . . or *a* $\in$ *Bji*(*j*) . every class A*i* of θ_P communicates to a feature vector (*ai*)1_*i*_*n*, where a $\in$ *Ai* if and as long as *fp*1 = *a*1 and & *fpn*(*a*) = *xn*, likewise, *a* $\in$ *Bj* if & provided that $f_k(a)$ = y*j* pro some *yj* $\in$ *Vk*. Equation leads to the a rule form: If *fp*1(a) = *x*1 and & *fpn*(*a*) = x*n* after that *fk*(*a*) = *yj*1 or . . . or *fk*(*a*) = y*ji*(*j*) .

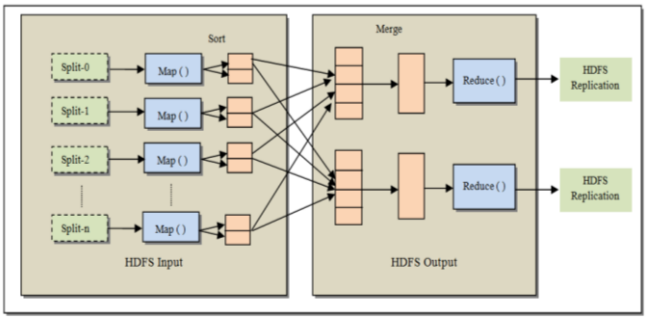
In the RST, distinguish two different kinds of rules: deterministic & non-deterministic rules: If a class A*i* of θp intersects precisely one B*j*, after that A*i* $\subseteq$ *Bj* , and the worth of *k* of any *a* $\in$ *Ai* is distinctively resolute. Or else *fk*(*a*) possibly N*i* contained in any class and it contain a suitable disjunction on the RHS of Equation. A class A*i* is nothing but deterministic if it included in some B*j* , or else call it indeterminisic. If every classes of A*i* be deterministic afterward $\theta p \subseteq \theta k$ as well as k is dependent on *P*.

According to us, the majority of conventional algorithms driven by rough sets are chronological algorithms and rough set tools available now can only run on a individual computer on its own to handle huge data sets. The rough set of parallel computation approximations is implemented in the direction of enhancing rough sets applications in the data mining field and handling huge data sets. And the use of MapReduce Technique[30] can achieve this parallel approximation.

map () input to the mapping function as a pair and generate the key, the value pairs. MapReduce clusters together all the values linked by the same k key and convert them to the reducer() function.

reduce () Reduce is the function that uses the output of the map function as input as key k and its associated values. These values are merged together to form a perhaps smaller set of values. By performing sorting and shuffling, it generates reduced values from the Map function.

The storage system which is used by Hadoop applications is the Hadoop Distributed File System (HDFS). A data block generates numerous reproductions of HDFS and assigns these reproductions to data nodes to allow reliable, extremely fast computing. Hadoop consists of two key elements: a file storage space and a dispersed processing system. The primary key element in file storage space is also called HDFS (Hadoop Distributed File System). It gives reliable, scalable, relatively inexpensive storage. HDFS stores the file crossways of all servers in a cluster. By repeatedly checking cluster servers and blocks that monitor HDFS data and ensuring ease of use of information. The subsequent very important element of Hadoop is nothing but "MapReduce" the parallel data processing framework. The code for HDFS (Hadoop distributed file system) and MR(Map Reduce) runs on the same set of nodes. It enables the implementation of java source code in MapReduce programming, as well as other languages.



The Map Reduce programming model

II. SURVEY ON ROUGH SETS BASED CLASSIFICATION

In the presence of vague and partial information, the categorization of rough sets was introduced. Core and Reduce are two key concepts in RST. The primary definition of a reduct is a subset of attributes that is appropriate to explain the verdict's attributes. An NPhard problem is to pronounce any set of reductions pro a set of data [31].

Set estimation algorithms are used to achieve a reduction [32]. The hub includes all the reducts. The hub corresponds to the main data in the actual data set. The hub is the intersection of every feasible reduction.

There are several efforts to apply RST to rule finding. Decisions and rules are drawn up on the basis of the reductions, are envoy of the knowledge that can be obtained from the data set. Two components are used in the mining process of association rule to support organizational decision-making and knowledge management in the contribution of Li and Cercone [33].

Customer records may be used for cluster sales events based on likeness in the characteristics of the SOMs. Each cluster may be subject to rules to make clear the RST associations. A.O et al. [34] shall be used for all data reducts that comprise a nominal subset of attributes linked to the class label for the classification of rough sets. RST can help resolve whether there is pleonastic information in the data in order that the data required for applications could be obtained. In conflicting data, RST-driven rule generation system may generate minor and non-redundant rule sets.

For example, in [35], the rough set method can be used to categorize diverse types of meteorological squall events that are accountable for ruthless summer climate. A rough set approximation-driven approach was presented in [36] to cluster web dealings from web access logs. Using this method, users can successfully find patterns of web page access to mine web log records.

In applications of technology, qualitative and quantitative data as of different sources perhaps associated; therefore the quantity of attributes can be increasing significantly [37].

Under the traditional RST, incremental mining algorithms for learning classification rules are resourcefully conferred in [38,39] as soon as the attribute set in the information system grows eventually.

Rules are the contemplation of certain principles on the basis of which decisions can be made for the purpose of making them. Rule is a declaration to launch a standard or principle, and as a standard to guide or mandate action or conduct. The rule is a qualified statement that can tell the system how to respond in an exacting situation. The generation of rules in data mining is in the form of association rules and was primarily introduced as a market-based analysis [31].

In the case of association rules, the creation of a rule is based on the idea of a frequent pattern of mining for the discovery of inspiring relationships and association between elements. Subsequently, techniques are developed for the mining classification rule [40]. Normally, rules are represented in the form of an IF-THEN set of patterns. The IF element (or LHS) of the pattern is referred to as an antecedent rule. The THEN component follows. In the precursor rule, the stipulation includes single or multiple attribute check which are logically linked to AND procedures.

The result includes the forecast of the class. Mining of brief rules is presented in the contributions of [41] a rough set-driven method of conflicting data. In this method, a heuristic algorithm is used to construct brief classification rules. In numerous decision-making and analytical applications, such as financial and economic research [42] and network security, rule-driven approaches have been successfully applied.

There are three different types of rule generation methods that are very common: frequent association rules for mining, unusual rule association for mining, and multi-objective rule for mining. inside Frequent Association Rule Mining (FARM)[31], if its support is not below a given threshold, an association rule is often called in this technique. And this can be true if it also has a min confidence over a user-set threshold.

Association rule r is referred to in Rare Association Rule Mining (RARM)[32] as rare or sporadic if its assistance is not more than the total aid given. This means rule r is uncommon if its funding is beneath the minimum amount of aid given. If it is sure, an uncommon principle of association is true.

The mining problem is known in Multi-Objective Rule Mining (MORM)[44] as a multi-objective problem instead of a single objective problem. Procedures for instance support count, comprehensibility and interest is measured to be diverse goals in judging a rule; the mining issue has to be addressed. Support count is the quantity of records that gratify every conditions present in the rule. Knowing it helps to measure the understandability of the rule that can be metric by the number of attributes took part in the policy. Value can be calculated by the shocking nature of the rule.

The extent of multimodality data is usually very huge in the era of big data; it takes time to perform efficient assessment and reduction of attributes. With regard to this MapReduce, there is an admired parallel computing model [45]. ZR Chen et al. In 2010, introduced a MapReduce-based attribute-reduction approach in the sense of rough data sets [46]. In 2012, TR li et al. suggested a parallel approach pro the estimation of rough approximation sets [47]. Such parallel algorithms calculate large-scale data well and efficiently.

The MapReduce framework [48] for programming was presented in 2004. It is a framework designed to process huge amount of data in a tremendously parallel manner, thus offering a platform for the simple development of applications that are scalable and fault-tolerant.

Apache Hadoop [49, 50] is an accepted MapReduce open source implementation. Mahout [51, 52] is a library for machine learning at the zenith of the Hadoop system. It includes a set of cluster algorithms, advice systems and classification issues. Similar to a Hadoop, Mahout is a freely available project as well. Mahout is responsible for a range of classification models, including Support Vector Machines, Bayesian models, Logistic Regression, and Random Forest. More recently, in the process of dealing with big data are other projects. a few of them are spark [53], a cluster computing system designed to speed up data analytics; storm [54], a disseminated and fault-tolerant real-time computing platform that makes it easy to reliably process unlimited data stream; dremel[55], a versatile, collaborative Ad-hoc search system pro read-only data analysis, and Apache Drill[56], a Distributed Framework enabling information-intensive distributed applications for large-scale, interactive data analysis.

Dean J and G Sanjay [57] gave a brief narrative of the MapReduce programming model with various programs as examples. And also provides a summary of implementation of the MapReduce programming model, with fault tolerance, task granularity and location. This gives explanations for the successful use of Google's MapReduce programming framework. The design conceals parallelization information, fault tolerance, and load balancing, making it accessible. And as MapReduce computations, a wide variety of problems are also presented effortlessly. To reduce the impact of slow machines and to handle system errors and data loss, superfluous execution can be used.

Zdzisław Pawlak and Andrzej Skowron[58] present key concepts for rough set theory and various directions for study and very good application based on the rough set approach.

In this paper, the style based on discernibility and Boolean reasoning for the ability to compute unlike entities, including reductions and rules of decision, was mentioned. It was explained that the rough set strategy be able to be used for synthesizing and analyzing notion approximations in the distributed ambiance of clever systems.

Knowledge acquisition using the MapReduce technique three rough set of methods are proposed by Zhang, T Li, Yi pan [59]. Testing the accomplishment of the projected parallel speedup techniques.

Exhaustive findings on actual and synthetic data sets have shown that a large number of data sets in data mining can be effectively processed by the proposed methods.

Junbo Zhang, Tianrui Li, Da Rusan[60] are suggesting a parallel technique for computing rough-set approximations. As a result, algorithms matching the parallel technique based on the MapReduce systems are provided to handle with the vast data. And in order to test the performance of suggested parallel algorithms, scaleup ,speedup and sizeup were also used.

III. PROPOSED ROUGH SET BASE CLASSIFICATION MODEL FOR HDFS

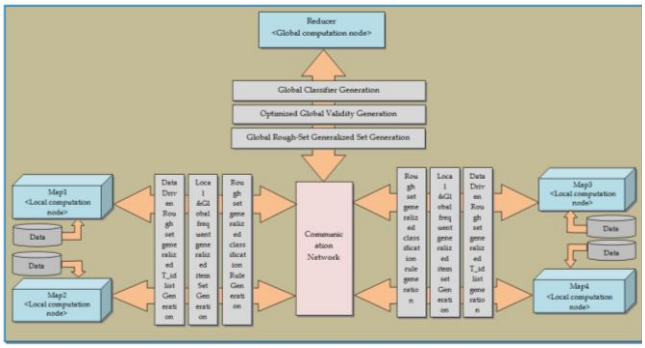


Fig 1. MapReduce Knowledge Discovery architecture based on associative classification

The proposed framework work constructs generalized Rough sets for the discovery of knowledge by MapReduce. The framework presumes that the data is fragmented horizontally between manifold shared nothing systems coupled to the distribution environment. The framework diminishes the time-complexity of the classifier generation process by using generalized Rough sets & parallel generation of items at local sites. A centralized site called Reducer Global Computation Node (Global_CM) will initiate and manage the total system operation.

Two MapReduce processes are initiated by the Global Computation Node (Global_CM) to generate targeted rough-set associative classification models. The MapReduce framework will parallel the computation and improve the computational efficiency of the model. The first MapReduce process initiated by Global Computation Node is "Class label-based generation of tid-list from HDFS data chunks" where basic class label base item sets are discovered by targeted dataset. The next MapReduce process is the "Rough set-Based Associative Classification Rules Generation" that applies the rough set to finalize rules by rough set approaches. In the subsequent sections, the detailed discussion of the MapReduce process will be discussed.

3.1 Class label based Tid-list generation of HDFS data:

The proposed approach adopts HDFS to manage data spread across the distributed file system, using MapReduce framework to extract associative classification rules. However, considering easy updating of processed results while dynamic increments and data decrements, HDFS will organize increments and data decrement in logic units called chunks. The HDFS uses the Map node to process each logical data chunk, so the output Map function corresponds to the output of the local data chunk. At Reducer node, the output of all local chunks will be aggregated, resulting in a global Tid-list.The other contribution of the paper is we make use of class label base Tid-list where a same attribute value with diverse class labels will be measured to be different. This is because the same value of the attribute in a different transaction can lead to a different label of the class. For example, the Rented-Home attribute can lead to a high income group in the case of some user data, where the same attribute can lead to a lower income group in some transactions.

According to the above strategy, the proposed model first uses Map Node Computation to generate Class Label Base Tid-list from each loaded data chunk, then this whole class label Tid-list will switch to the corresponding reducer node according to their class label to generate a global Tid-list. The merge Tid-list function is used to merge two different Tid-lists. In this approach, the number of Reducer Nodes dependence straightly on the number of dataset class labels. The combination of class labels Tid-list yields of all Reducer nodes is a global class label based on Tid-list. Using the resulting Tid-list, associative classification rules can be easily created using global support and confidence threshold. As parallel processing over the distributed file system mainly involves the generation of item set, this paper focuses on extending the MapReduce framework over the generation of item set. The process of generating a rule over a single processor follows the standard Tid-list Base Classification Method [61]. In the following algorithm, the MapReduce base algorithm for the global class tag base generation Tid-list is shown.

Algorithm3.1: Tid-list based on the globally supported generation of class item:

Input : Data set loaded in HDFS
Output : Global Tid-list and classification rules

Map: Local Class label base Tid_list generation

1. Read the data chunk assigned
2. for every transaction T_i do
   a. Generate a new row in Tid-list representing T_i
   b. If grouping of attribute value and class label of record T_i is not present
   c. Create same as a new column label in Tid-list
   d. If grouping of attribute value and class label is present in T_i then
   e. store value 1 at corresponding entry
   f. Else store value 0.
3. End

Reducer:

1. Read all local Tid-list
2. merge_Tid-list()
   a. create separate rows for all transactions in Global_Tid-list
   b. create separate columns for each attribute labels in Global_Tid-list
   c. if combination of transaction and attribute label supported by local tid-list
   d. place 1
   e. else place 0
3. Generate class label base association rules with global support and confidence.

3.2 Rough set based associative classification rules generation:

Once the associative classification rules generated from HDFS using map reduce framework, the same rules will be used to generate rough set rules. According to this, the Map process will read the generated rules and apply to them decision variables training data to generate lower boundary approximation rules, where the other Map-Process will generate Upper boundary approximation rules and finally these two sets of rules will be consolidated in the Reducer process. In turn to produce a rough set of boundary rules, we propose the MapReduce method shown in the following algorithm 3.2.

Algorithm3.2: Rough set based associative classification rules generation

Input : Class label based association rules in HDFS, Training data D

Output : Global rough set based classification rules

Map-1: Generation of classification rules using lower rough set approximations

1. Read the rule R_i
2. Apply on data D to generate lower approximation space (LR_i)

Map-2: Generation of classification rules using upper rough set approximations

1. Read the rule R_i
2. Apply on data D to generate upper approximation space (UR_i)

Reducer:

I. Read the associated rules from (LR_i & UR_i)

II. Generate the best boundary rules (BR_i)

IV. IMPLEMENTATION EVALUATION & RESULTS

The proposed methods were experimentally evaluated on 8 nodes out of these eight nodes 1 node acting as master node and other nodes acting as slaves. The nodes are rich in Pentium-i5 processor configuration and 8 GB RAM interconnected with 10 GBPS data cable and 1Tb hard disk accompanied. The cluster network formed on Ubantu 14.4 operating systems using the version of Hadoop2.0.0-cdh4.4.0. The experimental evaluation resulted in the measurement of system performance on accuracy and scalability with respect to dynamic increments and data decreases.

The proposed dynamic MR Tid-AC model was experimentally evaluated at three levels, including classification accuracy, time accuracy with respect to dynamic changes and dynamic scalability with respect to mapping nodes. In order to do so, out of the total data 70% of the data loaded and the classification rules generated as per the proposed model will remain 30% of the data loaded and the classification rules generated with the proposed approach and finally the accuracy will be calculated. Usually 10 cross 10 partition approach used to divide the data into data sets for testing and training. In order to develop classification rules for associative in the form of preparation, map nodes were first initiated and then local Tid-lists were generated which were farther used to generate classification rules supported globally as explained in the projected models. The number of counts of map nodes is calculated by the system itself in the experimentation derived from load where the reducer nodes are added founded on the number of classes in the data set of the evaluation.

The KDD-96 UC-Census dataset [62] is an open source classification data set with even more than 1, 41,544 data records for accurate testing. The best claimed accuracy of the standard classification models for 84.47% NBTree, Naive-Bayes 81.69% and C4.5 81.91% [KDD-1996] for the comparative assessment of the accuracy of the proposed model. Finally, the proposed model showed 83.91% significant performance compared to c4.5 and Naïve-Bayes but slightly more inaccurate than the NBTree model.

Third stage of experimentation was carried out toward assess the dynamical nature of the projected model. Two parts are divided into 70: 30 ratios in the 10cross1 validation system to determine the dynamical nature of the training data. The projected MapReduce Tid-AC process, captivating Eight-number of core systems, was initially performed resting on 70% of HDFS data dealings and apply to the training set. The training time of this stage is 921 seconds. In the subsequently, the outstanding 30% of dealings further to the HDFS incremental MapReduce Tid-AC technique were run at 386 seconds (total: 921sec+386sec=) 1307sec with a total of 8 accuracy. Instead of applying MR Tid-Ac if we go back to the total associative classification algorithm by remaining 30% of the data, it takes 1342 sec which results in (921 + 1342=) 2,263 seconds. These experimental results show, therefore, that the projected incremental MapReduce Tid_AC technique is well-suited in case of accuracy and offers a much enhanced time complexity compared to the re-run of the total dynamic dataset increment process.

As indicated by the purpose of the subsequent stage of the trial to determine the scalability of the suggested classification method, time efficacy is reported with regard to the changeable number of systems. For record the experiment performed on 8, 16, 32, 64 map sizes, a significant increase in time complexity was identified with the increased number of computing systems in the proposed model. It indicates the proposed system's scalability function. The product of the device scalability factor shown in Figure4.2

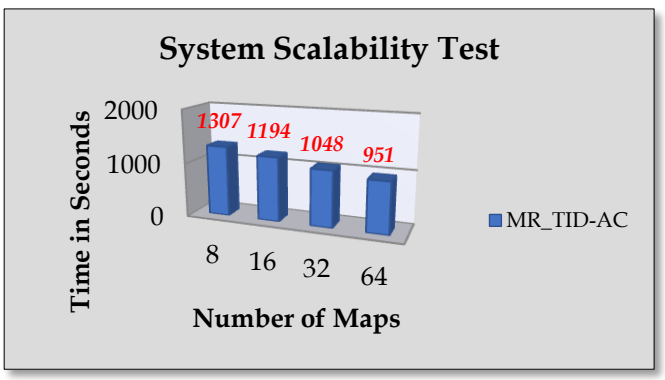


Figure.4.2: Scalability of the scheme by the varying number of maps.

V. CONCLUSION

Increasing data, known as Bigdata, is forcing researchers to adopt MapReduce computing techniques to process data. Realizing that we are proposing MapReduce base techniques for the generation of rough set base classification rules that can handle uncertainty in Big data that is a heterogeneous data collection. The suggested solution has two stages that include the generation of base item sets for the MapReduce base class label and rules generation for the rough set based classification. The proposed approach tested on the standard data set showed its effectiveness. The valid extension of the proposed approach could be proved by extending the system with fuzzy roughest.

REFERENCES

1. PawlakZ, Rough Sets "Theoretical Aspects of Reasoning about Data", Dordrecht Kluwer, 1991.
2. WangG,WuY,ChangL, "An approach for attribute reduction and rule generation based on rough set theory", Journal of Software 10 (11) (1999) 1206–1211.
3. Z.Pawlak,Roughsets, International Journal of ComputerInfSci11(1982)341–356.
4. M.Kryszkiewicz,Roughsetapproachtoincompleteinformations ystems,Info.Science112(1998)39–49.
5. TsoukisA, StefanowskiJ, "On the extension of rough sets under in complete information, in: New Directions in rough sets ,Data Mining , and Granular-Soft Com-puting,vol.1711,SpringerBerlinHeidelberg,1999,pp.73–81.
6. GrecoS, SlowinskiJ and BlaszczynskiJ, "Multi-criteria classification–a new scheme for application of dominance-based decision rules, Eur.J.Oper.Res.181 (2007) 1030–1044.
7. ZhuW, Generalized rough sets based on relations,Info.Science177(2007)4997–5011.
8. BryniarskiE, WybraniecU and Bonikowski Z, , -Skardowska, Extensions and intentions in the rough set theory,Info.Science107(1998)149–167.
9. XWuC,HuHQ,YuDR&LiuJF, "Neighbor hood rough set based heterogeneous feature sub set selection",Info.Science178(18)(2008)3577–3594.
10. YaoYY "Decision-theoretic roughest models", in Proceeding of the Second Int.Conf. on Rough Sets and Knowledge Technology, 2007, pp.1–12.
11. HerbertJ, YaoJT & AgameJ "theoretic perspective on roughest analysis", JChongqing University Posts Telecommunications.20 (2008)291–298.
12. PradeH and DuboisD, "Rough fuzzy sets and fuzzy rough sets, International. J. General Syst.17 (1990)191–209.
13. ZhangWX, WuWZ and LeungY, "Knowledge acquisition in incomplete information systems: A rough set approach, Eur. Journal. Oper. Res. 168 (2006)164–180.
14. WangXZ, He.Q and ChenDG, "FRSVMs:fuzzy rough set based support vector machines, Fuzzy Sets Systems.161(2010)596–607.
15. Wang WT, Dai JH, Tian HW, , Liu L,Decision rule mining using classification consistency rate,Knowl.BasedSyst.43(2013)95–102.
16. Skowron A and Swiniarski RW, "Rough set methods in feature selection and recognition, Pattern Recognition Lett.24 (2003)833–849.
17. ShenQ and JensenR, "New approaches to fuzzy-rough feature selection", IEEETrans. Fuzzy Systems17(4)(2008)824–838.
18. Liao XW, Xu WH, Li Y, , "Approaches to attribute reductions based on rough set and matrix computation in inconsistent ordered information systems", Knowl.BasedSystems.27(2012)78–91.
19. QinYK and haoH.Z, , "Mixed feature selection in incomplete decision table", Knowledge Based Systems57(2014)181–190.
20. MatarazzoB, Slowinski R,& Greco S, "Rough sets in decision making, Encycl.Com-plex.Syst.Sci.(2009)7753–7787.
21. Shen LX, Tay F, , "Fault diagnosis based on rough set theory", Eng. Appl. Artif. Intelligence.16 (1)(2003)39–43.
22. WakulicA Deja and PaszekP, "Applying roughest theory to medical diagnosing, Rough Sets" Intelligent. Syst. Paradig. 4585(2007)427–435.
23. Cao CG Jiang F, Sui YF, "A rough set approach to outlier detection", International J.Gen.Syst.37 (2008)519–536.
24. Chen YS, "Classifying credit ratings for Asian banks using integrating feature selection and the CPDA-based rough sets approach, Knowledge Based Syst. 26(2012) 259–270.
25. ZPawlak, "Rough Sets" International J. Computer Information Science. 198211 (5), 341– 356.
26. Z.Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishing, Dordrecht, 1991.
27. PawlakZ, "Rough set approach to knowledge-based decision support". Eur. J. Operation 1997 Res. 99, 48–57.
28. Busse Grazymala, ZPawlak SlowinskiJW and ZiarkoW "Rough sets", Communications of the ACM, Vol. 38, pp.88–95, 1995.
29. Rybinski H. and KryszkiewiczM, (1996b). "Reducing information systems with uncertain real value attributes". In 6th International Conferences, Information Processing and Management of Uncertainty in Knowledge- Based Systems, Proceedings (IPMU'96), Vol. II. July 1–5, Grenada. pp. 1165.
30. PawarBV and Dhande Varda C, "A Survey on Parallel Method for Rough Set using MapReduce Technique for Data Mining", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.
31. ImielinskiT, R Agrawal, & ASwami. "Mining association rules between sets of items in large databases", Proceedings of 1993 ACM SIGMOD, ACM, New York, NY, USA,1993, pp.207–216.
32. PK Reddy & RU Kiran, "Mining rare association rules in the datasets with widely varying items" frequencies Lecture Notes in Computer Science, Vol. 5981, 2010,pp.49–62.
33. NCercone and J Li "A rough set based model to rank the importance of association rules", Lecture Notes in Computer Science, 2005,Vol. 3642, pp.109–118.
34. SOFalaki, AOAdetunmbi, OSAdewale, and BKAlese, 'Network intrusion detection based on rough set and knearest neighbour', International Journal of Computing and ICT Research, 2008, Vol. 2, pp.60–66.
35. RamannaS, PetersJF, SurajZ, ShanS and N. Pizzi, "Classification of meteorological volumetric radar data using rough set methods", Pattern Recognition Letters 24 (6) (2003) 911–920.
36. Brodley CE and Dy J.G, "Feature selection for unsupervised learning", The Journal of Machine Learning Research archive 5, 2004, 845–889.
37. KusiakA, "Decomposition in data mining: An industrial case study", IEEE Transaction on Electronics Packaging Manufacturing 23 (4), 2000 345–353.
38. ChanCC, "A rough set approach to attribute generalization in data mining", Information Sciences 107, 1998, 177–194.
39. Xu Y & T Li, "A generalization rough set approach to attribute generalization in data mining", Journal of Southwest Jiaotong University 8 (1), 2000 69–75.
40. KamberM & JHan "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco, 2001,CA.
41. Huang J, Nie P, SaiY, and Xu R 'A rough set approach to mining concise rules from inconsistent data', Proceedings of IEEE GRC 2006, IEEE, Atlanta USA, pp.333–336.
42. R Brause and PaetzJ 'Rule generation and model selection used for medical diagnosis', Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology – Challenges for future intelligent systems in biomedicine, 2002 Vol. 12, No. 1, pp.69–78.

43. AAbraham & C Grosan 'Stock market modeling using genetic programming ensembles', Genetic Systems Programming: Theory and Experiences, 2006,Vol. 13, pp.133–148.
44. BT Nath & A Ghosh 'Multi-objective rule mining using genetic algorithms', Information Sciences: an International Journal, 2004, Vol. 163, pp.123–133.
45. Ma HF, ZhaoWZ and HeQ, "Parallel k-means clustering based on mapreduce," in Proc. 1st Int. Conf. Cloud Comput., 2009, pp. 674–679.
46. Wang GY, LiangZ,YangY and ChenRZ "Attribute reduction for massive data based on rough set theory and mapreduce," in Proc. Rough Set Knowl. Technol., 2010, pp. 672–678.
47. ZhaoCB, ZhangBJ, RTLi, Ruan, and GaoZ, "A parallel method for computing rough set approximations," Inf. Sci., vol. 194, pp. 209–223, 2012.
48. GhemawatS and DeanJ "Mapreduce: simplified data processing on large clusters", Commun. ACM 51 (1) 2008, 107–113.
49. Hadoop Apache-2013, "Hadoop Apache Project", <http://hadoop.apache.org/> accessed December-2013.
50. Hadoop, WhiteT, "The Definitive Guide", O'Reilly Media, Inc., 2012.
51. Apache Mahout-2013 "Apache Project Mahout", <http://mahout.apache.org/> accessed December- 2013.
52. DunningT, OwenS, AnilR, FriedmanE, "Mahout in Action", Manning Publications Company, 2012.
53. Spark-2013 <http://spark-project.org/> accessed December-2013.
54. Storm-2013 <http://storm-project.net/>accessed December-2013.
55. Romer.G, S.Melnik, J.Long,A.Gubarev, ShivakumarS, and Vassilakis DremelT, MTolton,: interactive analysis of web-scale datasets, in: Proc. of the 36th International Conference on Very Large Data Bases, 2010, pp. 330–339.
56. Apache Drill-2013 <http://incubator.apache.org/drill/> accessed December-13.
57. Dean J, GSanjay, "MapReduce: Simple Data Processing on Large Clusters" Proceedings To appear in OSDI 2004.
58. Zdzisław Pawlak, Andrzej S "Rudiments of rough sets" Proc. Elsevier accepted in 7 June 2006.
59. T Li, Zhang J, pan Yi "Parallel Rough Set Based Knowledge Acquisition Using MapReduce from Big Data" Proceedings ACM Big Mine, August 12th , 2012 Beijing, China.
60. J Zang, T Li, Da Rausan "A parallel method for rough set approximations" Proc. Elsevier accepted in 11 January 2012.
61. B.Raghuram, and Gyani J "Fuzzy Associative Classification Driven MapReduce Computing Solution for Effective Learning from Uncertain and Dynamic Big Data" International Journal of Database Theory and Application, Vol. 11, No.1 2018.
62. Data Set (URL): https:/archive.ics.uci.edu/ml/datasets/census+income.

AUTHORS PROFILE

Mr. Hanumanthu Bhukya is a full-time research fellow at the CSE Department, the University College of Engineering & Technology, Kakatiya University, Warangal, TS, and India. He published above 18 papers in peer-reviewed International journals and conferences. His research focuses on data analysis, web security and information security. He's an ISTE fellow.

Dr. Sadanandam M received his PhD degree from JNTU,Hyderabad, in 2013. He is currently working as Assistant Peofessor of CSE, Kakatiya University, Warangal. His research focuses on Speech Recognition and Processing, Image Processing, Data Analytics and Pattern Recognition. He published above 35 papers have been published in peer-reviewed IEEE and ACM journals and conferences. He served as HoD of Computer Science and Engineering, Chairman of BoS for CSE & IT, Joint Director for KU Engineering Hostels and Placement Officer for University College of Engineering, Kothagudem. He is also a member of IEEE (Indian Society for Technical Education), Member of the Institute of Engineers (FIE), Member of CSI, IAENG.

ACKNOWLEDGMENT

Our thanks to the members of the management and Principal of the Kakatiya Institute of Technology and Science-Warangal who facilitated reading and computing resources to develop this model and to narrate this paper. And our sincere thanks to Mr. B.RghuRam,Assistant Professor,CSE,KITSW for helping to bring this paper to successful conclusion, and also our gratitude to our beloved Principal Prof. P.Malla Reddy, UCE&T, Kakatiya University, who entertained us to research and publish this paper.