International Journal of Database Theory and Application
Vol. 11, No.1 (2018), pp. 9-20
hhtp://dx.doi.org/10.14257/ijdta.2018.11.1.02

Fuzzy Associative Classification Driven MapReduce Computing Solution for Effective Learning from Uncertain and Dynamic Big Data

Raghuram Bhukya[1] and Jayadev Gyani[2]

[1]Assistant Professor, Department of Computer Science and Engineering
Kakatiya Institute of Technology & Science, Warangal, India
[2]Assistant Professor, Computer Science and Information Technology College
Majmaah University, Majmaah, Kingdom of Saudi Arabia
[1]raghu9b.naik@gmail.com, [2]jayadevgyani@yahoo.com

Abstract

Handling uncertainty and dynamic changes in data sets supposed for analysis is always been challenging task for data analytics community. The same challenges even inherited to the embryonic big data analytics which are generally mentioned as veracity and velocity properties. Indeed, in case of big data, handling uncertainty and dynamism data could be more typical because of the scalability factor which is a result of data storage in distributed file system structure. In order to overcome difficulties of handling uncertainty and dynamic changes in big data analytics and considering efficiency provided by fuzzy associative classification techniques in handling uncertainty of data, we propose a dynamically scalable fuzzy associative classifier extraction model for Map-reduce framework. The important contributions of the paper is that it proposes a data driven fuzzy generalization approach for handling uncertainty, Tid-list based classification approach for easy scalable computation among multiple nodes and data chunks driven updating of classification model in case of dynamic changes to dataset with respect Map-reduce framework. The experimental evaluation results shows that the proposed Map-reduce model for fuzzy associative classification rule extraction can efficiently handle data uncertainty and dynamic changes to data stored in distributed file system, along with satisfying scalability factor.

Keywords: *Fuzzy associative classification, MapReduce, Data Veracity, Dynamic data*

1. Introduction

The digital data floating from various fields including e-commerce, social media, health and economic industry *etc.*, is continuously throwing challenges to knowledge processing community for providing effective storing and analytical techniques. The handling of this immense data resulted from human cognitive process is currently referred as big data analytics [1]. The primary challenge in big data analytics is handling scalability of ever increasing data. This scalability can be offered by distributed file system.

The distributed file system offers scalability by dividing the data into chunks and storing the chunks into disk and whenever the present disk is full the next chunk is stored in another disk. The other advantage of the distributed file system is its reliability which offered by storing same chunk of data into multiple disks by which even if a disk is get failed it can recover data from other disks to ensure high reliability. Out of different open source distributed file systems available Hadoop Distributed File System (HDFS) [2]

Received (December 29, 2017), Review Result (February 28, 2018), Accepted (March 15, 2018)

offered by Apache is shown its significance by huge number of applications and adaptation by data service provider. The flexibility of implementing on common service hardware and incorporating MapReduce frame work are primary reasons for success of HDFS.

MapReduce [3] is computing framework introduced by Google inc., for distributed processing of data files. The effectiveness of MapReduce is derived from its efficiency in scaling the parallel computation on large number of commodity systems. The MapReduce computing performs parallel processing in Map and Reducer phase. Where the Map phase perform local processing of data stored at local files and Reducer phase perform consolidated processing of global data gathered form Map phase. The MapReduce computing in-order to organize communication between local and global processing phases make uses of the <key,value> pair where Map phase takes <key, value> as input and generates intermediate <key,value> which will be used by Reducer phase. The dataset domain where ever MapReduce framework supposed to apply the programming should organize according <key,valure> pair methodology. The smooth transition and computation paradigm offered by <key,value> model making knowledge discovery filed to scale MapReduce framework for huge volumes of data processing. Along with HDFS the MapReduce is also being used in other specialized distributed processing applications like Apache Mahout, Apache Spark, MangoDB, Hive *etc*.

Machine learning with synonym of data mining is been part of human digital life with scientific and commercial fields including e-business, health care industry, Internet of the Things applications, home security systems *etc.*, [4]. The machine learning can be realized using techniques including association rule mining, classification, clustering and time series analysis. Out of this different technique classification found a major application in distributed learning with applications such as machine learning, recommender systems, targeted campaign, credit risk management, medical diagnosis *etc*. Considering real time significance of classification there are proposal in literature which can perform distributed learning from distributed file systems using classification techniques random forest, Ada boosting, SVM and rule base models [5]. Out different classification techniques scaled on distributed file system associative classification techniques shown their worth in accuracy and easy adaptation to different types of datasets [6]. The worthiness of associative classification technique derived from its adaptation of frequent item sets extraction methods for extracting relationship between attribute item sets and class label. Because of these exhaustive exploration associative classification techniques offers most accurate and easily interpretable results than any other classification technique.

In performing associative classification over big data stored in distributed file system the primary challenge is addressing voluminous nature of dataset. To address this problem there are proposal in literature which adopts scalability principle in conventional mining algorithms [7] to successfully drag the conventional mining algorithm on MapReduce architecture and success fully address the huge volume challenge of big data. At the same time one can notice most of the proposal in literature ignore the tow important birthright problems of big data those are uncertainty and dynamic increments in data.

Data uncertainty which present as unrelated or over fitted data values in big data sets are result of data collection from heterogeneous and unanimously situated sources in huge quantity. In the present scenario of Big data the uncertainty of data is defined as Veracity property out of popular 4-V properties. Mainly the data uncertainty can found in structural and unstructured data sources which are collection of Sensor networks, social networks and web sites [8]. In current research most of the proposals for Bigdata analytics concentrate on parallelizing conventional mining algorithms for MapReduce framework in order to handle its volume but at the same time they ignore Veracity properly.

In research literature we can found many strategies to efficiently handle or mitigate the uncertainty in data so that effective information can retrieve out of it [9]. Out of different statistical methods the fuzzy logic methods shows its significance by reflecting human perspective in handling uncertainty of data. By using fuzzy logic [10] uncertainty can be handled by mapping the data to human perspective label with mapping degree in between 0 to 1 which reflects the relatedness to respective label.

The other property of Bigdata which overlooked by many of the mining technique is velocity [11]. That is in real time big data applications data will dynamically incremented for the purpose of effectiveness in analytics. The situation will be worse with distributed file systems. There are effective data mining models like erasable item set mining [12] for incremental data updates and single scan to handle dynamic increment as well decrements [13] in data mining literature, but these models need to explore on MapReduce framework in order to handle ever scaling distributed file systems.

Considering real time importance of handling data uncertainty and dynamic increments in big data we propose fuzzy associative classification driven MapReduce model for effective learning satisfying velocity and veracity properties of big data.

2. Research Literature

In the resent research literature we can found various conventional machine learning algorithms extend to perform analytics over Bigdata. Even though the most common factor among these algorithms are MapReduce paradigm they use different computing models including Apache Spark and Mahout [14] for conventional data processing in quick mode, CGL-MapReduce [15] for handling data streams, Amazon EC2 [16] for processing in cloud environment *etc.* Using these computing models, in order to answer the voluminous properties of Bigdata we can found machine learning algorithms for frequent pattern extraction [17], classification [18], and clustering [19] algorithms extended on MapReduce computing paradigm.

Out of different classification models associative classification plays a vital role in learning and predicting application because of its accuracy and interpretability. Considering the importance of associative classification in real time applications we can found proposals to perform associative classification over big data which includes Tid-list based [20] based conventional associative classification model, and intuitive fuzzy associative classification model [21]. Even though these associative classification models for MapReduce computing paradigm successfully answer the voluminous property of big data they fail to answer the Veracity and Velocity properties of Bigdata. Considering this drawback this work attempts to propose associative classification model for MapReduce computing paradigm that can handle Veracity and velocity properties of Bigdata.

Veracity is the form of uncertainty which may happen due to less quality sensors, aggregated data for the purpose of privacy, usage of statically wrapped variables etc. In literature we can found machine learning applications for uncertain data including frequent patter mining from uncertain data [22], classification of uncertain data objects [23], clustering uncertain dataset [24], and outliner analysis of dataset [25]. Even though there are ample number of models to explore uncertain data there is lack of models to handle data uncertainty over MapReduce applications.

Dynamic data mining [26] which deals with efficient handling of increment and decrements of data is been a major research interest of knowledge engineering community. One can found dynamic data mining applications in all fields of data mining including association rule extraction [27], classification [28], and clustering [29]. Out of these various kinds of applications of dynamic data mining we can found proposals evolved for classification applications including incremental decision trees [30] and incremental concept induction [31]. Specific to associative classification there are applications called Associative Classification based on Incremental Mining [32] and incremental classification application for intruder detection [33] and incremental associative classification for distributed databases [34] but best of our knowledge there is no proposal which can perform incremental associative classification over MapReduce framework.

Considering the explored things mentioned in above literature survey *i.e.*, the lack of MapReduce base learning techniques satisfying uncertainty and dynamic nature of big data, this work proposes fuzzy associative classification technique for MapReduce framework for effective learning satisfying velocity and veracity properties of big data.

3. Fuzzy Associative Classifier Extraction from Uncertain Bigdata

The typical MapReduce structure for processing Distributed File System (DFS) is shown in the Figure 1. The same architecture is adopted for the proposed fuzzy associative classification extraction from uncertain big data. In a typical MapReduce framework the uncertain and ever incrementing data collected from heterogeneous sources will be stored in DFS which is popularly available as Hadoop distributed file system (HDFS). In HDFS there is a master node where the data supposed to be stored will be submitted which will store in 3 replicated copies in distributed mode. The applications of HDFS adopt the MapReduce computing paradigm for computation where the Map nodes will access the disks and process at local level in parallel. The local data processed at local level will consolidated at one or more Reducer nodes.

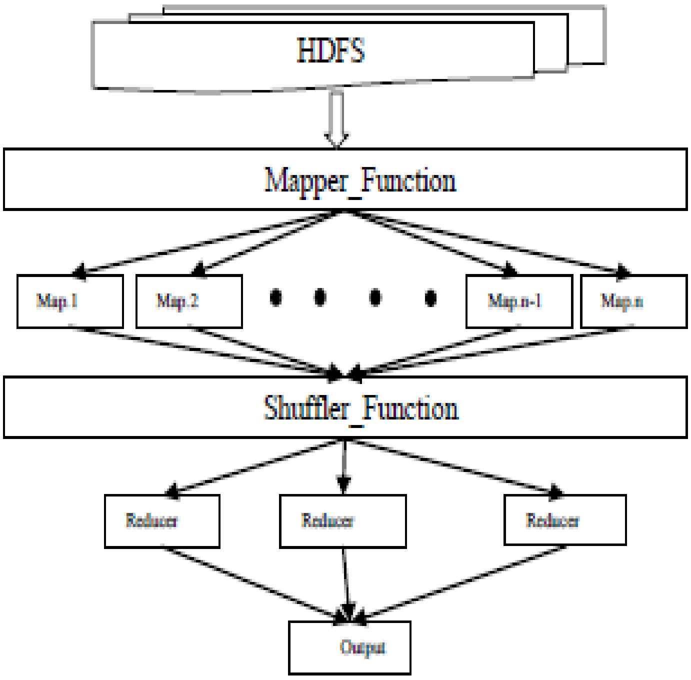

Figure 1. MapReduce Architecture for Processing Distributed File System

In order to realize effective learning rules generation from Bigdata satisfying veracity and velocity properties on MapReduce framework we propose 3-JOBs of MapReduce computation model (Dynamic MR_FAC) which includes

MRJob_1. Fuzzy associative classification rules generation by tackling uncertainty of data.

MRJob_2. Updating existing classification model when new data added to existing Bigdata.

MRJob_3. Efficiency calculation of test rules.

3.1. Fuzzy Classification Rule Generation from Uncertain Data on MapReduce Computing Paradigm

In order to overcome the short comes introduced by data uncertainty result of data extraction from heterogeneous sources we adopt fuzzy logic model which instead of crisply deciding class relatedness of an attribute value, it will generate value among 0 to 1 to show relatedness of variable value to class label. So that the data evaluator can experiment or decide about the degree of the variable to which level it can be considered for generating classification rules.

If an attribute value $x_i \in A_c$ can partition into fuzzy sets $L_j = \{L_1 \ldots L_k\}$ with a specific membership value given by attributes fuzzy membership function μ_c. where $L_j(x_i) = \mu_c\left(x_i^j\right)$ should satisfy equation (1).

$$0 \leqq \mu_c\left(x_i^j\right) \geqq 1 \;\&\; \sum_{j=1}^{k} \mu_c\left(x_i^j\right) = 1 \tag{1}$$

If an attribute Ac is categorical then the every unique value $x_i \in Ac$ will be considered as a fuzzy partition and $\mu_c(x_i) = 0$ or 1.

So by using fuzzy logic base class mapping given in equation (1) the level of uncertainty of a variable with respect to label can be decided. So then the evaluator can use threshold value between 0 to 1 to keep or sack the particular variable in taking business decisions. But the introduction of the fuzzy logic will make the basic changes in generating class label based association rules which is presented in following paragraphs.

The fuzzy support for a fuzzy set L_i with respect to a distributed data sub set D_y with n number transactions $\{t_1 \ldots t_n\}$ is given by equation (2).

$$FS_y(L_1, L_2) = \frac{\sum_{j=1}^{n} Min(L_1(t_j), L_2\,(t_j\,))}{n} \tag{2}$$

The fuzzy associative classification rule extraction is special case of fuzzy association rule extraction. The class label based fuzzy generalized set support calculation process the equation (2) will be updated as equation (3).

$$\mathrm{FS_y}(\mathrm{L_1}, \mathrm{L_2} \rightarrow \mathrm{c}) = \frac{\sum_{\mathrm{j=1\&t_j[A_c]=c}}^{\mathrm{n}} \mathrm{Min}(\mathrm{L_1}(\mathrm{t_j}), \mathrm{L_2}\,(\mathrm{t_j}\,))}{\mathrm{n}} \tag{3}$$

In case of distributed environment where data set D fragmented horizontally among N number of systems then the global fuzzy support of a class label based fuzzy generalized set $L \rightarrow C$ is calculated with equation (4).

$$\text{Global\_FS (L}\rightarrow\text{C)} = \frac{\sum_{y=0}^{N} FS_y(L \rightarrow c)}{N} \tag{4}$$

The fuzzy global confidence of the fuzzy generalized associative classification rule L→C is calculated by equation (5).

$$\text{FC (L}\rightarrow\text{C)} = \text{Global\_FS(L} \rightarrow \text{C)} \big/ \text{Global\_FS(L)} \tag{5}$$

Using the equation (4) and (5) the calculation can be performed for generating fuzzy class label based association rules. But considering target generating uncertain classification rules from big data using MapReduce framework the work proposes Algorithm-1 to fulfill the task.

Algorithm-1. Fuzzy class label base association rule generation using MapReduce framework for uncertain data.

```
Input:  Data set loaded in HDFS
Output:  Global Tid-list and classification rules
Map():
S1. Read the data chunk assigned
S2. for each transaction $T_i$ do
S3. Generate fuzzy membership values of each attribute using equ.1
S4. Create a new row in Tid-list representing $T_i$
S5. If combination of attribute value and class label of record $T_i$ is not present
    Create same as a new column label in Tid-list
S6. Else If combination of attribute value and class label is present in $T_i$ then
    store fuzzy value at corresponding entry
S7. Else store value 0.
S8. End

Reducer():
S1. Read all local Tid-list
S2. merge_Tid-list()
S3. create separate columns for each uniq attribute labels in Global_Tid-list
S4. create separate rows for all transactions in Global_Tid-list
S5. Generate class label base association rules with global support and
    confidence using (4) & (5).
```

According to the Algorithm-1 the proposed model first uses Map node computation to generate fuzzy class label base Tid-list from each loaded data chunk of Map node, then this entire class label Tid-list will shuffle to corresponding Reducer node according to their class label to generate global tid-list. For merging two different tid-list it make use of merge_Tid-list function. In this approach the number of Reducer Nodes is directly depends upon number of class labels of dataset. The combination of class labels Tid-list yielded by all Reducer nodes is global class label based Tid-list. Using the resulted Tid-list we can easily generate associative classification rules using global support and confidence threshold. The process of generating class label based item sets from generated Tid-list with single scan is explained in [35]. As the parallel processing over distributed file system mainly involve in item set generation this paper concentrates on extending MapReduce framework over item-set generation. The rule generation process which is done over single processor is follows the method stated in standard Tid-list base classification method [35].

3.2. Dynamic Updating of Classifier with MapReduce Framework:

The dynamic data updating is possible in two ways that is data insertion and removing. In real time the data removing and inserting will be carried out using data chunks and Tid-list of each data chunk is preserved in distributed file system. In order to update existing association rule according to change we introduce a pre processing step for updating existing Tid-list before extracting updated classification rules. The same procedure adopted for dynamic updating of classification rules shown in algorithm-2. If a data chunk removed from dataset then the corresponding Tid-list will be removed from set of Tid-list and remaining Tid-lists will redistribute to among Map nodes by name node. Instead if a new chunk of data is added to distributed file system then the corresponding Tid-list will generated by a Map node and added to Tid-list set. Once the new set of Tid-list is formed then the proposed MapReduce process for associative classifier generation will be initiated. The only additional burden in this model is maintain the set of Tid-list, but at same time the size of Tid-list file ignorable in comparison with actual data size.

Algorithm-2. Algorithm for processing of class label base Tid-list to generate updated class label based classification rule set.

```
Input  :  updated data chunks
Output :  updated global classifier rules
S1. If data chunk is removed
    Then remove corresponding transaction rows
    form Tid-list.
S2. If new data chunks added
    Then Load data chunk in a Map node and
    generate corresponding Tid-list
S3. Initiate updating associative classifier
    generation from new set of Tid-list
```

3.3. Efficiency Calculation of Resulted Classifier:

In order to test efficiency of generated rules the proposed approach follows the which is adhere to dynamic increment and decrements of data saved in distributed file system. According to the proposed approach when a data chunk is assigned to Map node for processing the 20% of the data will be collected as testing set in 10 X 10 approaches for evaluating testing classifier. So when the final classifier extracted from the respective chunks of dataset stored in distributed file system using proposed model, then the extracted test set used for classifier evaluation. When the dynamically a data chunk is added and removed accordingly inclusion and deletion will be performed with evaluation test set.

Algorithm-3. Algorithm for testing the dynamic classifier accuracy on MapReduce framework.

```
Input  : Testing data chunks, Classification rule set
Output :  Global classifier efficiency
Map():
S1. For each instance in test set apply the classifier rule
    set
S2. If rule is correctly classified
    Then improve true positive count
S3. send total test instances, true positive count to
    Reducer
Reducer():
S1. Generate final test instances and final true positive
    count by adding forwarded results
S2. Calculate classifier accuracy by dividing true
    positive count by total number of instances
```

According to the proposed approach for testing the classifier at first the Map nodes will apply the rule set on the test data allotted to us and it will check whether it is correctly classified or not. The rule mapping process at first divides the rules according to class label and applies each rule for testing transaction ignoring class label to check whether any rule matches or not. When any rule exactly matches to testing transaction then it will check class label of both are same or not, if class labels exactly match it will consider properly classified else that map will be ignored. When it is properly classified then it will be counted as true positive. Finally, all true positive counts generated by Map nodes and number of test instances will be sent to single Reducer node where all true positive count added which divided by total number of test instances is going to give the classifier accuracy. The proposed algorithm for testing dynamic classifier accuracy is shown as Algorithm-3.

4. Experiment and Evaluations

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

The proposed Dynamic-MR_FAC model experimentally evaluated in three levels which includes classification accuracy over uncertain data, time accuracy with respect to dynamic changes and scalability of dynamism with respect to map nodes. In order to do so out of total data 70% of data loaded and classification rules generated as per proposed model then remaining 30% of data loaded and classification rules generated with proposed approach and finally accuracy is been calculated. The standard 10 cross 10 division method used for dividing the data into training and testing datasets. To produce associative classification rules form given training set first map nodes process initiated and local tid-lists were generated which further used for generating globally supported classification rules as explained in proposed models. In the experimentation the number of Map nodes count is decided by the framework itself based on load where the reducer nodes were initiated based on number of classes in the evaluation dataset.

In order to perform accuracy test with respect to handling data uncertainty we make use of KDD-96 UC-Census dataset [36] which is open sources classification data set with more than 1,41,544 data records. The data fuzzy portioning and the level of abstraction are have been generated by appropriately defining fuzzy sets on crisp partitions obtained by applying classical discretization algorithm [37] for continuous attributes.

The best claimed accuracies of the standard classification models are C4.5 is 81.91%, for Naive-Bayes it is 81.69%, and for NBTree it is 84.47%, [38] which are used for comparative evaluation of accuracy of proposed model. The final observation is that the proposed model shown 83.91% appreciable performance with respect to c4.5 and Naïve-Bayes but slightly lower accuracy than NBTree model.

The second phase of experiment conducted in order to evaluate the dynamic nature of the proposed model. In order to evaluate dynamic nature the training data divided two sections in 70: 30 ration in 10cross1 validation method. Taking 8-number of core systems, at first the proposed Dynamic-MR_FAC method was run on 70% of data transactions in HDFS and applied on training set. For this step the training time was 921 seconds. In the next step remaining 30% of transactions added to HDFS incremental Dynamic-MR_FAC method was run with 386 seconds (total: 921sec+386sec=) 1307sec with overall 8 accuracy. Instead of applying MR_Tid-Ac if we go rerun total associative classification algorithm by remaining 30% of data it is taking 1342 sec which results in (921+1342=) 2,263 seconds. So these experimental result show that the proposed incremental Dynamic-MR_FAC method compatible in case of accuracy and gives much better time complexity in comparing to re-running total procedure for dynamic dataset increment.

As per the intent of third phase of experiment to evaluate scalability of the proposed classification model, the time efficiency with respect to varying number of system is recorded. In order to record experiment conducted on 8, 16, 32, 64 number of maps and it was observed that the proposed model showing the consistence growth in time complexity as number of systems for computing grows. This shows the scalability factor of proposed system. The result of scalability factor of the system is shown in Figure 2.

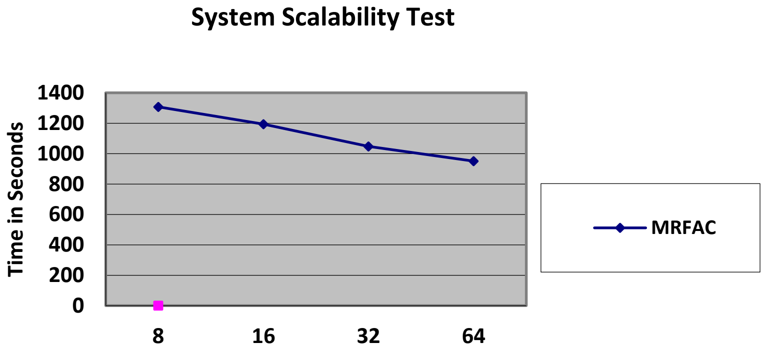

Figure 2. Scalability Efficiency with Respect to Varying Number of Maps

5. Conclusion

The efficiency of HDFS in processing big data stored distributed file system resulting in scaling of conventional data mining algorithms to MapReduce framework, but most these proposals ignoring two important factors that is uncertainty and dynamic nature of data which referred as veracity and velocity properties of Bigdata. Considering the needs of data mining proposals able to handle data uncertainty and dynamic changes in data and efficiency of associative classification technique in real-time learning assignments this paper proposes a fuzzy logic adopted associative classification approach for handling data uncertainty in Big data and also perform seem less transformation of approach to handle dynamic increments as well decrements of data. The experimental results of proposed model for extracting fuzzy class label association rules shows that it can scale successfully on MapReduce architecture to handle data uncertainty and manage dynamic increments and decrements to data stored in Hadoop distributed file system.

References

[1] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen and S. Belfkih, "Big Data technologies: A survey", In Journal of King Saud University - Computer and Information Sciences, vol. 1, **(2017)**.

[2] "Apache Hadoop Project", http://hadoop.apache.org/.

[3] J. Dean and S. Ghemawa, "MapReduce: Simplified Data Processing on LargeClusters", Google Labs, OSDI, **(2004)**, pp. 137-150.

[4] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE Communications Surveys & Tutorials, vol. 18, no. 2, **(2016)**, pp. 1153-1176.

[5] S. R. Upadhyaya, "Parallel approaches to machine learning—A comprehensive survey", In Journal of Parallel and Distributed Computing, vol. 73, no. 3, **(2013)**, pp. 284-292.

[6] N. Abdelhamid, "Associative Classification Approaches: Review and Comparison", Journal of Information & Knowledge Management, World Scientific Publishing Co, vol. 13, no. 3, **(2014)**.

[7] A. Bechini, F. Marcelloni and A. Segatori, "A mapreduce solution for associative classification on bigdata", Information Sciences, vol. 332, no. 1, **(2016)**.

[8] http://www.ibmbigdatahub.com/infographic/four-vs-big-dataUncertainty survey.

[9] C. C. Aggarwal and P. S. Yu, "A Survey of Uncertain Data Algorithms and Applications", IEEE Transactions On Knowledge and Data Engineering, vol. 21, no. 5, **(2009)**.

[10] L. A. Zadeh, "Fuzzy sets", Journal of Inf. Control, vol. 8, **(1965)**.

[11] J. Fong, H. K. Wong and S. M. Huang, "Continuous and incremental data mining association rules using frame metadata model", In Knowledge-Based Systems, vol. 16, no. 2, **(2003)**.

[12] G. Lee, U. Yun, H. Ryang and D. Kim, "Erasable itemset mining over incremental databases with weight conditions", In Engineering Applications of Artificial Intelligence, vol. 52, **(2016)**.

[13] U. Yun, H. Ryang, G. Lee and H. Fujita, "An efficient algorithm for mining high utility patterns from incremental databases with one database scan", In Knowledge-Based Systems, vol. 124, **(2017)**.

[14] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, I. Stoica and Spark, "Cluster computing with working sets", Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, Hot Cloud'10, USENIX Association, Berkeley, CA, USA, **(2010)**.

[15] J. Ekanayake, S. Pallickara and G. Fox, "Mapreduce for Data Intensive Scientific Analyses", Proc. IEEE Fourth International conferences on Sciences, **(2008)**, pp. 277-284.

[16] http://aws.amazon.com/ec2/.

[17] W. Dai and W. Ji, "A MapReduce Implementation of C4. 5 Decision Tree Algorithm," International Journal of Database Theory and Application, SERSC, vol. 4, **(2014)**.

[18] M. GH. AL Zamil, "The Application of Semantic Big Data based Classification", International Conference on Information and Communication Systems, **(2014)**.

[19] H. Yu, Jiong Yang and Jiawei Han, "Classifying Large Data Sets Using SVMs with Hierarchical Clusters," SIGKDD '03 Washington, DC, USA, **(2003)**.

[20] F. Thabtah, S. Hammoud and H. Abdel-Jaber, "Parallel Associative Classification Data Mining Frameworks Based MapReduce", Journal of Parallel Processing Letters, vol. 25, no. 02, **(2015)**.

[21] R. Bhukya and J. Gyani, "Fuzzy clustering driven fast and intuitive classifier learning with Mapreduce framework", Journal of Theoretical and Applied Information Technology, vol. 18, no. 1, **(2017)**.

[22] C.-K. Chui, B. Kao and E. Hung, "Mining Frequent Itemsets from Uncertain Data", Proceedings of 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining, **(2007)**.

[23] J. Bi and T. Zhang, "Support Vector Machines with Input Data Uncertainty", Proc. Advances in Neural Information Processing Systems, **(2004)**.

[24] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data", Proceedings of 11th ACM SIGKDD International conference on Knowledge Discovery in Data Mining, **(2005)**.

[25] C. C. Aggarwal and P. S. Yu, "Outlier Detection with Uncertain Data", Proceedings of SIAM International conference on Data Mining, **(2008)**.

[26] V. Raghavan and A. Hafez, "Dynamic data mining", Lecture Notes in Computer Science book series, vol. 1821, (**2004)**.

[27] S. Brin and R. Motwani, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", SCM Special Interset Group on Management of Data, vol. 26, no. 2, **(1997)**.

[28] A. Sultan and Al-Hegami, "Classical and Incremental Classification in Data Mining Process", IJCSNS International Journal of Computer Science and Network Security, vol. 7, no. 12, **(2007)** December.

[29] M. Ester, H.-P. Kriegel, J. Sander, M. Wimmer and X. Xu, "Incremental Clustering for Mining in a Data Warehousing Environment", Proceedings of the 24th VLDB Conference New York, USA, **(1998)**.

[30] D. Kalles and T. Morris, "Efficient Incremental Induction of Decision Trees", Machine Learning, vol. 24, **(1996)**.

[31] J. C. Schlimmer and D. Fisher, "A Case Study of Incremental Concept Induction", In Proceedings of the 5th National Conference on Artificial Intelligence, **(1986)**.

[32] M. H. Alnababteh, M. Alfyoumi, A. Aljumah and J. Ababneh, "Associative Classification Based on Incremental Mining", International Journal of Computer Theory and Engineering, vol. 6, no. 2, **(2014)**.

[33] F. Shaorong and H. Zhixue, "An incremental associative classification algorithm used for malware detection", 2nd International Conference on Future Computer and Communication, Wuhan, **(2010)**.

[34] Raghuram and J. Gyani, "Incremental associative classification on distributed databases", IEEE-Conf convergence to technologies, Pune-India, **(2014)**.

[35] F. Thabtah, P. Cowling and S. Hammoud, "MCAR: multi-class classification based on association rules", Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications, Washington, DC, USA, **(2005)**.

[36] C. Merz and P. Murphy, "UCI machine learning repository", University of California, Irvine, CA, USA, **(1996)**.

[37] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous valued attributes for classification learning", Proceedings of International Joint Conference on Artificial Intelligence, **(1993)**.

[38] R. Kohavi and S. Blvd, "Scaling up the accuracy of Navi-Bays classifiers: a Decision tree hybrid", Proceedings of Knowledge discovery and data mining, **(1996)**.

Authors

Raghuram Bhukya, is a Ph D candidate in faculty of computer science and engineering at the Jawaharlal Nehru Technological University, Hyderabad, India. He received his M.Tech in computer science and engineering from Pondicherry Central University, Pondicherry, India. Currently he holds the position of Assistant professor in department of computer science and engineering at Kakatiya Institute of Technology and Science, Warangal, India. His research interest includes Bigdata analytics, fuzzy logic, distributed computing, Machine learning algorithms and their applications.

Dr. Jayadev Gyani, currently holds the position of Assistant Professor in Majmaah University, Majmaah, Kingdom of Saudi Arabia. He received his Doctoral degree in computer science and engineering from Hyderabad Central University, Hyderabad, India. His research interest includes Design patterns, Bigdata analytics, Distributed computing, Machine learning algorithms and their applications.