Journal of Theoretical and Applied Information Technology
30th April 2017. Vol.95. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FUZZY CLUSTERING DRIVEN FAST AND INTUITIVE CLASSIFIER LEARNING WITH MAPREDUCE FRAMEWORK

RAGHURAM BHUKYA, DR. JAYADEV GYANI
[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
Many business organizations exploring their big data resources using Hadoop distributed file system due to its capability in offering fast processing, easy scalability and higher extend reliability on low cost commodity hardware. The efficiency of Hadoop distributed file system is the result of using Mapreduce frame work, which is leading research communities to scale conventional data mining techniques to fit on Mapreduce framework. So far in the research we can found proposals of data mining on Mapreduce frame work concentrating only on scalability factor, but at the same time they are ignoring two important factors which can affect the efficiency, those are reducing number of scans to database and lacking of intuitiveness in obtained data mining results. In order to overcome such short comes and considering importance of classification learning techniques in real time environment, we propose an fuzzy associative classifier learning model using Mapreduce framework which take advantage of Tid-list representation to extract the total classifier with in single scan to database and provides intuitiveness using data driven fuzzy clusters. The experimental results show the proposed model successfully enhanced the fast and intuitive efficiencies of classification techniques for big data analytics without compromising the accuracy.
Key Words: *Associative Classification, Mapreduce, Distributed File Systems, Fuzzy Clustering, Tid-List Representation, Result Intutiveness.*

1. INTRODUCTION

In the present era of data explosion, a massive amount of data accumulating from different sources like office automations, financial services, internet technologies, mobile users, social networks and sensors networks etc. According to literature studies, worldwide the volume of business data itself doubles every 1.2 years and this time range will come down in near feature. This massively accumulating data which is a result of human cognitive process is being referred as big data throwing a tough challenge to knowledge management community that is with respect to retrieving useful information within tolerable time along with maintaining consistency with commodity hardware. There is no formal definition to big data [1] but it referred by it 4V's characteristics namely, huge Volume, huge Velocity, high Variety and low veracity. As the conventional hardware and data base management technologies can't scale to handle big data, the distributed file system is suitable for persistent and scalable storage.

In real time application in order store and mange ever scaling big data, most of the data service provider adopting The Hadoop distributed file system (HDFS) [2]. HDFS is a scalable distributed file system offered by Apache Hadoop open source project, which is adopted by more than 50 fortune companies for their data source management. The HDFS can scale and store extreme size of data just using commodity hardware connected in dedicated networks. The Hadoop provides greater extent reliability and fault tolerant because it maintains 3 replicated copies of the data by default. The other reason for popularity of HDFS is its efficiency in fast processing of big data which is a result of using Mapreduce framework for computation.

The Mapreduce [3] is a framework introduced by the Google to efficiently process the huge data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

stored in scalable distributed file system. The simplicity of the Mapreduce framework is it can easily deploy on commodity hardware and its efficiency is a result of parallel processing. The parallel processing paradigm of Mapreduce follows two phases namely Map and Reduce. The Map function will executes at different machines in parallel and generates local processing results. These local processed results will be combined at other machines and processed by Reducer function to obtain global set of results. The total computation is organized around set of input (key, value) value and output (key, value) pairs which makes the easy for developers to implement parallel algorithms on framework. The flexibility of storing and processing big data using distributed file system is making Apache Hadoop Mapreduce execution environment more popular for big data analytics. Mapreduce base computation paradigm influencing many fields of computer science including data mining.

Data mining [4] which is a process of extracting useful and previously unknown knowledge from data sources is wide spread in digital life due to its potential applications in various fields including decision support systems, social networking, e-commerce, Bio-informatics and etc. Data mining offers information processing using many techniques including association rule mining, clustering, classification and text mining. Out of different information processing techniques the classification plays important role by providing automated learning and prediction offers to user. The classification techniques evaluated with respect to efficiency and interpretability. In comparison with other classification techniques like decision trees, Bayesian classification, neural network base classification the associative classification techniques are proved to be more accurate and easily interpretable [5]. At the same time the normal associative classification techniques are suffer with the problem of huge number of mined rules which decrease their time complexity and intuitiveness. This problem could be answerable by introducing fuzzy stets [6] concepts in classification rules mining which leads to a new technique known as fuzzy associative classification techniques [7].

Fuzzy logic is a fuzzy set technique which quantifies the values to upper hierarchy in semantic perspective using the membership value in between 0 to 1. In case of fuzzy generalization of balance attribute example a balance value can generalized to different label with a specific membership value as shown in figure 1. In crisp quantification monthly balance less than $2000 will considered low, in between $2000 to $5000 considered as moderate and above $5000 will be considered as high. The drawback of crisp quantification approach is it quantifies values to sharp boundaries, which is against human semantic perspective. For example balance $1995 which is considered as low balance where it can be considered as moderate balance with slight deviation with respect to human perspective. This kind human perspective quantification can be achieved by fuzzy logic, where the balance between $1000 to $3000 can quantify as low balance and as moderate with respect to a specific fuzzy membership value. Whereas balance between $3000 to $5500 can quantify as moderate balance and as high with respect to a specific fuzzy membership value between 0 to 1. For a value the fuzzy membership value 1 specifies exact relatedness to quantifier, whereas value 0 specifies not relatedness and value in between 0 to 1 specifies corresponding relatedness. The user can decide cutoff value to be considered for specific quantifier.

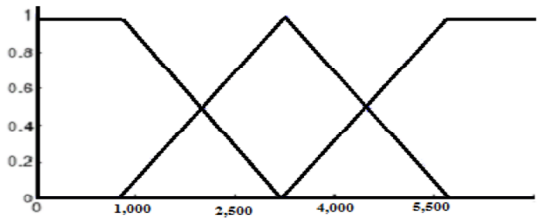

Figure 1: Fuzzy Partitions Of Balance Attribute

Considering the significance of fuzzy sets and associative classification there are proposals in the literature which explores fuzzy associative classification technique on Mapreduce frame work [8], but one important thing it ignores is data discretization into fuzzy sets are based on parameters suggested by human expert and the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

fuzzy discretization of data is also computed on single system instead of parallel Mapreduce framework. As the size of data is very high and distributed it is not reliable to depend on human experts to cluster data into fuzzy set and single computing engaging on fuzzy discretization of such massive dataset will incur huge computation cost. In order to address this problem we adopt fuzzy c-means [9] based data clustering techniques, which decides the clusters of data based up on data values and perform fuzzy discretization of data on Mapreduce framework. At the same time considering fuzzy c-means can't handle categorical as well binary data, our proposed pre processing model includes mechanism to handle intuitively these kinds of data. The other issue which could be overlooked in pre processing step is effect on time and space efficiency due to transforming whole data set into discretization representation. The proposed model reduces this effect by merging data transformation and processing operations under same module.

The other challenge to be address in exploring classical data mining techniques on Mapreduce framework is reducing number of scans to database. As the size of data base is massive in distributed file system even reduction in single scan will also show great impact on improving time efficiency. In conventional mining techniques it is already proved that Tid-list based vertical representation [10] can provide the frequent patterns even with single scan. The contributions of the work is we proposes scalable and intuitive fuzzy classification rule generation model with MapReduce framework using data driven fuzzy clusters and Tid-list based representation which can overcome difficulties in extending fuzzy associative classification method on MapReduce framework.

2. FUZZY ASSOCIATIVE CLASSIFICATION NOTATIONS & DEFINITIONS

Let data set D with total T number of transactions provided to extract classifier and when it submitted to distributed file system it is horizontally fragmented among N number of distributed systems working under common distributed file management master. Where each distributed file system consist subset of $D = \{D_1, D_2 \ldots\ldots D_N \mid D = \cup D_i$ where i=1 to N} such that transactions $T_i \subset T = \{T_1, T_2 \ldots\ldots T_n \mid T = \cup T_i$ where i=1 to n} can present only in one subset of D. Let transaction set T consist of set of m number of attributes $A = \{A1, A_2 \ldots\ldots A_m\}$ and each attribute A_c take k number of different unique values $x = \{x_1, x_2 \ldots\ldots x_k\}$. That is each value x_i taken by attribute A_c further generalized to any one of fuzzy partition $\{L_1, L_2 \ldots\ldots L_k \mid A_c = \cup L_j$ where c = 1 to m and j=1 to k}.

The hierarchical taxonomical tree for the balance attribute is shown in figure1. For example the attribute account_balance value < $ 1,000 generalized to its higher hierarchy Low_balance label and account_balance value >1,000 to 5,000 generalized to label Mid_balance and account_balance value >5,000 generalized to High_balance label.

Figure 2: *Taxonomical Tree For Balance Attribute*

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 $L_j(x_i) = \mu_c(x_i^j)$ should satisfy equation (1).

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

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.

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).

$$FS_y(L_i) = \frac{\sum_{j=1}^{n} (L_i(t_j))}{n} \quad (2)$$

The fuzzy support for a combination generated from fuzzy sets of different attributes with respect to a distributed data sub set D_y given by equation (3).

$$FS_y(L_1, L_2) = \frac{\sum_{j=1}^{n} Min(L_1(t_j), L_2(t_j))}{n} \quad (3)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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) & (3) will be updated as equation (4) & (5).

$$FS_y(L_i \rightarrow C) = \frac{\sum_{j=1\&t_j[A_C]=c}^{n}\left(L_i(t_j)\right)}{n} \quad (4)$$

$$FS_y(L_1, L_2 \rightarrow c) = \frac{\sum_{j=1\&t_j[A_C]=c}^{n} Min(L_1(t_j), L_2(t_j))}{n} \quad (5)$$

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 (6).

$$Global\_FS\ (L \rightarrow C) = \frac{\sum_{y=0}^{N} FS_y(L \rightarrow c)}{N} \quad (6)$$

The fuzzy global confidence of the fuzzy generalized associative classification rule L→C is calculated by equation (7).

$$FC\ (L \rightarrow C) = Global\_FS(L \rightarrow C) \Big/ Global\_FS(L) \quad (7)$$

3. RELATED WORK

In recent time several open source computing environments have been proposed based on Mapreduce computing paradigm in order to effectively handle big data which includes Apache Spark [11] for fast data processing, ApacheS4 [12] & CGL-Mapreduce [13] for stream data processing, Amazon EC2 [14] for cloud support and etc. Out of different computing environments Apache HADOOP is more popular because of scaling capacity on commodity hardware.

Considering impact on Mapreduce computing paradigm we can found several mining algorithms which explores conventional data mining algorithms on Mapreduce framework. The association rule finding algorithm are proposed in [15,16], k-means clustering for map reduce framework proposed in [17], the basic machine learning classification models on map reduce explored by CT-Chu [18] and further major classification work's proposed like SVM[19,20], KNN[21],boosting[22]. Even though these proposals successfully explored their respective classification techniques on scalable Mapreduce architecture, they also suffer with problems due like their conventional methods for example moderate classification accuracy, harder in interpreting the technique and assessing the results. These problems in conventional classification algorithms overcame by introducing associative classification technique.

The first associative classification technique for single class label was proposed by Liu [23] which is further updated with multi class association rule by Li [24]. The other improvements includes Yang [25] model to classify with minimum number of rules Hu &Li [26] model proposed model to handle missing values, fadi's [27] model extract associative classifier with single scan to database. The associative classifications also explored with respect to application on different fields like web mining [28], XML document classification [29], text analysis [30], image processing [31] and in [32,33,34] we can found associative classification models for distributed data mining environment. These conventional models suffer with a main problem that is lack of intuitiveness in classification result set. This lack of intuitiveness can overcome by fuzzy associative classification models proposed in [7, 35] and even fuzzy associative classification models also explored on distributed data mining environment in [36, 37, 38].

In literature we can found in Apriory algorithm [39] based associative classification model [40], fp-groath model [41] based associative classification model [42] with respect to Map-reduce architecture but the primary drawback of these models are they won't consider intuitiveness of generated results. Even there exist even fuzzy sets [8] but this proposal has drawbacks like, the model won't apply standard fuzzy clustering technique for fuzzy discretization of the data set which could affect its efficiency. The data discretization computation of the model is based on single system instead of distributed computing and even number of scans to data base is more than one which could affect time efficiency of model. Considering these drawbacks we propose fuzzy clustering driven fast and intuitive fuzzy classification rule generation model with Mapreduce framework.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. FUZZY ASSOCIATIVE CLASSIFIER EXTRACTION MODEL FOR MAPREDUCE FRAMEWORK

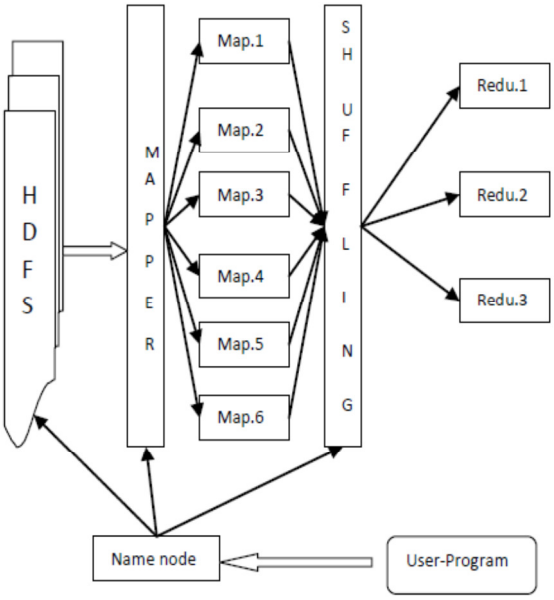

Figure 3: Mapreduce Architecture For Extracting Fuzzy Associative Classifier From Hdfs

The proposed Mapreduce model for extracting fuzzy associative classifier from HDFS is shown in the figure.1. In the Mapreduce architecture the training data will be stored in HDFS and the name node is the master node where the user program will be submitted. Once if the user program is submitted to the name node then it initiates the Map nodes based up on data load of HDFS and assign a logical data block of HDFS for each Map node for further processing. The Map nodes process the corresponding HDFS data blocks in parallel. The intermediate results obtained from Map nodes will shuffled to different Reducer nodes to process for consolidated results. The number of Reducer nodes will decide manually based up on the number of class labels in training dataset. In order to extract fuzzy classification rules from given dataset the proposed Mapreduce computation model perform following Mapreduce jobs.

MR_Job1. Data driven fuzzy quantifier parameters generation using Fuzzy c-means

MR_Job2. Tid-list base Fuzzy associative classification rules generation

MR_Job3. Classification rules pruning

MR_Job4. Classifying test instance and efficiency calculation

The comprehensive explanation of Mapreduce jobs presented in following sections.

4.1 Data driven Fuzzy quantifier parameters generation using Fuzzy c-means clustering on Mapreduce framework:

Once if the task submitted to the name node it will initiate the Mapreduce job of data pre processing, which generates parameters to transform the crisp representative data records into the fuzzy representation. While in distributed file system as the huge data is distributed among multiple number of systems, depending on human experts to decide fuzzy quantification parameters is unreliable [43].In order to obtain data driven fuzzy quantification parameters the proposed model adopts fuzzy c-means algorithm. In this Mapreduce job experts only restricted to decide number of partitions and suitable semantic names, referred as fuzzy set labels for partitions with respect to attributes where the fuzzy quantification parameters driven by data stored in HDFS will be extracted by proposed fuzzy c-means approach for Mapreduce.

The Fuzzy c-Means(FCM) is a clustering method which allows a data element to be belongs to two or more clusters with specific membership value. The advantage of FCM is that it could categorize the data according to human semantic perspective. The objective of FCM for data points {x1,x2…xn} with expected clustered centers { v1,v2,…vc } is based on minimization of function (8) and membership value μ given by (9).

$$J_m(\mu,V)=\sum_{i=0}^{c}\sum_{i=1}^{n}\mu_{i,j}^{m}\left(x_i-v_j\right)^2,\ \mathrm{m}\geq 1$$

(8)

$$\mu_{i,j}=\left[\left(\sum_{k-1}^{C}\frac{\|x_i-c_j\|}{\|x_i-c_k\|}\right)^{2/m-1}\right]^{-1}$$

(9)

The challenges in applying fuzzy c-means on map reduce architecture are the parallel processing of data and in sufficiency in main memory. Our proposed approach addressed these challenges by adopting data chunks model proposed for clustering distributed datasets proposed in oKFc-Means [44] and in order to get coordinate among parallel processed data chunks we make use of map-reduce architecture effectively. We assume each HDFS data block assign to map node as data chunk, which will be used as unit of data for generating distributed fuzzy C-means clustering parameters, in line with oKFc-Means. In order to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

explore fuzzy c-means on map reduce framework to extract fuzzy quantifier parameters from distributed file system, our proposed approach at map nodes applies fuzzy c-means algorithm on all the data chunks in parallel to generate local centroids for all numeric attributes and local centroids of each attributes collected at a corresponding reducer node where again fuzzy c-mea ns algorithm will be applied on them to decide global centroids.

The proposed Mapreduce approach for Generating numeric attributes fuzzy partitioning parameters using fuzzy c-means presented in algorithm.1. According to proposed algorithm at first step the map node using assigned data node generates fuzzy centroids matrix for all corresponding numerical attributes by applying fuzzy c-means algorithm and stores them as intermediate result. The process will be carried out by all of the map nodes in parallel until completion of the all data chunks from Hadoop data bases. Once the fuzzy centroids matrix generation of all numerical attributes is been completed by map nodes they submit the status to name node, which will redistribute the centroids matrixes to reducer nodes. The re distribution will be done by attribute wise, that is all centroids matrices of an attribute will be directed to a reducer node where by combining these again fuzzy c-means algorithm will applied and final globally optimized fuzzy partitioning values will be obtained with respect to assigned attribute. The process will be continued for all attributes at other reduce nodes individually. The final combination of the reducer results are nothing but fuzzy partitioning parameters of the all numerical attributes of the data set.

Algorithm.1: Mapreduce Job 1: Generating Fuzzy Partitioning Parameters Of Numeric Attributes Using Fuzzy C-Means

Input : Key: Training Data set, Value: Data record values and initial Vectors V_i for attributes A_i
Output : <Key, Value> pair, Where Key = attribute A_i & Value = cluster centroids vector V_i of A_i

Map_Procedure:

1. For each quantitative attribute A_i initiate the vector $V_i=\{v_{i1},v_{i2},\ldots v_{ic}\}$ where c = number of fuzzy sets satisfying conditions $\sum_{i=1}^{c}\mu_i=1$.
   a. For each transaction T_n read the attribute value A_i and apply fuzzy c-Means algorithm and update the local V_i vector
2. Using final set of V_i of all attributes form the centroids matrix M_n for the data block where n= number of map nodes.

Reduce_Procedure:

1. For each attribute A_i assigned
   a. Read attribute A_i corresponding centroids vector V_{in} from all centroids matrix M_n
   b. Calculate global fuzzy cluster centroids vector V_i of A_i by applying fuzzy c-Means on all V_{in} vectors.

Handling categorical attributes: In order to handle the categorical and binary attributes which can't be generalized, the proposed model will considers every individual value taken by such attributes as a unique quantifier for corresponding attributes. For a transaction while calculating fuzzy membership of a categorical attribute, the membership for quantifier for exact match with attribute value will considered as 1 and for all reaming quantifier the fuzzy membership value will be considered as 0. In order to realize this approach on Mapreduce frame work there should be manual input indicating the categorical and binary attributes in total training dataset. In fuzzy quantifier generation the map node in a chunk if it encounters categorical attribute it just store all the unique value of corresponding attribute. At reducer phase for a categorical attribute all the corresponding values collected from processing the entire chunk will be unified and considered at global fuzzy quantifiers of attribute.

4.2 Tid-list base Fuzzy associative classification rule generation on Mapreduce framework:

Using the fuzzy partitioning values obtained in the previous step and corresponding fuzzy set labels expert , the fuzzy associative classification process on Mapreduce architecture will be initiated in this job. The proposed Mapreduce approach for generating fuzzy associative classification rules presented in Algorithm.2. In this job as like previous job the dataset in the HDFS will be partitioned into logical blocks and each block will assign to a Map node where the fuzzy associative classification rules will be generate in parallel. In order to overcome problems of multiple scan to database and transforming entire dataset into fuzzy discretization representation our approach at first generate class label based fuzzy Tid-list at Map node for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

corresponding data block with a single scan, which is used for frequent class label based item sets. In order to do so each map node at first creates the Tid-list vector by taking all attributes fuzzy partitioned set labels on row and for each transaction creates a new column. Then for each attribute item of transaction generate fuzzy membership values with respect to corresponding fuzzy partitions and store at corresponding label in the column. After generating fuzzy Tid-list all map nodes will generate all possible levels of local fuzzy item-sets along with their support count. The detail of fuzzy itemsets generation from fuzzy Tid-list is detailed in our previous publication [38].

Once the local fuzzy item-sets generation process is been over at map nodes then each level of item-sets will be directed to a corresponding reducer node. Then for each allotted level of item set, the reducer node will club the same local item-set and find their global support count of each. After this the reducer will apply global support cutoff, where the fuzzy itemsets with less than global threshold value will be dropdown. Then class label based association rule generated and corresponding confidence value will be calculated in conventional CBA [23] approach. Once the rule generation is been over then the rules not satisfying global cutoff threshold will be drop. The process will repeated for all level of item sets at every reducer node in parallel which results into the final set of globally valid associative classification rules.

Algorithm.2: Mapreduce Job 2: Globally Supported Fuzzy Associative Classification Rules Generation

Input : Key: Training Data set, Value: Data record values and Centroids Vector V_i and fuzzy set label vector L_i of attributes A_i

Output : Fuzzy associative classification rules

Map_Procedure:

1. Create Tid-list$_n$ vector by taking all fuzzy labels of L_i on row
2. For each transaction t_i
   a. Create new column in Tid-list$_n$ do
   b. For each attribute A_i of transaction t_i do
      i. Generate membership values μ_i using centroids vector V_i using equation (1)
      ii. Store each μ_i value at corresponding label in the column
3. Generate local class label based fuzzy itemsets using Tid-list$_n$
4. Calculate support count using (5)

Reduce_Procedure:

1. For each level of allotted class label based fuzzy itemsets do
   a. Combine all same itemsets and generate global itemsets
   b. Calculate global support using (6)
   c. Drop itemsets not satisfying global support threshold
   d. Generate fuzzy associative classification rules using globally supported itemsets
   e. Generate confidence of fuzzy associative classification rules using (7)
   f. Drop rules without satisfying confidence threshold.

4.3 Rule pruning:

The Mapreduce job 3 is targeted to prune the redundant and less effective classification rules to get compact rule set. In order to prune the generated classification rules, at map node the rule set divided according to their targeted class label and each set will sort in ascending order according their rank. The rank of the rule decided based on confidence and support factor. That is a rule with high confidence will be given higher rank than rules with lower rank. If confidence of two rules are equal then their support will be compared and rule with high support value will be given higher rank and in case if two rules found same confidence and support then the rule whichever is generate first it will be given highest rank. At reduce procedure level each reducer directed to handle rule set of a single class at time where subset rules with less rank than its super set will be pruned.

4.4 Classifying test instances and efficiency calculation:

The final Mapreduce job is to classify test instance and calculate efficiency of fuzzy classifier. In order to classify test instance at first the test instances will loaded in HDFS and based up on load assigned to the map nodes where the fuzzy

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

classifying rule set will be applied on records and all the applicable rules along with their firing strength will be gathered. The class of the training record decided by the sum of firing strengths of rules with respect to each class label, that the applicable class label with highest firing strength will considered as class label of record. Before generating target class label the class label attribute of the transaction will be intentionally ignored and if once the class label of record is decided then the actual class label will be compare with obtained class label and correctness will be measured. The process will repeat for all the records in parallel at all map nodes in parallel. The consolidated accuracy and error rate will be calculated at reducer node using all the success and error scores of the all map nodes. Where the classification accuracy given by dividing correctly classified instance with total number of training instances and error rate given by dividing incorrectly classified instance by total number of training instances. The Mapreduce job 4 process is presented in algorithm.3.

Algorithm.3: Mapreduce Job 4: Fuzzy Classifier Accuracy Calculation

Input : Key: Testing data set, Value: Data record

Output : The accuracy and error rates of classifier

Map_Procedure:

1. For each test instance T_i do
   a. Find all rules that applicable to T_i and place in R
   b. If all the rules in R indicating same class then assign that class label to record
   c. Else calculate sum of firing strengths with respect to each class label
   d. Assign class label to record with highest firing strength
   e. Update success and error count by comparing actual class label with assigned class label.

Reduce_Procedure:

1. Calculate consolidated classification efficiency and error rate by combining success and failure counts of all the map nodes.

5. IMPLEMENTATION AND EVALUATION

The evaluation process considers accuracy and speed as factors to examine the performance of the proposed fuzzy associative classification algorithm implementation on map reduce architecture.

The Mapreduce framework established with maximum of 12 number of system in system where one system act as server node and other as computing nodes. Each of this system with Pentium-i5 2.67 GHz processors, 4GB RAM, 1Gbps Ethernet connection and hard disk with 500GB capacity. The MapReduce implemented with software version is Hadoop2.0.0-cdh4.4.0 and Mapreduce runtime (Classic) running on Ubuntu 14.4 operating system. .The maximum numbers of map task are dynamically decided by the master node where maximum number of reducers given is 4.

The first phase of experiment conducted to record classification accuracy of the proposed model, by taking 6 number of cores where one node considered as master node. The data set divided in training and testing data set in 10 cross 10 validation method. In order generate fuzzy associative classification rules all the nodes transferred into corresponding fuzzy representation using proposed Mapreduce base fuzzy preprocessing technique using which the globally optimized fuzzy classification rules and accuracy with respect training dataset was computed by applying subsequent algorithm on Mapreduce framework. The experiment repeatedly conducted for 3 different times and corresponding results are recorded, finally the average of recorded values used for evaluation.

In order to evaluate accuracy of our model we compare the accuracy of the our proposed fuzzy tree based global classifier model with respect to other famous classification models which are carried out on same KDD99 database[45]. The accuracy of these standard models on KDD99 database collected from literature. Considering that the accuracy of model should not be compromised we compare the accuracy our model to best reported accuracies even though they are centralized models. There huge number of models claiming accuracy on KDD99 dataset but we consider only such databases which make use of total number of available data records and attributes. The comparative study of results shown in Table.2 In comparison with Winner KDD[46]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and Runner KDD[47] models our proposed MR-FAC model shown comparably well accuracy at same time it maintained highest intuitiveness with an average of 3-rules for each class. Along with these models we implemented distributed version of standard associative classification model CBA[23] on Mapreduce referred as MR_CBA with same interaction protocol and applied on KDD99 data set. In comparison with accuracy with respect MR_CBA also the proposed model proved its comparability. All to gether the accuracy levels with respect to Normal, Dos, and Probe classes are appreciable but in case of U2R and R2L classes it shown lower accuracy level. The reason to have lower accuracy in case of U2R andR2L may be due to lower number of training instances.

Table 1: The comparative test results on KDD-99 dataset

Model Name	Normal	Probe	DoS	U2R	R2L
Winner-KDD	99.5	83.3	97.1	13.2	8.4
Runner-KDD	99.4	84.5	97.5	11.8	7.3
Dist_CBA	99.7	87.8	96.9	22.6	9.2
MR-FAC	98.2	81.3	96.7	11.3	7.1

The main objective of second phase of experiment is to evaluate time efficiency advantageous of adopting fuzzy model in comparison with non fuzzy model. In order to so we run MR_CBA and proposed model on 6, 8,10,12 number of nodes with Mapreduce frame work and which proves the time efficiency and scalability of our proposed approaches.

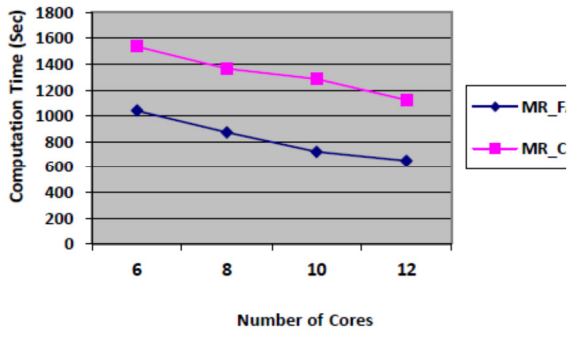

Figure 4: Computation Efficiency Of Methods With Respect To Number Of Nodes.

6. CONCLUSION

Considering the efficiency of Mapreduce framework and importance of fuzzy associative classification technique this paper proposes a fast and efficient fuzzy associative classification algorithm which can scale on Mapreduce framework. In the proposed model the accuracy of classifier acquires by using data driven fuzzy clusters and time efficiency acquired using Tid-list based representation. The experimental results of proposed fuzzy associative classification model shows that it can successfully scale on Mapreduce architecture to handle data stored in HDFS and generate intuitive classification rules without compromising the efficiency. Extracting classifiers on multidimensional and dynamic datasets with Mapreduce framework can be worth full extension the proposed model.

REFERENCES:

[1]. C.L. Philip Chen , Chun-Yang Zhang, .Data-intensive applications, challenges, techniques and technologies: A survey on Big Data Information Sciences 275 (2014) 314–347

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

[3]. MapReduce: Simplified Data Processing on LargeClusters, Jeffrey Dean and Sanjay Ghemawa, Google Labs, pp. 137–150, OSDI 2004

[4]. C.W.Tsai, C.F.Lai, M.C.Chiag, L.T. Yang " Data mining for internet of the things" IEEE communications Survery & Tutorials, 2014, Vol.16.Issue.1, 77-97.

[5]. Neda Abdelhamid, "Associative Classi¯cation Approaches: Review and Comparison", Journal of Information & Knowledge Management, Vol. 13, No. 3 (2014) World Scienti¯c Publishing Co.

[6]. Zadeh, L. A.: Fuzzy sets. Inf. Control, 8, 338–358 (1965).

[7]. F.P.Pach, A.Gyenesei, J.Abonyi, Compact fuzzy association rule-based classifier, Expert Systems with Applications, Vol.34, 2008, pp.2406–2416

[8]. Pietro Ducange, Fransceco Marcelloin, Armado Segatori, " A MapReduce-based Fuzzy Associative Classifier for Big Data ".

[9]. J.C.Bezdek, Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York, 1981

[10]. Savasere.A, Omiecinski.E, Navathe.S An efficient algorithm for mining association rules in large databases, 1995, proceedings of the 21st conference on very large data bases 95 zurich, 432-444.

[11]. M.Zaharia,M.Chowdhury,M.J.Franklin,S. Shenker,I.Stoica,Spark:Clustercomputingwithworkingsets,in:Proceedingsofthe2ndUSENIXConferenceonHotTopicsinCloudComputi

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ng,HotCloud'10,USENIXAssociation,Berkeley,CA,USA,2010.10–10

[12]. L.Neumeyer,B.Robbins,A.Nair,A.Kesari, S4:distributedstreamcomputingplatform,in:Proceedingsof2010IEEEInternationalConferenceonDataMiningWorkshops(ICDMW),,2010,pp.170–177,doi:10.1109/ICDMW.2010.172

[13]. J. Ekanayake, S. Pallickara, and G. Fox, "Mapreduce for Data Intensive Scientific Analyses," Proc. IEEE Fourth Int'l Conf eScience, pp. 277-284, 2008.

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

[15]. M.-Y.Lin,P.-Y.Lee,S.-C.Hsueh,Apriori-based frequent itemset mining algorithms on MapReduce, in: Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, ICUIMC'12, ACM, NewYork, NY, USA, 2012, pp.76:1–76:8.

[16]. H.Li,Y.Wang,D.Zhang,M.Zhang,E.Y.Chang,PFP:parallelFP-Growthforqueryrecommendation,in:Proceedingsofthe2008ACMConferenceonRecommender Systems, RecSys'08, ACM, NewYork, USA, 2008, pp.107–114.

[17]. W.Zhao,H.Ma,Q.He,ParallelKMeans clustering based on MapReduce, in : Lecture Notes in Computer Science, 5931, Springer Berlin Heidelberg, 2009, pp.674–679.

[18]. C.-T.Chu,S.K.Kim,Y.-A.Lin,Y.Yu,G.R.Bradski,A.Y.Ng,K.Olukotun,Map- Reduce for machine learning on multicore , in: B.Schölkopf, J.C.Platt, T.Hoffmanan(Eds.), Advances in Neural Information Processing Systems19,Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems Vancouver, British Columbia, Canada, December4-7,2006,MITPress,2006,pp.281–288.

[19]. Q. He,C.Du,Q.Wang,F.Zhuang,Z.Shi,A parallel incremental extreme svm classifier,Neurocomputing74(16)(2011)2532–2540.

[20]. G. Caruana,M.Li,Y.Liu, An ontology enhanced parallel SVM for scalable spam filter training, Neurocomputing108(2013)45–57.

[21]. C.Zhang,F.Li,J.Jestes,EfficientparallelkNNjoinsforlargedatainMapReduce,in:Proceedingsofthe15thInternationalConferenceonExtendingDatabaseTechnology,EDBT'12,ACM, NewYork,NY,USA,2012,pp.38–49,doi:10.1145/2247596.2247602.

[22]. I. Palit, C.Reddy, Scalable and parallel boosting with mapreduce ,IEEETrans. Knowl.Data Eng.24(10)(2012)1904–1916.

[23]. B.Liu, W..Hsu& Y.Ma." Integrating classification and association rulemining" Jouranl of Knowledge discovery and data mining, KDD-98 , 1998.

[24]. W.Li, J. Han & J.Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules ", International conference on data mining, pp. 369–376, 2001.

[25]. J.Yang, " Classification by association rules: The importance of minimal rule sets", The twentieth international conference on machine learning, 2003

[26]. H.Hu, J.Li. "Using association rules to make rule-based classifiers robust". Proceedings of the sixteenth Australasian database conference pp. 47–54, 2005.

[27]. F.Thabtah, P.Cowling and S.Hammoud, "MCAR: multi-class classification based on association rules", Proceedings of the ACS/IEEE 2005 International Conference on Computer Systems and Applications, Washington, DC, USA, pp. 33–38, 2005.

[28]. Ajlouni, M. I. A., Hadi, W., & Alwedyan, J. (2013). Detecting phishing websites using associative classification. European Journal of Business and Management, 5, 36–40.

[29]. Costa, G., Ortale, R., & Ritacco, E. (2013). X-class: Associative classification of XML documents by structure. ACM Transactions on Information Systems, 31, 3.

[30]. Yoon, Y., & Lee, G. G. (2013). Two scalable algorithms for associative text classification. Information Processing and Management, 49, 484–496.

[31]. Jabbar, M., Deekshatulu, B., & Chandra, P. (2013). Knowledge discovery using associative classification for heart disease prediction. In Intelligent informatics (pp. 29–39). Springer.

[32]. G.Thakur, C.J. Ramesh.2008 A Framework For Fast Classification Algorithms, International Journal Information Theories & Applications V.15, pp. 363-369.

[33]. D.Mokeddem, H Belbachir.2010. Distributed classification using class association rules mining algorithm. IEEE International conference on Machine and web intelligence, Algeria.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[34]. Raghuram and Jayadev Gyani, "incremental associative classification on distributed databases", IEEE-Conf.

[35]. Xin Wang, Xiaodong Liu, Witold Pedrycz, Xiaolei Zhu, Guangfei Hub, " Mining axiomatic fuzzy set association rules for classification problems", European Journal of Operational Research 218 (2012) 202–210.

[36]. B.RaghuRam and G.Aghila " Mobile agent based distributed fuzzy associative classification rules generation for OLAM", IEEE conference on IAMA, Chennai, India, 2009.

[37]. Raghuram and Jayadev Gyani, B.Hanumnthu, "Fuzzy associative classifier for distributed databases", proceedings of ICWET-2012, Mumbai, India, IJCA, 2012.

[38]. Raghuram and Jayadev Gyani, "Fuzzy generalised classifier for distributed knowledge discovery", International Journal of Business Intelligence and Data Mining, Vol.8, No.3, pp.227 – 243, 2013

[39]. R.Agrawal,T.Imieli´nski,A.Swami,Mining associationrulesbetweensetsofitemsinlargedatabases,SIGMODRec.22(2)(1993)207–216,doi:10.1145/170036.170072.

[40]. Fadi Thabtah, Suhel Hammoud, Hussein Abdel-Jaber, "Parallel Associative Classification Data Mining Frameworks Based MapReduce", Journal of Parallel Processing Letters, Volume 25, Issue 02, June 2015.

[41]. J.Han, J.Pei, Y.Yin, R.Mao, Mining frequent patterns without candidate generation : a frequent-pattern tree approach, DataMin.Knowl.Discov.8(1) (2004) 53–87, doi:10.1023/B:DAMI.0000005258.31418.83.

[42]. Alessio Bechini, Francesco Marcelloni, Armando Segatori, " A mapreduce solution for associative classification on bigdata ", Information Sciences 332 (2016) 33–55.

[43]. Ashish Mangalampalli, Vikram Pudi, Fuzzy Association Rule Mining Algorithm for Fast and Efficient Performance on Very Large Datasets, Fuzz-IEEE, 2009, South korea.

[44]. Hore, P., Hall, L., Goldgof, D., Gu, Y., Maudsley, A.: A scalable framework for segmenting magentic resonance images. J. Signal Process. Syst. 54(1-3), 183–203 (2009).

[45]. Merz, c. and Murphy, P. 'UCI machine learning repository' (1996) University of California, Irvine, CA, USA

[46]. B.Pfahringer, Winning the KDD99 classification cup: bagged boosting, ACM SIGKDD Explorations, Vol.1 (2), 2000, pp.65–66.

[47]. Levin, KDD-99 classifier learning contest LLSoft's results overview, ACM SIGKDD Explorations, Vol.1 (2), 2000, pp.67–75.