Result Analysis of Ave and Max Classification Algorithms

Maya A.Gharat
Dept.of Computer Science and Engineeering
Shri. L.R Tiwari College of Engineering
Mira Road,Mumbai,India
mayagharat1229@gmail.com

Sharamila S.Gaikwad
Dept.of Computer Science and Engineering
Rajiv Gandhi Institute of Technology
Andheri,Mumbai,India
sharmila_gaikwad@yahoo.com

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 Hadoopframework. 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.

*****

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 volumesof data from different sources quickly, Hadoop is used [2]. 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].

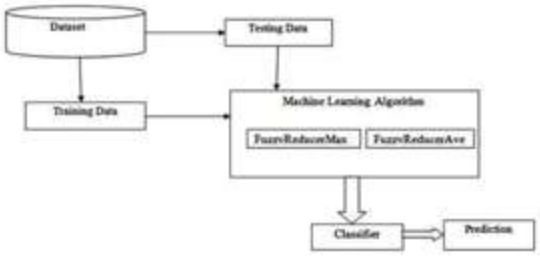


Fig. 1. Classifier Steps

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

A. Building the Classifier or Model

- Learning phase
- Building the classifier using algorithm.
- Classifier built from training set
- Training set referred as category /class

B. Using Classifier for Classification

A classifier is derived after the learning algorithm works on training set of data. Using the test set the 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 [12] 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.

D. *A Linguistic Fuzzy Rule Base Classification*

.

A FRBCS is composed by two elements: Inference system and the knowledge base. In a linguistic FRBCS the KB is formed from the database which contains the membership functions of the fuzzy partitions associated to the input attributes and the rule base associated with the fuzzy rules 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.

III. MAP REDUCE PRINCIPLE USED IN BIGDATA

The MapReduce 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 traNsferred 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 intermediate key-value pairs calculated previously summed up by the key values and generate an output value that becomes the output of the method.

Fig 2 below depicts a standard MapReduce technique with its map and reduce steps.k and v indicates to the original key-value pair. k^i andv^i 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.

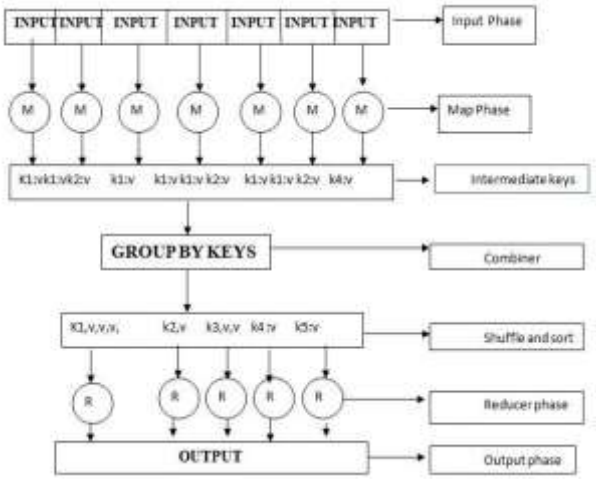


Fig. 2.Mapreduce programming model

Hadoop is the most suitable and very known among developers for implementation of the MapReduce programming model [12]. 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 [16], 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

A. *A linguistic fuzzy rule-based classification method*

In 2013, Jose Antonio Sanz et al. [] proposed a linguistic fuzzy rule-based classification method based on a new completely interval-valued fuzzy reasoning method was proposed. This inference process used interval-valued restricted equivalence functions to increase the relevance of the rules in which the equivalence of the interval membership degrees of the patterns and the ideal membership degrees is greater, which is a desirable behavior. Interval-valued fuzzy sets have proven to be an appropriate tool to model the system uncertainties and the ignorance in the definition of the fuzzy terms. Furthermore, the parameterized construction of these

functions allows us to compute the most suitable set of IV-REFs to solve each specific problem.

B. A fuzzy association rule-based classification method

In 2011, Jesus Alcal a-Fdez et al.[],proposed a fuzzy association rule-based classification method for high dimensional problems and accurate and compact fuzzy rule based classifier established. The cost was low for computation. The method relies on Improved Weighted Relative Accuracy measure in which relevant rules which were best were preselected.

C. A Rule-Based Classification Algorithm

In 2009, Biao Qin et al. [15], proposed a Rule-Based Classification Algorithm for Uncertain Data, an approach for handling uncertain data. 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.

D. A Fuzzy Unordered Rule Induction Algorithm

In 2009, Jens Hühn and EykeHüllermeier [16], proposed a Fuzzy Unordered Rule Induction Algorithm, which builds upon the RIPPER interval rule induction algorithm was proposed. The model built by FURIA uses fuzzy rules of the form given in (1) using fuzzy sets with trapezoidal membership functions. Specifically, FURIA builds the fuzzy RB by means of two steps.

1) 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.

2) Obtain the fuzzy rules by means of fuzzifying the final rules from the modified RIPPER algorithm in a greedy way.

E. An upgraded version of FRBCS

In 2005,H. Ishibuchi et al.[17] proposed adopted from Ishibuchi and Nakashima. 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.

F. Ant Colony-based Data Miner

In 2002, R. S. Parpinelli et al. [18] proposed an algorithm for data mining called Ant-Miner (Ant Colony-based Data Miner). The goal of Ant-Miner is to extract classification rules from data. Real ant colonies and data mining concepts is focused for the algorithm. Ant-Miner seems particularly advantageous when it is important to minimize the number of discovered rules and rule terms,for improvement of clarity of established knowledge.

G. An adaptive method to construct FRBCS

In 1996,Nozaki and Ishibuchi [19] have proposed an adaptive method to construct a fuzzy rule based classification system where an error correction based learning procedure adjusts the level of uncertainess of each fuzzy rule by performance of classification.

V. PROPOSED METHODOLOGY

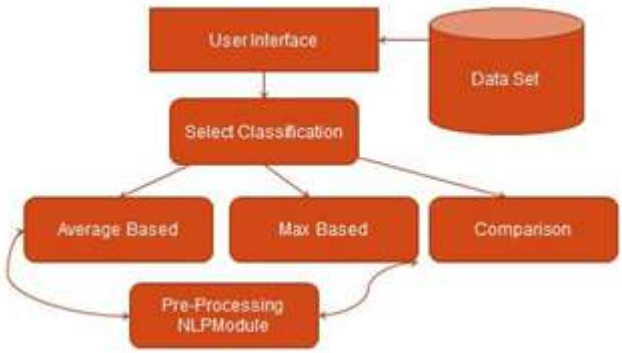


Fig. 3. Implementation flowchart

It is classified into 5 Modules:

a) User Interface for following Inputs

- Data Set file in Specific comma Separated Value format.
- Number of Entries required.
- Number of Training Set given.

b) Pre-processing NLP Module

c) Apply Semantic Classification for Average calculation

d) Apply Semantic Classification for aximum Calculation

e) 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.

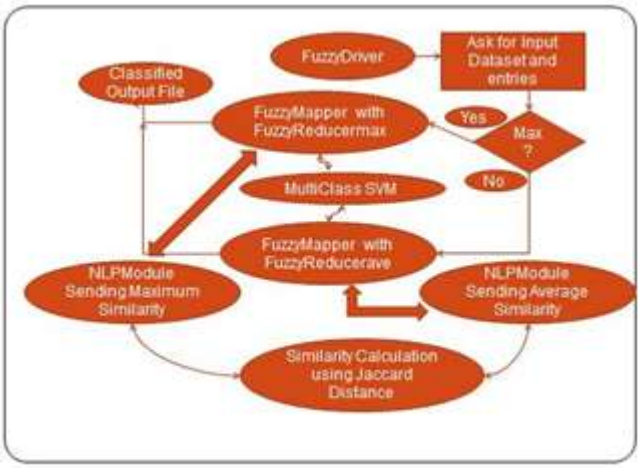


Fig. 4. Call Hierarchy of Modules

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:

1) Average Calculation :- This process is implemented in three steps:-
a) Separate strings into tokens
b) Apply Multiclass Support Vector Machine Approach
c) Calculate Similarity based using Jaccard Distance for textual data and using Euclidian distance for categorical data not containing action words and which will return a average similarity number.
2) Max Calculation:- This process is implemented in three steps
   a) Separate strings into tokens
   b) Apply Multiclass Support Vector Machine Approach
   c) Calculate Similarity based using Jaccard Distance for textual data and using Euclidian distance for categorical data not containing action words and which will return a maximum similarity number.

For detections of similarity of strings here RiWordnet library is used which makes groups of similar words based on key patterns provided as input.

VI. RESULT ANALYSIS

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

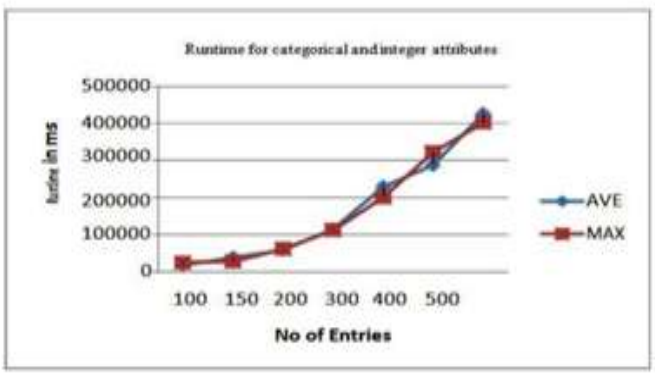
Accuracy Calculations:-

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

TABLE I

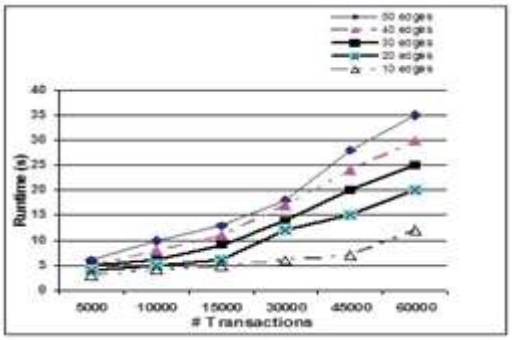
Runtime and Accuracy for AVE and MAX with Categorical, Integer attributes

	Ave		Max	
No of entries	**Runtime**	**Accuracy**	**Runtime**	**Accuracy**
500	17962 ms	100	24678ms	99
1000	38685 ms	99	26459 ms	98
2000	57758ms	100	60707 ms	89.77
3000	115448 ms	99.83	111634 ms	91.51
4000	229252 ms	100	200583 ms	92.25
5000	287253 ms	100	322781ms	88.81
6000	426260ms	100	402417ms	90.92
Average	293154.5ms	99.83	164180ms	92.89



The above table depicts the runtime and accuracy for number of entries from the dataset file having categorical and integer type of attributes, for both the algorithms AVE and

MAX.Further figure shows the graph of runtime for both the algorithms for comparison.



Runtime without Map Reduce Implementation

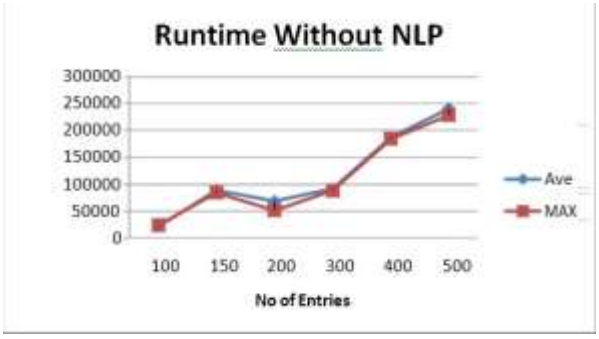
The above graph is an excerpt of classification technique used in Web Usage mining where Mapreduce concepts are not infused in implementation.

TABLE II

No of entries	AVE		MAX	
	Runtime	Accuracy	Runtime	Accuracy
100	21596ms	90.47	23893ms	57.14
150	87725ms	38.70	83915ms	38.70
200	67818ms	43.90	50855ms	39.02
300	91533ms	59.01	87781ms	47.54
400	186632ms	49.32	184496ms	41.97
500	240200ms	33.67	228173ms	33.67
Average	115917.3ms	52.51	109852.2ms	43.00

Time taken and Accuracy without Pre Processing for Ave and MAX

The above table shows the the runtime and accuracies for various entries in the text file when pre processing is not applied.It is observed that time taken by records to classify is more in both the case of algorithm AVE and MAX.Also,it is observed that AVE is having better accuracy than MAX.If we compare the time spent by the algorithms to classify the datasets,AVE is taking more time than the algorithm MAX.Hence,we use a pre processing technique, which classifies the records within less time.



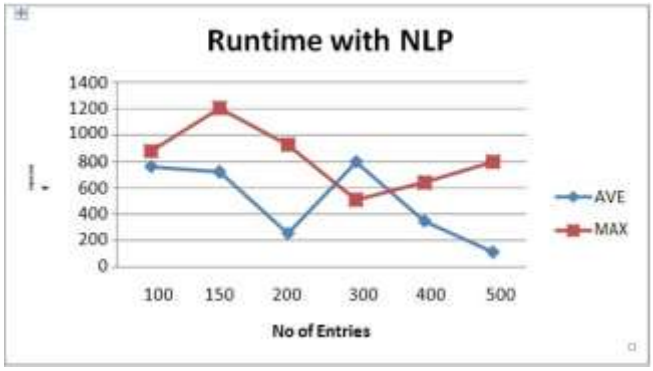
Runtime of data entries without pre-processing

TABLE III

	Ave		Max	
No of entries	Runtime	Accuracy	Runtime	Accuracy
100	760 ms	61.904762	877 ms	57.142857
150	720 ms	54.83871	1204 ms	45.16129
200	254 ms	53.658535	925 ms	53.658535
300	797 ms	31.666666	507 ms	30.0
400	346 ms	26.58228	639 ms	11.111111
500	112 ms	15.306123	797ms	13.26
Average	498.16ms	40.65	824.83ms	35.05

Time taken and Accuracy using NLP Pre Processing for Ave and MAX

The Table III. shows the runtime and accuracy for number of entries from the datset file.From the table it is clear that the time taken to classify the datasets is less because NLP technique is applied.Where as the Table II shows that the time taken is more for classifying the datasets.Also ,the average shows that the algorithm AVE is performing better in terms of both time and accuracy than the algorithm MAX.Accuracy is considered for the number of dataset entries correctly classified.Also,the table is depicted in the form of a graph as shown below.



Runtime of data entries with NLP pre-processing

VII. Conclusion And Future Work

With the help of experimental results we have shown the impact of mapreduce concept over a classification of big set of data with short span of time. It speed up the process without compromising with the accuracy of the desired outcome. It is observed that for the network dataset which contains categorical and integer attributes with each record in comma separated format with the class values. The algorithm AVE performs better in terms of accuracy than algorithm MAX, whereas in terms of time spent for processing the dataset, algorithm MAX performs better than the algorithm AVE. The algorithm MAX takes less runtime as compared to AVE. When the Reuters dataset is considered, in which classification

is performed on dataset using both algorithms, with NLP and without NLP, it is observed that with NLP as a pre processing technique both the algorithms perform better in terms of time spent than the classification without pre processing technique.

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 Systems.,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] Victoria L´opez, Sara del R´ıo, Jos´e Manuel Ben´ıtez and Francisco Herrera, "A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules," IEEE International Conference on Fuzzy Systems.,2014.

[7] T. White, "Hadoop, The Definitive Guide,"Oreilly Media, Inc., (2012)

[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] 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.

[10] Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simplified Data Processing on Large Clusters," Google, Inc.

[11] Jasmine Zakir , Tom Seymour and Kristi Berg,"Big data analytics," Issues in Information Systems, vol 16, Issue II, pp. 81-90, 2015

[12] S. Owen, R. Anil, T. Dunning and E. Friedman, Mahout in Action. Manning Publications Co., 2011

[13] Jose Antonio Sanz, Alberto Fernandez, Humberto Bustince and Francisco Herrera,IVTURS: a linguistic fuzzy rule-based classification system based on a new Interval-Valued fuzzy reasoning method with Tuning and Rule Selection

[14] 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.

[15] Biao Qin, Yuni Xia, "A Rule-Based Classification Algorithm for Uncertain Data" Department of Computer Science, Indiana University ,Purdue University Indianapolis, USA

[16] Jens Hühn and Eyke Hüllermeier, "FURIA: An Algorithm For Unordered Fuzzy Rule Induction" Philipps-Universität Marburg,Department of Mathematics and Computer Science,2009

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

[18] Rafael S. Parpinelli, Heitor S. Lopes, "Data Mining With an Ant Colony Optimization Algorithm", IEEE Transactions On Evolutionary Computing, Vol. 6, No. 4, August 2002

[19] K. Nozaki, H. Ishibuchi, H. Tanaka, 1996Adaptive Fuzzy Rule-Based Classification System",Dept. of Ind. Eng., Osaka Prefecture Univ., Japan vol. 4, no. 3, pp. 238-50, Aug. 1996

[20] DanilZburivisky. "Hadoop Cluster Deployment" PacketPublishing, Mumbai"

[21] Janmenjoy Nayak, Bighnaraj Naik and H.S.Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks : Applications & Challenges" International Journal of Database Theory and Application Vol. 8 , No.1, pp.169-186,2015

[22] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya, "Preprocessing Techniques for TextMining - An Overview, International Journal of Computer Science & Communication Networks,Vol 5(1),7-16

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

[24] S.Jusoh and H.M. Alfawareh, Natural language interface for online sales, in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007). Malaysia: IEEE, pp. 224–228 ,November 2000.

[25] 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.