RESEARCH ARTICLE OPEN ACCESS

Survey on Enhancing the Speed of Mapreduce Classification Algorithms Using Pre-Processing Technique

Maya A. Gharat [1], Sharmila S. Gaikwad [2], Saurabh Suman [3]
PG student [1], Assistant Professor [2] & [3]
Department of Computer Science and Engineering [1] & [3]
Shree L. R. Tiwari College of Engineering
Department of Computer Science and Engineering [2]
Rajiv Gandhi Institute of Technology
Mumbai -India

ABSTRACT

The world now contains an unimaginably vast amount of electronic brain information which is getting ever infinite ever more rapidly. There are many reasons for the information explosion. The most obvious one is technology. Despite the affluence of tools to apprehand, process and distribute all this information—sensors, computers, mobile phones it already overreaches the available storage area. The quantity of data is growing very fast and we need to reflect how data should be processed and what kind of information to extract. The consequence is being felt worldwide, from industry to science, from government to the arts. Scientists and computer engineers have introduced a new word for the wonder: "Big Data"in small terms big data relates to information that can't be processed or examined using conventional processes or tools. Classification with big data has become one of the most recent vogue when talking about learning from the available information.

Keywords :- Big Data, classification ,Fuzzy Logic, Fuzzy Rule Based Classification System, Hadoop, key-value, Map Reduce

I. INTRODUCTION

Now a days, data has become a very important part of our life. The data from various sources is collected and stored somewhere in the data warehouse. For identifying the datasets that are of large size and have greater complexity we use concept of big data. Big data is a group of large amount of assembled and unassembled data from different sources comprises of Big data. Data coming from social network, machine generated data, and traditional enterprise are various sources of big data[1].Storing the data, managing the stored data and analyzing, using traditional techniques of data mining is not possible. Therefore it is necessary that for predicting the future trends, useful information has to be extracted from these data sets. The concept of Hadoop is used for processing large volumes of data [2]. There are various effective and accepted tools for pattern recognition and classification which are used now which work under the framework of Map Reduce which is a programming model and the tools are called as Fuzzy Rule Based Classification Systems [3]. They are able to obtain a good precision using these tools and they provide the end user with a model by making use of labels called as linguistic labels. Also management of various traits such as uncertainty, ambiguity or vagueness can be done in a very effectual way. When it comes to dealing with big data it becomes interesting, as this situation comes with built-in ability. The big data contains commonly greater number of instances and/or features. The exponential growth of the search space affects the initial learning quantity of FRBCSs. A rule set which cannot be interpreted may be generated this growth complicates the learning process and it can lead to unfamiliar problems or complex problems [4]. For processing large bulk of data in parallel the work can be divided into a set of individualistic tasks using a model which is designed is called Map Reduce model of programming model [5]. Classification consists of phases: the first phase is called as phase learning process in which a huge training data sets are analysed and then it create the patterns and the rules. The second phase is evaluation or testing and recording the accuracy of the representation of categorical patterns. The purpose of classification is to be able to use its model to predict the class label of objects whose class label is unknown.

In this work we will present a FRBCS through which interpretable model will be obtained and we have named the method as Chi-FRBCS-BigData. A classical FRBCS learning method called as the Chi et al's approach is used [9], which has been modified with which big data can be dealt using the programming model i.eMap Reduce. Two different versions, ChiFRBCSBigData-Max and Chi-FRBCS-BigData-Ave,will be developed and both of them will differ in their "reduce"

operation through which results will be compared and analyzed.

II. INTRODUCTION TO BIG DATA

There are data sets which are larger in size and traditional data mining techniques and various tools of software cannot be used to manage them. Information is hidden in datasets of big data in the form of large volumes and can't be examined using latest algorithms.

2.1 . BIG DATA CHARACTERISTICS [1]

Big data consists of 3Vs Volume, Variety and Velocity:

Data Volume: As the Big Data tsunami hit the data stores most organizations were already struggling with the increasing size of their databases. Nobody really knows how much data is there because the volume is growing so fast. According to computer giant IBM, 2.5 exabytes - that's 2.5 billion gigabytes (GB) - of data was generated every day in 2012. That's big by anyone's standards. "Most of the data is unstructured, coming from origin such as text, voice and video. "And as mobile phone penetration is forecast to grow from about 70% by 2017, those figures can only grow. The US government's open data project already offers more than 120,000 publicly available data sets.

Data Variety: A number of applications are gathering data from emails, documents, or blogs. Now data comes in the form of unstructured text, photos, emails, and videos, monitoring devices, PDFs, audio and also structured data. This unstructured data is difficult for analyzing.

Data Velocity: The speed of data creation, streaming and aggregation is measured in data velocity. The flow of data is continuous and in massive amount. If the velocity can be handled, the researchers and businessmen can make proper decisions which are of advantage to them.

Data Veracity: Biases, noise and abnormality are all weak points of big data. It is necessary to identify for whether the problem being analyzed is the data which is stored and mining of importance or not. There must be a team supporting all to clean the data and that dirty data cannot be accumulated in the system.

2.2. Map Reduce programming model [2]

A programming model designed for processing large volumes of data in parallel by dividing the work into a set of individual tasks. Developers are allowed to use map reduce in a serviceable programming style to create a map function that formulates a key-value pair related with the data which is given as input to produce a set of intermediate key-value pairs, and a minimize function that combines all intermediate values related with the same intermediate key and the engine consists of one Job Tracker and a number of Task Trackers. The client submits Map Reduce jobs to the Job Tracker Node. With a rack-aware file system, the Job Tracker pushes work out to available Task Tracker nodes, which contain the data or close data. The details are automatically managed by the system like partitioning the input data, scheduling and implementing tasks across a processing groups, and organizing the communications between points. The large distributed processing environment can be easily utilized by the developers not having any kind of experience. Map Reduce is a software scheme introduced by Google in the year 2004 to support dispersed computing on large data sets on collection of computers **[3]**. The scheme inspired by the map and the reduce tasks are commonly used in practical programming **[4].** When framework interacts with user's mappers and reducers, the scheme uses typed information every time. Details read from files into Mappers, sent by mappers to reducers, and send out by reducers into resultant files are stored in HDFS.

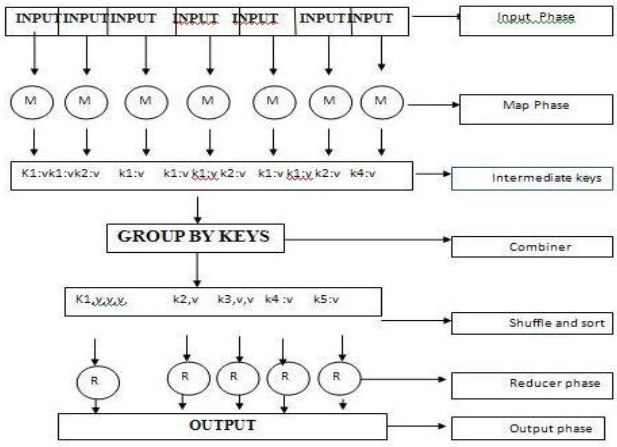

Fig 1.Map Reduce programming model

As shown in above figure a typical MapReduce program with its map and reduce steps is shown.

- **Input Phase** − Here there is a Record Reader that converts each record in an input file and forwards the parsed data to the mapper in the form of key-value pairs.

- **Map** − A series of key-value pairs are taken by the map and processed. Also key-value pairs are processed for generating zero or more key-value pairs.

- **Intermediate Keys** − The key-value pairs generated by the mapper are known as intermediate keys.

- **Combiner** −A local reducer which is called combiner, group's identical data from the map phase into recognizable sets. It takes the intermediate keys from the mapper as input and applies a user-defined code to collect the values in a small room of one mapper.

- **Shuffle and Sort** – Next, the step of shuffling and sorting is started. The sorted key-value pairs are downloaded onto the native machine where sorting initiated. The independent key-value pairs are sorted by key into an enormous data list. The data list groups the similar keys into one so that their values can be repeated easily in the Reducer function.

- **Reducer** − The grouped key-value paired data is taken by the reducer as input a Reducer task is run on each one of them. Here, the data can be assembled, percolated, and grouped in a number of ways, and it requires a different levels of processing. Once the implementation is over, it gives zero or more key-value pairs to the last step.

- **Output Phase** − There is a output designer that converts the last key-value pairs from the Reducer task and writes them onto a file with the help of a record writer.

In order to overcome these problems, several approaches have been proposed to deal with big data as substitutes for Map Reduce and Hadoop.

III. CHI-FRBCS –BIG DATA: A LINGUISTIC FUZZY RULE BASED CLASSIFICATION SYSTEM FOR BIG DATA [5]

In this section, we will introduce two versions of a linguistic FRBCS that manage big data. To do so, first, we present some definitions related to FRBCSs and the fuzzy learning algorithm that has been adapted in this work, Chi- FRBCS. Then, we will describe how this method is adapted for big data using a MapReduce scheme that is modified to produce two variants that will provide different classification results.

A. *Fuzzy Rule Based Classification Systems*

A FRBCS is composed by two elements: the Inference System and the Knowledge Base (KB). The formation of KB is helped by Data Base(DB), containing the functions of membership of the fuzzy sub division associated to the input attributes, and the Rule Base (RB), which contains the fuzzy rules that outlines the complication. Traditionally, specialist information for building the KB is not there and therefore, a machine learning strategy needed to form the KB from the convenient examples [6].

A classification problem is usually defined by m training samples $x_p = (xp_1, xp_2 \ldots xp_n)$, p = 1,2.. m from M category where x_{pi} is the result of attribute i (i=1, 2... n)of the p-th training sample. In this work, we use fuzzy rules of the following form to build our FRBCS:

Rule R_j : If x1 is Aj^1 and...and x_n is A j^nthen Class = $C_j$with $RW_j$where R_j is the label of the jth rule, x = (x_1,x2,x3... x_n) is an-dimensional pattern vector, $A^ij$is an antecedent fuzzy set,$C_j$is a class label, and $RW_j$is the rule weight [7]. We use triangular membership functions as linguistic labels. There are many alternatives that have been proposed to compute the rule weight [18]. Among them, a good choice is to use the heuristic method known as the Penalized Certainty Factor (PCF) [8].

$$RW_j = PCF_j = \frac{\sum xp \varepsilon Cj \mu Aj(xp) - \sum xp \varepsilon Cj \mu Aj(xp)}{\sum_{p=1} m \mu Aj(xp)}$$

where $\mu A_j(xp)$ the membership grade of the xp p-th example of the training set with the antecedents of the rule and $C_j$is the successive class of rule j . When predicting a class using the built KB for a given example we use the fuzzy reasoning method of the wining rule [9].

A relationship between the input attributes and the classes space is found for generating fuzzy KB using this generation method using following steps.

1) Linguistic fuzzy partitions are build: This step builds the fuzzy DB from the domain associated to each attribute A_i using equally distributed triangular membership functions.

2) A new fuzzy rule associated to each example $x_p = (x_{p1}; : : : ; xp_n, C_p)$ is generated.

a) Matching degree x_p of example is calculated with respect to the fuzzy labels of each attribute using a conjunction operator.

b) The maximum membership degree obtaining a fuzzy region is selected which is corresponding to the example.

c) Building of a new fuzzy rule whose antecedent is calculated corresponding to the previous fuzzy region and whose successive is the class label of the example Cp.

d) Rule weight is computed finally.

Now, several rules with the same antecedent can be built. If they have the same class in the successive, then, same rules are deleted if having the same class in the successive. However, rule with highest weight is maintained in the RB if the class in the successive is not same if the class in the successive is different, only the rule with the highest weight is maintained in the RB.

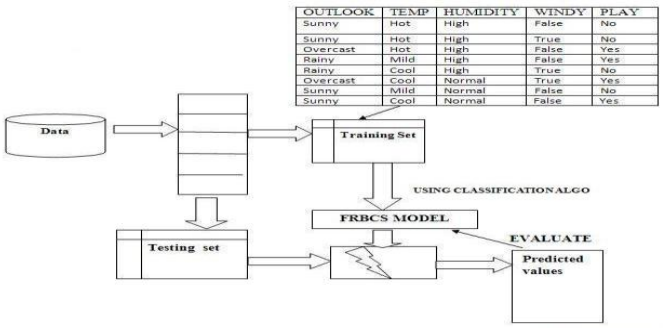

Fig 2 Classification Model

C. The Chi-FRBCS-Big Data algorithm: A MapReduce Design

In this section, we will present the Chi-FRBCS Big Data algorithm that we will develop to deal with big data classifications problems. To do so, this method uses two different MapReduce processes. As shown in above figure this is the proposed model for classification.

- One MapReduce process is committed to the fabrication of the model from a big data training set.
- The other MapReduce process is used to estimate the class of the examples belonging to big data sample.

Building the knowledge base for the Chi-FRBCS Big Data using a Map Reduce design. This procedure is divided into the following phases:

1) Initial: In this first phase, the method computes the domain associated to each attribute $A_i$using the whole training set. With that information, the fuzzy DB is created using equally distributed triangular membership functions as in Chi-FRBCS.
The original training dataset is segmented into independent data blocks automatically which are transferred to the different processing units together created DB.

2) Map: Using the available data each processing unit works separately for building associated RB which follows the original Chi-FRBCs method. Then related fuzzy rule is created. Using the example values , membership degree of the fuzzy labels are calculated then, the fuzzy area that obtains the highest value is selected to become the precursor of the rule; next, the class of the example is assigned to the rule as successive, and finally, the set of examples which belong to the current map process ,the rule weight is computed. After the rules have been created and before finishing the map step, Now, each map process searches for rules with the same successive. Only one rule is preserved if the rules share the same successive, rules having differentsuccessives, the rule with the highest weight is kept in the mappers RB .

3) Reduce: In this third phase, a processing unit receives the results obtained by each map process (RBi) and combines them to form the final RB .The combination of the rules is straight-forward: the rules created by each mapper RB_1,RB_2; $RB_n$are all integrated in one RB,RB_R. However, contradictory rules (rules with the same antecedent, with or without the same successive and with different rule weight) may be created. Therefore, specific procedures to deal with these contradictory rules are needed. Precisely, these procedures define the two variants of the Chi-FRBCS-Big Data algorithm.

(a)Chi-FRBCS-Big Data-Max: Rules with the same antecedent are searched in this approach. The final RB, RB_R will contain the rule having greatest weight. Among these rules, only the rule with the highest weight is maintained in the final RB, RB_R. Only the most powerful rules will be maintained therefore not necessary to check the consequence if same or not. The rules can have same successive and

antecedent and will be computed in different mapper processes with different training sets.

(b) Chi-FRBCS-Big Data-Ave: Here also rules having same antecedent are searched. Rules having the same successive are then calculated .The rules with same antecedent and successive can have the same weight have different weights as they are built over different training sets. Finally, the rule with the greatest average weight is kept in the final RB, RB_R.

4) Final: Finally in this phase, the results that were calculated in the previous phases serve as the output of results evaluated. Absolutely, the generated fuzzy KB is composed by the fuzzy DB built in the "Initial" phase and the fuzzy RB, RB_R, is finally obtained in the "Reduce" phase. The model that will be used to predict the class for new example will be this KB now.

IV. CLASSIFYING BIG DATA SAMPLE SETS

Chi-FRBCS-Big Data uses another MapReduce process to estimate the class of the examples that belong to big data classification sets using the KB built in the previous step. This approach follows a similar scheme to the previous step where the initial dataset is distributed along several processing units that provide a result that will be part of the final result.

This mechanism does not include a reduce step as it is not necessary to perform a computation to combine the results obtained in the map phase. Specifically, this class estimation process is depicted in Figure 5 and follows the phases:

Initial: In this first phase, the method does not need to perform a specific operation. The system automatically segments the original big data dataset that needs to be classified into independent data blocks which are automatically transferred to the different processing units together with the previously created KB.

Map: In this second phase, each map task estimates the class for the examples that are included in its data partition. To do so, each processing unit goes through all the examples in its data chunk and predicts its output class according to the given KB and using the fuzzy reasoning method of the wining
rule. Please note that Chi-FRBCS-Big Data- Max and Chi-FRBCS-Big Data-Ave will produce different classification estimations because the input RBs are also different, however, The class estimation process followed is exactly the same for both approaches.

Final: In this last phase, the results computed in the previous phase are provided as the output of the computation process. Precisely, the estimated classes for the different examples of the big data classification set are aggregated just concatenating the results provided by each map task.

V. CONCLUSION

In this work, we presented the importance of Big Data and how fuzzy rule-based classification algorithm with linguistic variables will be implemented. Using this algorithm we will try to obtain an interpretable model that will be able to handle big collections of data. We will use Map Reduce Programming model. In this way, our model distributes the computation using the *map* function and then, combines the outputs through the *reduce* function. The Chi-FRBCS-Big Data algorithm is developed in two different versions: Chi-FRBCSBigData- Max and Chi-FRBCS-BigData-Ave and compared for their performance.

REFERENCES

[1] Dr. Arvind Sathi,*Big Data Analytics*,Disruptive Technologies for Changing the Game

[2] Lijuan Zhou, Hui Wang, Wenbo Wang, "Research on Parallel Classification Algorithms for Large-scale Data", *Journal of Convergence Information Technology(JCIT)* ,Volume 7, Number 21, Nov 2012.

[3] Jing Jin, Zhen Qin, Xin Li, Shanzhi Chen, "A Sustainable Scalable Framework for Map Reduce", *International Journal of Advancements in Computing Technology*, vol.4, No.9, 2012

[4] Jeffrey Dean, Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", In Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI), pp.137-150, 2004

[5] Victoria L´opez, Sara del R´ıo, Jos´e Manuel Ben´ıtez and Francisco Herrera," On the use of MapReduce to build Linguistic Fuzzy Rule

Based Classification Systems for Big Data",*International Journal of Computational Intelligence Systems, Vol. 8, No. 3,* July 6-11, 2014.

[6] Pedro Villar, BartoszKrawczyk, Ana M. Sanchez, Rosana Montes and Francisco Herrera ,"Designing a compact Genetic Fuzzy Rule-Based System for One-Class Classification",*IEEE International Conference on Fuzzy Systems* ,2014.

[7] H. Ishibuchi and T. Nakashima, "Effect of Rule Weights in Fuzzy Rule-Based Classification Systems,"*IEEE Transactions on Fuzzy Systems*,vol. 9, no. 4, pp. 506–515, 2001.

[8] H. Ishibuchi and T. Yamamoto, "Rule Weight Specification in Fuzzy Rule-Based Classification Systems*," IEEE Transactions on FuzzySystems*, vol. 13, no. 4, pp. 428–435, 2005.

[9] O. Cord´on, M.J. del Jesus and F. Herrera, "A proposal on Reasoning Methods in Fuzzy Rule-Based Classification Systems," *InternationalJournal of Approximate Reasoning*, vol. 20, no. 1, pp. 21–45, 1999.

[10] Z. Chi, H. Yan and T. Pham, "Fuzzy algorithms with applications to image processing and pattern recognition", *World Scientific*, 1996

[11] L.X. Wang and J.M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 6, pp. 1414–1427, 1992