International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 10472-10476


# Mining Association Rules from No-SQL data bases using Map-Reduce Fuzzy Association Rule Mining Algorithm

**Chatakunta Praveen Kumar [1], Pole Anjaiah[2], Santosh Patil, Ediga Lingappa[3] and Mothe Rakesh[4]**
[1, 2, 3 ,4, 5] *Assistant Professor, Department of Computer Science and Engineering, Institute Of Aeronautical Engineering, Jawaharlal Nehru Outer Ring Road, Dundigal, Hyderabad, Telangana 500043, India.*
[1]*Orcid: 0000-0002-6391-7970*

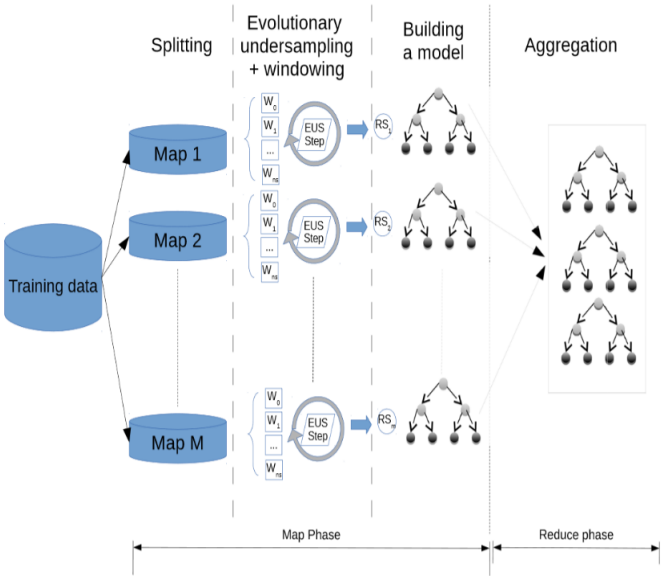
**Abstract**

Big Data trepidations enormous tome, intricate, emergent records sets with several, self-directed bases. With the fast growth of networking, data storage, and the data pool capacity, Big Data is now quickly escalating in all learning and engineering purviews, including physical, organic and biomedical sciences. In order to get use of these huge and heterogeneous data, data mining is a technique to get utilization of such data. Data mining is also called as facts discovery in databases which is the non-trivial process of detecting the valid, novel, hypothetically beneficial and eventually comprehensible acquaintance in outsized scale data. But the existing data mining techniques are not suitable for such heterogeneous data. In this paper we are proposing a map-reduced based fuzzy based data mining technique.

**Keywords:** big data, data mining, fuzzy, Map, Reduce

## INTRODUCTION

21st century is an era of data evaluation, big data is definitely another milestone in all the aspects of 21st century people's daily activities. Big data sets can benefit us a lot in different aspects, for illustration, medical, banking, IT field and so on. In order to get benefit of these data we need to mine the patterns from the big data, in order to mine patterns from the data fuzzy based approaches gives better flexibility. A large portion of the strategies and calculations in the field of affiliation decide mining expect that the trait esteems are Boolean in nature. These algorithms consider only the presence or absence of items in a transaction. Important factors like number of items in a transaction or the profit of items are not taken into consideration for finding associations between items. Weighted fuzzy association rule mining techniques are capable of considering these factors while forming association rules. The Boolean association rule mining techniques cannot include quantitative or categorical attributes for processing. Converting to equal intervals or Clustering can be used with quantifiable aspects to segment these morals into groups having akin properties. The old-fashioned clustering strategies have a constraint that they require the client to determine the quantity of groups as an underlying parameter. Discovering this number stays as a dreary errand. The unsupervised grouping calculations are most suited for recognizing normal bunches from informational index without giving any underlying parameters.




The fuzzy threshold based unsubstantiated cluster estimation algorithm, which is described in the last chapter, is capable of converging to a finest total of clusters. In this paper we are proposing a fuzzy based rule mining algorithm to get the patterns from big data. To mine big data with the fuzzy algorithm we use map-reduce frame work for parallel processing of the data on multiple number of nodes which helps to improve the performance of the proposed method and handles the hetroginious data. The rest of the paper is as follows section-2 describes the different fuzzy mining techniques and section-3 illustrate the proposed method, section-4 deals with the performance of the proposed method and finally section concludes the paper.

## METHODS

Datamining is the analysis of vast quantity of data to discover hidden information. Dynamic inquire about on Datamining has been continuing for quite a while. There are numerous information mining techniques or calculations that exist for mining information to get designs. Here we exhibit a survey

of various fluffy related techniques.

Mining Fuzzy Frequent point set utilizing Compact Frequent Pattern(CFP) tree Algorithm K.SuriyaPrabha, R.Lawrance,et. al. composed a novel strategy for era of solid run the show. The proposed development calculation for building a Fuzzy CFP tree from a quantitative database is depicted in this area. The proposed approach coordinates the fluffy set ideas and the variety of the exemplary FP-tree-like way to deal with productively locate the fluffy incessant itemsets from the quantitative exchanges. The Fuzzy FP-tree development calculation is first intended to assemble the tree structure for the fluffy regular 1-itemsets. Every hub in the tree structure keeps a fluffy continuous 1-itemset, its participation esteem, and the enrollment estimations of its super-itemsets in the way as indicated by the crossing point administrator, which is the base administrator here. this proposed work coordinates the fluffy set ideas in the recently proposed CFP-tree calculation by developing a minimal sub-tree for a fluffy successive thing, creating hopefuls in bunch from the smaller sub-tree and later discharge the ebb and flow subtree from memory leaving the space for next subtree accordingly essentially beats alternate calculations on both execution times, memory utilizations and lessening the pursuit space at long last bringing about the revelation of fluffy incessant itemsets.[2]

Fleecy Weighted Associative Classifier: A Predictive Technique for Health Care Data Mining Sunita Soni and O.P.Vyas et al. made an estimation for fleecy weighted alliance course of action in which they expand the issue of request using Fuzzy Association Rule Mining and propose the possibility of Fuzzy Weighted Associative Classifier. Characterization in perspective of Association rules is believed to be effective and gainful all around. Partnered classifiers are especially fit to applications where the model may help the region authorities in their decisions. Weighted Associative Classifiers that endeavors weighted Association Rule Mining is starting at now being proposed. Regardless, there is in like manner called "sharp farthest point" issue in connection rules mining with quantitative trademark spaces. This paper proposes another Fuzzy Weighted Associative Classifier that produces portrayal rules using Fuzzy Weighted Support and Confidence structure. By then aïve approach can be usual to delivering strong rules as opposed to weak irrelevant standards. Where cushy method of reasoning is used as a part of dividing the domains.The issue of Invalidation of Downward Closure property is understood and the idea of Fuzzy Weighted Support and FuzzyWeighted Confidence system for Boolean and quantitative thing with weighted setting is summed up [3].

Visit Item sets from Multiple Datasets with Fuzzy information Praveen Arora, R. K. Chauhan and Ashwani Kush et.al.proposed a Traditional methodologies handles fresh and fluffy information extremely well however less distributed outcomes demonstrate that databases that contain various tables with fluffy information having scientific categorization can be dealt with productively. The Proposed calculation is found by broadening these customary calculations and finds the multi level fluffy affiliation controls in Entity – Relationship displayed databases, which is able to deal with different tables. The investigation dissects how the properties of a few elements seem together. The Study likewise investigates the principles as for the connections existing between the substances and their predecessors. On the off chance that few connections exist between at least two substances, at that point the fluffy affiliation administers between their qualities and predecessors are analyzed regarding each such relationship. The found calculation utilizes the join and substance bolsters in deciding continuous thing sets. By considering the substance bolster it doesn't dispose of from the outcome element thing sets that are visit regarding their element table yet not as for the relationship table and it likewise permits the calculation of right help and certainty for rules existing among qualities of a similar element table [4].

An Enriched System for Mining Association Rules in Large Databases Farah Hanna AL-Zawaidah, YosefHasanJbara and Marwan AL-Abed Abu-Zanona et. al. show a novel connection choose mining approach that can adequately discover the alliance controls in tremendous databases. The expected technique is gotten from the standard Apriori approach with features added to upgrade data mining execution. They had performed wide investigations and differentiated the execution of the computation and existing figurings found in the written work. Trial comes to fruition show that the approach beats distinctive philosophies and exhibit that approach can quickly discover visit itemsets and effectively mine potential connection rules.

An Algorithm For Mining Fuzzy Association Rules Reza Sheibani , Amir Ebrahimzadeh ,Member, IAUM presents a paper , in this paper, we presentan effective calculation named Fuzzy Cluster-Based AssociationRules. The FCBAR strategy is to make group tables by checking thedatabase once, and after that bunching the exchange records tothe k_th bunch table, where the length of a record is k.Moreover, the fluffy huge itemsets are created by contrastswith the fractional group tables. This prunes considerableamount of information, decreases the time expected to perform information scansand requires less difference. Examinations with the genuine lifedatabase demonstrate that FCBAR outflanks fluffy Apriori_likealgorithm , a well– known and broadly utilized affiliation rulesalgorithm. In this paper we proposed the productive calculation for mining fluffy affiliation rules. The FCBAR calculation makes group table to help revelation of fluffy vast itemsets. Complexities are performed just against the halfway bunch tables that were made ahead of time. Investigations with the genuine database demonstrate that FCBAR beats Apriori_like calculation, a notable and broadly utilized affiliation rule.[6]

Proficient Parallel Pruning of Associative Rules with Optimized Search The principle center of this examination work is to propose an enhanced affiliation control mining calculation to limit the quantity of applicant sets while producing affiliation rules with effective pruning time and inquiry space advancement. The relative relationship with decreased applicant thing set lessens the general execution time. The versatility of this work is measured with number of itemsets utilized as a part of the exchange and size of the informational collection. Assist Fuzzy based lead mining rule is adjusted in this work to acquire more useful cooperative principles and regular things with expanded touchy. The prerequisite for touchy things is to have a semantic association between the parts of the thing esteem sets.

FPrep: Fuzzy Clustering driven Efficient Automated Pre-preparing for Fuzzy Association Rule Mining Ashish Mangalampalli, Vikram Pudi proposed the technique for preprocessing of Fuzzy Association Rule Mining. This paper portrays an approach, called FPrep, to do this pre-preparing, which initially includes utilizing fluffy bunching to produce fluffy parcels, and after that uses these allotments to get a fluffy variant (with fluffy records) of the first dataset. At last, the fluffy information (fluffy records) are spoken to in a standard way with the end goal that they can be utilized as contribution to any sort of fluffy ARM calculation, independent of how it works and procedures fluffy information. We additionally demonstrate that FPrep is significantly quicker than other such equivalent change methods, which thus rely upon non-fluffy strategies.

**PROPOSED METHOD:**

Fuzzy association rule mining (FAR) algorithm

Input: S, a binary database; minsup, a pre-defined minimum support threshold; minconf, a pre-defined minimum confidence threshold, MFs, a pre-defined membership function

Output: A set of fuzzy association rules

1 for each transaction $tr_i$ in S do

2 for each item(attribute) $D_j$ do

3 convert quantity $q_{ij}$ by MFs as $(f_{ij}/D_j.R_1+f_{ijn}/D_j.R_2+\ldots..+f_{ijn}/D_j.R_n)$

4. end for

5. end for

6. calculate count($D_j.R_k$):=sum{$f_{ijk}$}.

7. Set MAXCount($D_j.R_k$):=max{count($D_j.R_k$)}.

8. set $L_1$<-{$D_j.R_k$| MAXCount($D_j.R_k$)>=minsup * |S|}.

9. set r:=2.

10. while $L_{r-1}$!=null do

11. set $C_r$<-{a U b|a,b ∈ $L_{r-1}$, a !E b}.

12. for each transaction $tr_i$ in S do

13. set $C_{ij}$<-{Z|Z E $C_r$ ^ z <= $tr_i$}.

14. for each z E $C_{ti}$ do

15. calculate count ($tr_i$. z) = {min($f_{ijx}$, $f_{ijy}$)|x,y E z, x!E y}.

16. end for

17. end for

18. calculate count(z):= sum{count($tr_i$.z)}

19. set $L_r$<-{z|count(z)>=minsup * |S|}.

20. r:=r+1

21. end while

22. set $C_{FARS}$<-{$L_1$^$L_2$^….^$L_r$->$L_q$|q = 1 to r}.

23. for each w E $C_{FARS}$ do

24. calculate conf(w).

25. set $FAR_s$<-{w|conf(w)>=minconf * |S|}

26. end for

27. return $FAR_s$.

In the FDTA calculation, the amount estimation of every thing (trait) of every exchange in the databases is first changed over in view of the given participation capacities (Lines 1– 5). The enrollment estimations of a semantic term of every thing (Dj.Rk) are summed together in the database (Line 6). From that point forward, the phonetic term of every thing with the most extreme cardinality is found to speak to this thing for later mining process (Line 7). In the event that the spoke to cardinality of the spoke to etymological term is at the very least the given least help tally, it is then put into the arrangement of fluffy continuous (extensive) itemsets (Line 8). From that point forward, the Apriori-like system is performed to locate the fluffy successive itemsets (Lines 10– 21). This procedure is rehashed until no competitor itemsets are produced. The found fluffy regular itemsets are then framed to figure their certainty for initiating fluffy affiliation rules (Lines 22– 27). Since every thing utilizes just a phonetic term with the most extreme cardinality, the quantity of things in the first database is the same as the uncovered etymological terms. Accordingly, the combinational blast issue for mining fluffy affiliation principles can be maintained a strategic distance from.

**MRFAR Algorithm:**

**Map step:**

1. Separately node applies the map() function to the native data,
2. applies the Fuzzy association rule algorithm on each node
3. writes the output to a momentary storage.
4. A controlling node ensures that only one copy of dismissed input data is administered.

**Shuffle step:**

1 Operative hubs redistribute information in light of the yield keys (delivered by the "guide ()" work)

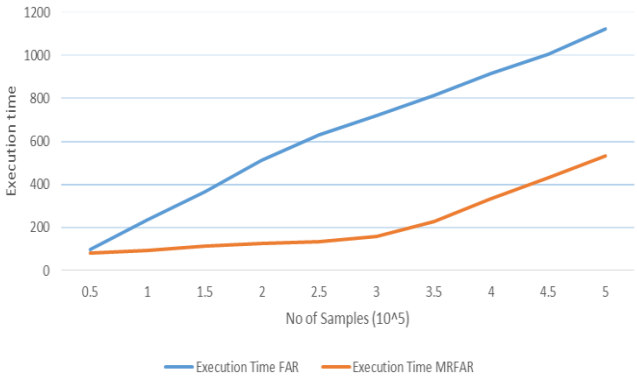
2 to such an extent that all information having a place with one key is situated on a similar laborer hub.

**Reduce step:**

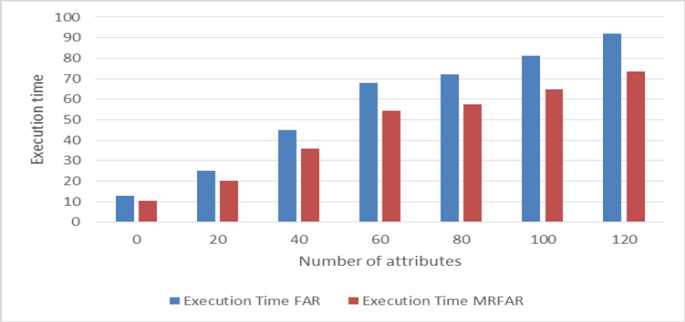
1 Operative nodes these days practise each group of output data, per key, in equivalent.

**Performance of proposed method:**

In edict to appraise whether the anticipated Map-reduced fuzzy association rule mining (MRFAR) algorithm, organisation might switch the large gage pool of heterogeneous data, we first adapt it with MapR_educe context called MRFAR, to improve the processing speed. The twitter test dataset utilized as a part of the analysis is a 1% inspecting gathered from Twitter.com by a planned web crawler framework. It consolidates each one of the tweets from September 2016 to March 2017, which is monstrous data contains more than 220 billion printed records, and set away it in Mongo-DB, which is a "No-SQL" open premise databank in order to better setting up the colossal data. In MongoDB, the data is secured in the JSON-style which offers both ease and control, and has expansive qualities for record support. The earth used for the examination is 4 PCs with the CPU of Intel i5 QM and 8GB of Random access memorey the OS Ubuntu16.04 X86/64 GNU/Linux; and Java jdk 1.7.0_23. The Hadoop bundle contains four Node Data's with a consider exchange speed for end-to-end TCP connections up to 200 MB/s, and for every Node Datas the best number of strings is 10.




In fig-2 shows the comparative execution time performance of FAR and MRFAR algorithms in mining different sample sizes. Here in FAR takes very huge amount of time to compute. But MRFAR shows very fast accessing because of map reduce processing, map-reduce algorithm do parallel processing of the DB that will reduces execution time in a drastically manner.




In fig-3 shows the comparative execution time performance of FAR and MRFAR algorithms in mining different attributes. Here in FAR takes very huge amount of time to compute. But MRFAR shows very fast accessing because of map reduce processing, map-reduce algorithm do parallel processing of the DB that will reduces execution time in a drastically manner.

**CONCLUSION**

Conventional rule mining techniques, are typically precise, yet have hard and delicate operations. Fluffy Based calculations then again give a powerful and effective way to deal with investigate substantial pursuit space. As of late various works have been completed utilizing Fuzzy calculation for mining affiliation rules. As many works have been carried out on mining association rules with Fuzzy algorithms this paper we presented a MRFAR algorithm which works with parallel processing of large amount of heterogeneous data and gives accurate results and also the results shows that it computes very fast manner than compared to the traditional methods.

**REFERENCES**

[1] Han, J., Kamber, M. (2001). "Data Mining: Concepts and Techniques", Harcourt India Pvt. Ltd.

[2] K.SuriyaPrabha, R.Lawrance," Mining Fuzzy Frequent itemset using Compact Frequent Pattern(CFP) tree Algorithm" International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[3] SunitaSoni, O.P.Vyas,"Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012.

[4] Praveen Arora, R. K. Chauhan and AshwaniKush ," Frequent Itemsets from Multiple Datasets with Fuzzy data", International Journal of Computer Theory and Engineering, Vol. 3, No. 2, April 2011.

[5] Farah Hanna AL-Zawaidah, YosefHasanJbara and Marwan AL-Abed Abu-Zanona, "An Improved Algorithm for Mining Association Rules in Large Databases", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 2011, pp. 311-316.

[6] Reza Sheibani, Amir Ebrahimzadeh ,Member, IAUM," An Algorithm For Mining FuzzyAssociation Rules", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I, March, 2008,pp.486-490.

[7] K.Sangeetha,Dr.P.S.Periasamy , S.Prakash,"Efficient Parallel Pruning of Associative Rules with Optimized Search", IOSR Journal of Computer Engineering (IOSRJCE) ,volume no.3,pp.26-30.

[8] AshishMangalampalli, VikramPudi,"FPrep: Fuzzy Clustering driven Efficient Automated Pre-processing for Fuzzy Association Rule Mining", IEEE Intl Conference on Fuzzy Systems (FUZZ-IEEE), July 2010. [9] G Vijay Krishna,PRadhaKrishna,"A Novel Approach for Statistical and Fuzzy association Rule Mining on Quantiative Data ",Journal of scientific and industrial Reasearch ,vol no.67,jul2008,pp.512-517.

[9] Centola D. 2010, The spread of behavior in an online social network experiment, Science, vol.329, pp.1194-1197.

[10] Chang et al., 2009, Chang E.Y., Bai H., and Zhu K., Parallel algorithms for mining large-scale richmedia data, In: Proceedings of the 17th ACM International Conference on Multimedia (MM '09), New York, NY, USA, 2009, pp. 917-918.

[11] Chen et al. 2004, R. Chen, K. Sivakumar, and H. Kargupta, Collective Mining of Bayesian Networks from Distributed Heterogeneous Data, Knowledge and Information Systems, 6(2):164-187, 2004.

[12] Chen et al. 2012, Yi-Cheng Chen, Wen-Chih Peng, Suh-Yin Lee, Efficient algorithms for influence maximization in social networks, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 577-601

[13] Chu et al., 2006, Chu C.T., Kim S.K., Lin Y.A., Yu Y., Bradski G.R., Ng A.Y., Olukotun K., Mapreduce for machine learning on multicore, In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS '06), MIT Press, 2006, pp. 281-288.

[14] Cormode G. and Srivastava D. 2009, Anonymized Data: Generation, Models, Usage, in Proc. Of SIGMOD, 2009. pp. 1015-1018.

[15] Das et al., 2010, Das S., Sismanis Y., Beyer K.S., Gemulla R., Haas P.J., McPherson J., Ricardo: Integrating R and Hadoop, In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10), 2010, pp. 987-998.

[16] Dewdney P., Hall P., Schilizzi R., and Lazio J. 2009, The square kilometre Array, Proc. of IEEE, vol.97, no.8.

[17] Domingos and Hulten, 2000, Domingos P. and Hulten G., Mining high-speed data streams, In: Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD□00), 2000, pp. 71-80.

[18] Duncan G. 2007, Privacy by design, Science, vol. 317, pp.1178-1179. 20) Efron B. 1994, Missing data, imputation, and the Bootstrap, Journal of the American Statistical Association, vol.89, no.426, pp.463-475.

[19] Ghoting et al., 2009, Ghoting A., Pednault E., Hadoop-ML: An infrastructure for the rapid implementation of parallel reusable analytics, In: Proceedinds of the Large-Scale Machine Learning: Parallelism and Massive Datasets Workshop (NIPS-2009).