ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 10, October 2016

Enhancing Speed of Map Reduce Classification Algorithms Using Pre-Processing Technique

Maya A.Gharat, Sharmila S.Gaikwad, Saurabh Suman

***Abstract*— With exponentially increasing electronic data day by day, Big Data is gaining attention for solving faster access and summarization problems. However, this huge amount of data with heterogeneous formats compelled us to renovate our traditional use of learning algorithms and ponder about new techniques which are challenging and complex. To solve problem of big data, we propose a linguistic fuzzy rule based classification system, which mainly consist of two methods viz. FuzzyReducerMax and FuzzyReducerAve. As name fuzzy suggest vague and uncertain in the similar way it is dealing with uncertainty that is essential to the diversity and authenticity of big data and because of the procedure of linguistic fuzzy rules it is capable to render a recognizable and operational classification model. This process is established on the MapReduce framework, which are very popular and frequently used to handle big data by Hadoop framework. The performance measure is done on these methods by using a Data set of networking attack logs. The result shows its capability to provide accuracy on classification with both the approaches and runtime analysis which shows its speed improvement.**

***Index Terms*— Big Data,Classification,Hadoop,Map Reduce NLP, Parts of Speech Tagging**

I. INTRODUCTION

Data has become a very important part in the field of function, industry, organization, economy, business, and individual. The data from various sources is stored somewhere in the data warehouse. Big Data is a new word used to classify the datasets that are of very large in size and have bigger complexity. Big data is a collection of huge volumes of structured and unstructured data from heterogeneous sources. The heterogeneous sources of big data are such as data generating from social network, data coming by, traditional enterprise and machine [1]. This data cannot be stored, managed and analyzed using traditional techniques of data mining. Useful information has to be extracted from these data sets for predicting the future trends. To process large volumes of data from different sources quickly, Hadoop is used [2].

Fuzzy Rule Based Classification Systems (FRBCSs) systems which follows the MapReduce principle for imbalanced big data are effective and accepted tools for pattern recognition and classification [3]. Furthermore, the FRBCSs can manage ambiguity, vagueness, or uncertainty in a very effective way.This trait is important when big data problems are handled, as uncertainty is inherent to this situation. However, when handling big data, the information at end usually have a large number of instances or/and features. In this case the inductive learning capability of FRBCSs is affected by the exponential growth of the search domain. This growth increases the complexity of the learning process and it may have complexity problems or scalability problems in future while generating a rule set that is not interpretable [4]. MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks [5].

***Manuscript received Oct 2016*.**

***First Author name**, Maya A.Gharat, Shree L.R Tiwari College of Engineering, Mumbai, India, Mob No.9970407910*

***Second Author name**, Sharmila S.Gaikwad, Rajiv Gandhi Institute Of Technology, Mumbai, India.*

***Third Author name**, Saurabh Suman, Shree L.R Tiwari College of Engineering,Mumbai , India.*

II. CLASSIFICATION

A model or classifier is constructed to predict the categorical labels. The data classification process includes two steps –

- Building the Classifier or Model
- Using Classifier for Classification

1. *Building the Classifier or Model*

This step is the learning step or the learning phase. In this step the classification algorithms build the classifier. The classifier is built from the training set made up of database tuples and their associated class labels. Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

2. *Using the Classifier for Classification*

The classifier is trained using the training set. Based on the test data, the runtime and accuracy performance are measured. The runtimes of both the techniques are compared using Natural Language Processing and without using Natural Language processing. A classifier is derived after the learning algorithm works on training set of data. Using the test set we test the classifier. If the data set is classified in most of test sets we assume the future data will also be tested correctly.

C. Support Vector Machine (SVM)

A [6] support vector machine is a classification type of technique used to examine data and identify patters in classification and regression analysis. Support vector machine (SVM) is utilized when your data is classified as two category. An SVM identify and isolates similar data by finding the best hyper plane that isolate all data points of one

category from those of the other category. Accuracy improves when margins are larger between category. A margin should not have points in its interior part. The support vectors are the data coordinates that are on the boundary of the margin. Mathematical functions are used in SVM design which is frequently used to model real world problems. Its performance magnify with number of attributes.

III. Linguistic fuzzy rule base classification[7]

A.Fuzzy Rule Based Classification Systems

A FRBCS is composed by two elements: the Inference System and the Knowledge Base (KB). In a linguistic FR-BCS, the KB is formed from the Data Base (DB), which contains the membership functions of the fuzzy partitions associated to the input attributes, and the Rule Base (RB), which comprises the fuzzy rules that describe the problem. Traditionally, expert information to build the KB is not available and therefore, a machine learning procedure is needed to construct the KB from the available examples. A classification problem is usually defined by m training samples x_p =(x_{p_1} , . . . , x_{pn} , C_p), p = 1, 2, . . . , m from M classes where xpi is the value 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 Rj: If x_1 is A_j^1 and ……and x_n is A_j^n then class C_j with RW_j where *Rj* is the label of the *j*-th rule, X=x_1……..x_n is a *n* dimensional pattern vector, A_j^i antecedent fuzzy set, *Cj* is a class label, and *RWj* is the rule weight We use triangular membership functions as linguistic labels.

B.Map reduce principle used in BigData

The MapReduce[8]model is based on necessary data struc-ture that is generally known as a key-value pair. All the data processed, the intermediate results and the final output are expressed in this key-value form. In this manner, the map and reduce technique that appears in a MapReduce procedure are:

- **Mapfunction:** In the map function the master link makes an automatic separation of the data into self-directed data blocks which are then distributed and dedicated to the sub task performer nodes. Each sub nodes executes in-dependently its data and generates a result that is trasferred back to the master link node. In terms of the key-value pairs, it is said that the map function receives a key-value pair as input and generates a list of intermediate key-value pairs. These intermediate key- value pairs are then automatically shuffled and ordered according to the intermediate key to speed up the reduce step.
- **Reduce function:** In the reduce function, the master link gathered the outcomes generated in the previous phase and then, uses them in some way to get the ultimate result of the algorithm. Again, in form of the key-value pairs, the reduce function got the interme- diate key-value pairs calcuated previously summed up by the key values and generate an output value that becomes the output of the method.

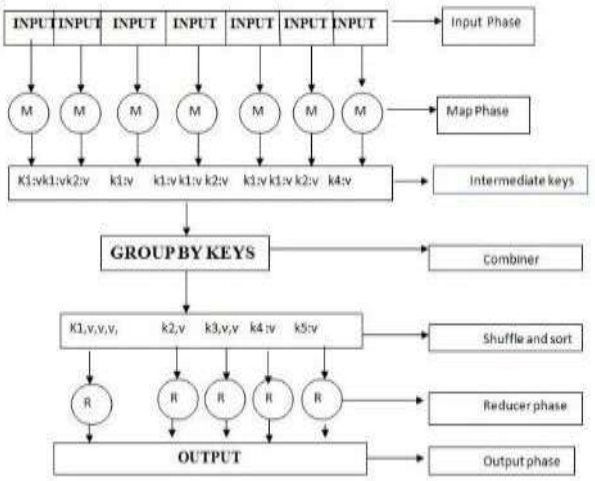

Fig. 2.Mapreduce programming model

Fig.2 depicts a standard MapReduce technique with its map and reduce steps. *k* and *v* indicates to the original key value pair respectively ; k^i andv^j are the intermediate mid meta data key-value pair that is created after the utilization of the map function; and v^{ii} is the ultimate treated as a form of result value of the algorithm.

Hadoop is the most suitable and very known among developers for implementation of the MapReduce programming model [9]. It is an open-source project coded in Java and maintained by the Apache software foun-dation that attempts to provide solution for management and processing of huge datasets in a distributed way. It provides similar services analogous to Google's Map Reduce technique.

Machine learning technique have also began to be associated using the MapReduce principle to handle big data. The Mahout project , also maintained by the Apache soft -ware foundation, is a machine learning library that has scalable machine learning capable applications over Hadoop or other scalable systems.

IV. Literature survey

In [10], IVTURS, which is a linguistic fuzzy rule-based classification technique based on a new completely interval-valued fuzzy reasoning method. This inference process utilize interval-valued limited equivalence functions to magnify the appropriateness of the rules in which the equivalence of the ideal membership degrees of the patterns and the common membership degree is greater, which is acceptable behavior. Interval-valued fuzzy sets have tested to be an appropriate tool to design the system intricacies and the ignorance in the definition of the fuzzy terms. Furthermore, the parametrized construction of these functions permits us to figure out the most appropriate set of IV-REFs to work out with each peculiar problem. An IVFS renders an interval, instead of a single digits, as the membership degree of each component to this set. More the equivalence between the example and the antecedent, greater will be the importance of the rule in the decision execution.

In [11], FARC-HD a fuzzy association rule-based classification method for more dimensional problems supported on three steps to get an accurate and precise fuzzy rule based classifier with a minimal computational cost was proposed. This technique restricted the order of the associations in the association rule extraction and believes the use of subgroup searching based on an improved Weighted Relative Accuracy measure to preselect the most appropriate rules before a processor rule selection with genetic post-processing.

In [12], Rule-Based Classification Algorithm for Uncertain Data, a approach for handling uncertain data was proposed by. A new rule-based algorithm for classifying and predicting both certain and uncertain data. A rule-based classifier is a technique for classifying records using a collection of "if ... then ..." rules. The uncertain data model was integrated with rule based mining algorithm. A new measure for generating rules was also introduced which was called as probabilistic information gain. For handling the data uncertainty the rule pruning measure has also been extended optimizing rules, and class prediction for uncertain data. This algorithm follows the new paradigm of directly mining uncertain datasets.

In [13],Fuzzy Unordered Rule Induction Algorithm, builds upon the RIPPER interval rule induction algorithm was proposed. The model built by FURIA uses fuzzy rules of the form given in using fuzzy sets with trapezoidal membership functions. Specifically, FURIA builds the fuzzy RB by means of two steps:

- Learn a rule set for every single class using a one-versus-all decomposition. To this aim, a modified version of RIPPER is applied, which involves a building and an optimization phase.
- Obtain the fuzzy rules by means of fuzzifying the final rules from the modified RIPPER algorithm in a greedy way. At classification time, the class predicted by FURIA is the one with maximal support. In case the query is not covered by any rule, a rule stretching method is proposed based on modifying the rules in a local way to make them applicable to the query.

Disadvantages:
1. It may come along with an unwanted bias since classes are no longer treated in a symmetric way.
2. Sorting rules by priority compromises comprehensibility.

In 2001 [14] Ishibuchi's method with weight factor was proposed. It implements the second type of FRBCS which has certainty grades (weights) in the consequent parts of the rules. The antecedent parts are then determined by a grid-type fuzzy partition from the training data. The consequent class is defined as the dominant class in the fuzzy subspace corresponding to the antecedent part of each fuzzy IF-THEN rule. The class of a new instance is determined by the consequent class of the rule with the maximum product of its compatibility and certainty grades. The compatibility grade is determined by aggregating degrees of the membership function of antecedent parts while the certainty grade is calculated from the ratio among the consequent class.

V. Proposed methodology

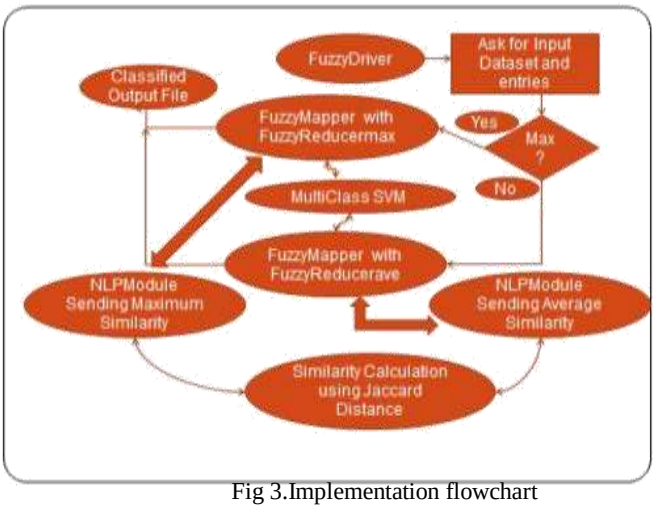

Fig 3.Implementation flowchart

It is classified into 5 Modules

1. User Interface for following Inputs
   - Data Set file in Specific comma Separated Value format.
   - Number of Entries required.
   - Number of Training Set given.
2. Pre-processing NLP Module
3. Apply Semantic Classification for Average calculation.
4. Apply Semantic Classification for Maximum Calculation.
5. Comparison of Both classification results with Accuracy

User Interface: - It is developed to take input file from user for classification and apply two methods of classification for observation of better one between them.

Preprocessing NLP Module:-It is use to handle similarity among words.

Semantic classification: - This is the last step where based training set classification is done by using Support vector machine.

PoS Tagging[15][16]: POS tagging is a very important preprocessing task for language processing activities.PoS tagging is the process of identifying and marking the words in the sentence with its Part of Speech(PoS)category. Part of speech (POS) tagging is the process of labeling a part of speech or other lexical class marker to each and every word in a sentence. Unique tag to each keyword reduces the number of parses. It is similar to the process of tokenization for

computer languages. POS tagging is considered as an important process in speech recognition, natural language parsing, information retrieval and machine translation. POS tagging is a well-understood problem in NLP, to which machine learning approaches are routinely applied.

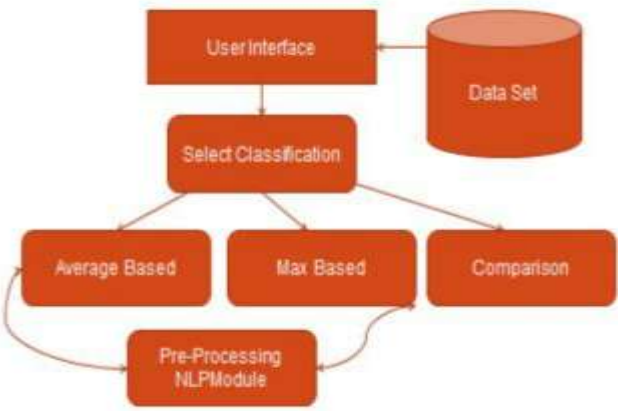

Fig. 4. Call Hierarchy of Modules

Sequences of operations are depicted in call hierarchy diagram. Fuzzy Driver GUI asks for input files of data set and checks its specific format i.e. Comma separated values. Then select the size of training set used from that data set so that pattern can be designed for further classification. It is recommended to use one fourth size of data set for training is good because less the training set it will take less time for learning patterns and if we increase the size of training set, though it will take longer time in learning patterns, there is no significant change in classification of patterns in output.For Classification, Support vector machine is used as it can support classification of more than one class. For handling big data sets mapreduce principle is used in FuzzyMapper with reducer versions of average and maximum. Differences in calculations of both the versions are as follows:

For Classification, Support vector machine is used as it can support classification of more than one class. For handling big data sets mapreduce principle is used in fuzzymapper with reducer versions of average and maximum. Differences in calculations of both the versions are as follows:

- Average Calculation :- This process is implemented in three steps :-

1. Separate Strings into tokens
2. Apply Multiclass Support Vector machine approach.
3. Calculate similarity based using Jaccard Distance and return average similarity number.

- Max Calculation :- This process is implemented in three steps

1. Separate Strings into tokens
2. Apply Multiclass Support Vector machine approach.
3.Calculate similarity based using Jaccard Distance and return maximum similarity number.The Jaccard [18]coefficient, which is sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms[17].For detections of similarity of strings here RiWordnet library is used which makes groups of similar words based on key patterns provided as input.

For measurement of performance and tolerance, sample of testing data set is gradually increased and checked the accuracy factor by executing a formula:

Accuracy Calculations:-

$$\frac{100*\text{Number of Correct Classification}}{\text{Number of Correct Classify + Number of Incorrect Classify}}$$

After execution of both the methods it has been observed that there are marginal changes when ReducerAve method and ReducerMax method is used. And its analysis is depicted and figure and the above table with two parameters considered here viz. accuracy and runtime execution.

The comparison is done using the standard dataset collected from UCI repository which shows significance difference in the runtimes and accuracy when both the algorithms are compared.The speed of FuzzyReducer Ave is more as compared to Fuzzy ReducerMax.Conversely ,the accuracy is more when the FuzzyReducer Ave algorithm is run for the dataset as compared to Fuzzy Reducer Max.

The techniques are also compared with NLP and without NLP.With NLP the runtimes are better as compared to running the algorithm without NLP.As NLP is a pre-processing module the runtimes improve for both the algorithms.The Table I shows the comparison of both the algorithms with respect to NLP and without NLP.

No of entries :2000
No of entries for training set:200
No of entries for testing:1800

Table I

Algorithms	Using NLP	Without Using NLP	Accuracy
FuzzyReducerAve	47.896 sec	262.183 sec	99.779
FuzzyReducerMax	45.896 sec	254.608 sec	99.27

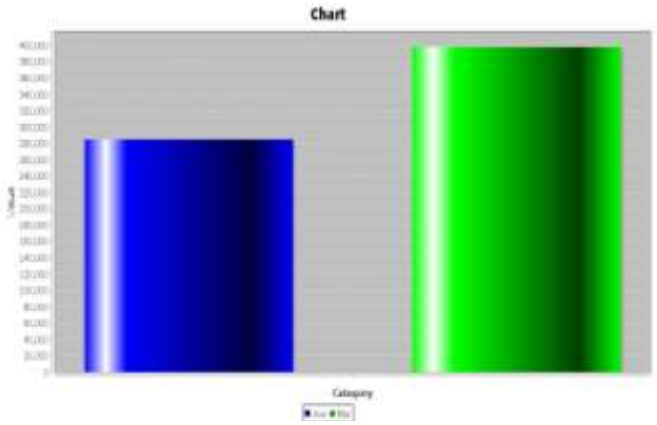

Fig 5.Runtime analysis without NLP

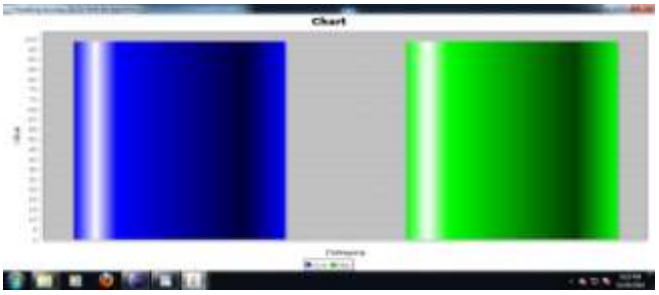

Fig 6.Runtime analysis with NLP

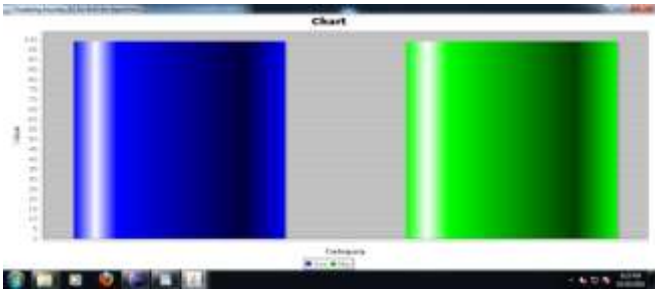

Fig 7.Analysis of Accuracy

VI CONCLUSIONS

The impact of mapreduce concept over a classification of big set of data with short span of time is seen. It speed up the process without compromising with the accuracy of the desired outcome. In future this principle can be used to classify issues related to disease prediction like coronary diseases , diabetes, in phonetic analysis like detection and training of voice for executing your commands by handicap person.

REFERENCES

[1] P. Zikopoulos, C. Eaton, D. DeRoos, T. Deutsch and George Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill,2011.

[2] S. Madden,2012"*From Databases to Big Data*," IEEE Internet Computing,vol. 16, no. 3, pp. 4–6.

[3] Y. Jin, "Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement," IEEE Trans. Fuzzy Syst. ,vol. 8, no. 2, April 2000.

[4] H. Ishibuchi, T. Nakashima and M. Nii, *Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining.* Springer–Verlag,2004.

[5] Y. Jin, "Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 2, pp. 212–221, 2000.

[6] Durgesh k. Srivastava, Lekha Bhambhu "*Journal of Theoretical and Applied Information Technology*", Journal of Theoretical and Applied Information Technology.

[7] Pedro Villar , Alberto Fern´andez , "Francisco Herrera Studying the Behavior of a Multiobjective Genetic Algorithm to design Fuzzy Rule-Based Classification Systems for Imbalanced Data-Sets," in IEEE Int. Conf. on Fuzzy Systems., 2011,pp.1240.

[8] Victoria López ∗, Sara del Río, José Manuel Benítez, Francisco Herrera, "*Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data*" *Dept. of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology),University of Granada, Granada, Spain*

[9] J. Alcala-Fdez, R. Alcala, and F. Herrera, "A fuzzy association rule based classification model for high-dimensional problems with genetic rule selection and lateral tuning*," IEEE Transactions on Fuzzy Systems*,vol. 19, no. 5, pp. 857–872, 2011.

[10] Jose Antonio Sanz, Alberto Fern ´ andez, Humberto Bustince "IVTURS: A Linguistic Fuzzy Rule-Based Classification System Based On a New Interval-Valued Fuzzy Reasoning Method With Tuning and Rule Selection,"*IEEE Transactions On Fuzzy Systems,* Vol. 21, No. 3, June 2013

[11] Jesus Alcal ´ a-Fdez, Rafael Alcal ´ a, and Francisco Herrera ,"A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems with Genetic Rule Selection and Lateral Tuning," *IEEE Transactions On Fuzzy Systems.*

[12] Biao Qin, YuniXia,SunilPrabhakar,YichengTu, "A Rule-Based Classification Algorithm for Uncertain Data," IEEE International Conference on Data Engineering.,2009.

[13] Jens Hühn, EykeHüllermeier, "FURIA: An Algorithm For Unordered Fuzzy Rule Induction," Data Mining and Knowledge Dicovery,19(3),293-319(2009).

[14] H.Ishibuchi andT. Nakashima, "Effect Of Rule Weights In Fuzzy Rule-Based Classification Systems,"*IEEE Transactions on Fuzzy Systems* , Volume: 9, Issue: 4,PP. 506 – 515, Aug 2001 .

[15] Daniel Jurafsky & James H. Martin.,"*Part-of-Speech Tagging", Speech and Language Processing, Draft* of February 19, 2015.

[16] Tao Jianchao ,"An English Part Of Speech Tagging Method Based On Maximum Entropy," *Intelligent Transportation, Big Data and Smart City (ICITBS)*, International Conference on 21 January 2016

[17] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu, "Using of Jaccard Coefficient for Keywords Similarity,*"Proceedings of the International MultiConference of Engineers and Computer Scientists* , Vol I, Hong Kong, IMECS 2013, March 13 - 15, 2013.

[18] Sapna Chauhan, Pridhi Arora ,Pawan Bhadana, *" Algorithm for Semantic Based Similarity Measure" International Journal of EngineerinG Science Invention*, Volume 2 Issue 6, PP.75-78, June. 2013.