Ashu Jain et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 1, March 2018, pp. 123-126

Scrutinizing Big Data using Machine Learning Classifiers

Ashu Jain[1#], Soumya Jain[2*], Charul Dewan[3%]
[1,3]Department of Information Technology, Northern India Engineering College, New Delhi, India
[#]ajainashu@gmail.com
[%]charularora@gmail.com
[2]University School of ICT, Guru Gobind Singh Indraprastha University, New Delhi, India
[2]soumya.jain06@gmail.com

Abstract—**The massive unstructured and semi structured heterogeneous data generated from devices, household appliances and from day to day activities namely sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few collected together is what we call as big data. Machine learning's supervised and unsupervised learning techniques can be used to process this large amount of data using the various classifiers to derive useful insight from the data and to predict the future patterns and trends. Supervised learning techniques uses the concept of train and test data i.e., some portion of data is first trained with expected results for a given input and then the remaining chunk of data is used to test the algorithm for prediction accuracy. J48, IBK & Naïve Bayes are few of the supervised learning classifiers used in this comparative analysis. The classifiers are used on the multivariate big data set selected under different models of training and test data to obtain the best performing classifier among them on basis of correctly classified instances (one of the performance measuring criteria). On the basis of the experiments conducted, J48 is the classifier which is generating the maximum number of correctly classified instances. Hence it can be considered as the best classifier to process the data of given kind.**

Keywords—***Big Data, Machine learning, Accuracy, Classification, WEKA***

I. INTRODUCTION

With regards to two of the most discussed innovations in current times, Big Data and Internet of Things are maybe comfortable steerage. Furthermore, finished the most recent couple of years, Big Data has gained ground in various areas. Also, despite the fact that Internet of Things happens to appear as something else, it is enormously connected to Big Data. According to IBM's examination, consistently, we make 2.5 quintillion bytes of information — so much that 90% of the information on the planet today has been made over the most recent two years alone. This information originates from all over: sensors used to assemble atmosphere data, presents via web-based networking media destinations, computerized pictures and recordings, buy exchange records, and mobile phone GPS signs to give some examples. This information is the thing that we call as large information.

We can unravel that Big Data is fundamentally a ton of information that is in complex edge. In particular, it likewise alludes to the utilization of prescient examination and techniques that permits extraction of significant data from such data. This permits better basic leadership, lessening in operational expenses and dangers for associations.

Gartner's Analytics predicted in their literature[1] about the trend of Internet of things(IoT) which would be just cheap, small devices which will be having a radio and GPS capability. They could be referred to as self-assembling mesh networks which would be aware of their location. These things would be connected through a network i.e., either through the Internet, in that case would be having an IP address or through a connection of sensor's which would make these devices trackable. These devices would include day to day usage household devices spread everywhere be it the newly launched automobile or the street lights. It's not a single technology, it's a concept. Driving the pattern are things like installed sensors, picture acknowledgment, enlarged reality, close field correspondence.

The huge unstructured and semi organized heterogeneous information produced from these gadgets is additionally alluded to as large information. In this manner IoT and huge information have an interdependency. The result is situational decision support, asset management and more transparency.

IoT would be and is generating numerous business opportunities but it all adds to the increasing complexity of IT.

The rest of the paper is organized as follows: Section 2 describes the related work done in the field of big data. Section 3 gives the overview of the classifiers used. Section 4 discusses the empirical data collection along target variable selected.. Section 5 includes the results. In section 6, conclusion and directions of future work are provided.

II. RELATED WORK

In this section, work done in the field of big data has been presented. Chen et al[2] reviewed about the big data and introduced the technologies related to big data. They concluded with the open problems and future scope of big data. Gohar et al[3] presented the big data analytics architecture for internet of small things (IoST) they

presented the detailed analysis of a big data implementation of the IoST used to track humidity and temperature via Hadoop. Huang and Gong[4] studied about the data mining engine based on big data. They designed and implemented some parallel data mining algorithms by using spark as the engine core and programming model. Malhat et al[5] proposed an operational and unified framework to balance between reduction rate and classification accuracy. The results suggested that the class-balanced splitting process is preferred regarding the classification accuracy criterion. Pradeep et al[6] discussed the growth, life cycle stages, handling of huge data on Hadoop framework, publication frequencies, advantages and challenges in handling big data. Dissanayeke and Jayasena[7] focused on how to analyze the massive and heterogeneous data of the internet of things in a proper way. In this paper big data has been analyzed using the machine learning classifiers.

III. CLASSIFIERS USED

A. *J48*

The J48 Decision tree classifier takes a shot at the accompanying calculation: with a specific end goal to group another thing, it first needs to make a decision tree in view of the quality estimations of the accessible preparing information. Thus, at whatever point it experiences an arrangement of things (preparing set) it distinguishes the property that segregates the different instances generally plainly. This element that can reveal to us most about the information instances with the goal that we can order them the best is said to have the most noteworthy data pick up. Presently, among the conceivable estimations of this element, if there is any an incentive for which there is no vagueness, that is, for which the information instances falling inside its classification have a similar incentive for the objective variable, at that point we end that branch and dole out to it the objective esteem that we have gotten.

For alternate cases, we at that point search for another quality that gives us the most elevated data pick up. Consequently, we proceed in this way until the point when we either get a reasonable decision of what blend of traits gives us a specific target esteem, or we come up short on properties. If we come up short on characteristics, or on the off chance that we can't get an unambiguous outcome from the accessible data, we relegate this branch an objective esteem that most of the things under this branch possess. Now that we have the decision tree, we take after the request of trait choice as we have acquired for the tree. By checking all the individual traits and their esteems with those found in the decision tree display, we can appoint or anticipate the objective estimation of this new example.

B. *IBK*

In this order strategy, classifier stores the highlights and the class name of the preparation sets. New questions are characterized in view of the voting criteria. It gives the most extreme probability estimation of the class. Euclidean separation measurements is utilized for allotting articles to the most as often as possible named class. Separations are ascertained from all preparation articles to test question utilizing proper K esteem. In this paper K esteem is doled out to 100 which demonstrate that the picked class mark will choose preparing object inside its 100 closest neighbors.

C. *Naïve Bayes*

The Naïve Bayes classifier takes a shot at a straightforward, yet nearly instinctive idea. Additionally, now and again it is likewise observed that Naïve Bayes outflanks numerous other similarly complex calculations. It makes utilization of the factors contained in the information test, by watching them exclusively, free of each other. It depends on the Bayes manage of contingent likelihood. It makes utilization of the considerable number of characteristics contained in the information, and investigations them exclusively just as they are similarly critical and autonomous of each other.

IV. EMPIRICAL DATA COLLECTION

A. *Dataset used*

For this performance analysis task of various classifiers on a big data set, the data containing general demographic information on internet users in 1997 which was collected and obtained as part of a survey conducted by the Graphics and Visualization Unit at Georgia Tech October 10 to November 16, 1997[8].

This data set has following features:

FEATURES OF THE KDD INTERNET USAGE DATASET

Data Type:	Multivariate, Nominal
Number of Instances:	10108
Number of Attributes:	72
Number of Classes:	Varying as per attribute

As part of this survey by GVU, the internet users were given a survey to fill and the data was then classified into 72 attributes based upon the values provided as answers to the survey questions.

We will use the provided attributes to classify the variables.

V. PERFORMANCE ANALYSIS OF THE CLASSIFIERS USING WEKA

For the analysis of data, machine learning tool weka is used. After loading the data, we are selecting a target attribute so that we can divide the data into possible number of classes.

Selecting the Target attribute as: Major_Occupation, We can see the dataset gets classified into 5 different classes, as represented in Fig 1. Also we can visualize the same through graph presented in Fig 2.

VI. RESULTS

In table II, the performance of the machine learning classifiers on the selected data has been presented.

TR – represents the Correctly classified instances obtained from the training data.

TST – represents the Correctly classified instances obtained from the test data

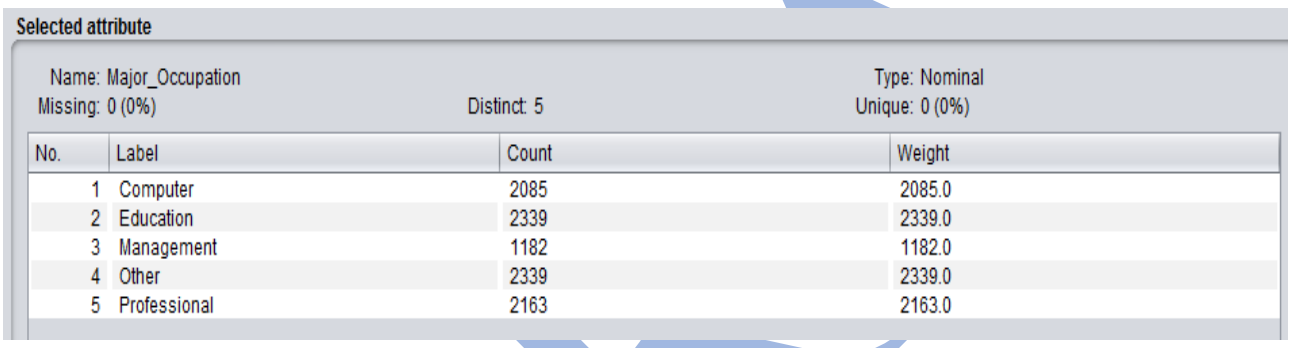

Fig.1 Classification of big data set for Target attribute : Major_Occupation

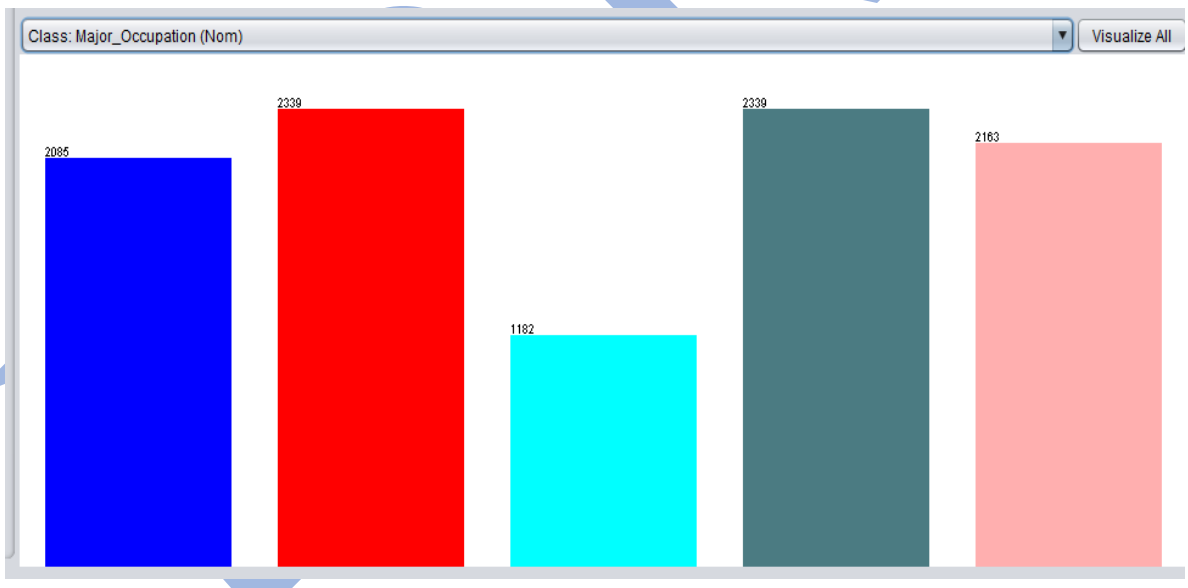

Fig. 2 Visualization of the data set on various classes obtained from target attribute

PERFORMANCE ANALYSIS OF THE CLASSIFICATION TECHNIQUE

Classification Technique used	Correctly classified instances		
	10-Fold Cross-Validation Model (%)	Train and Test Model (70-30 Ratio	
		TR (%)	TST (%)
J48	82.2517	88.9203	82.2559
IBK	57.222	57.6314	56.8931
Naïve Bayes	78.3043	82.6314	78.1332

Table III represents the Mean Absolute Error(MAE) and Root mean squared Error(RMSE) values on the considered dataset.

MEAN ABSOLUTE ERROR AND ROOT MEAN SQUARED ERROR

Classification Technique used	Mean Absolute Error	Root Mean Squared Error
J48	0.083	0.23
IBK	0.25	0.34
Naïve Bayes	0.09	0.25

VII. CONCLUSION AND FUTURE SCOPE

In this comparative analysis of the results obtained by applying some of the classification techniques to the big data set selected under different models of training and test data (part of supervised learning), we can infer that J48 classification technique (tree based) provided the highest number of correctly classified instances under both the models in consideration i.e., the 10 Fold cross-validation model and the manual train and test model where 70% of the dataset instances were selected as part of training data and rest as part of test data. Also, it could produce 0.0042% improvement even when the training data was reduced by 20% of the actual data. The minimum amount of deviation in producing correctly classified instances from the test data after training can be seen for IBK classification technique which signifies the accuracy of the training technique.

On the basis of MAE and RMSE values, it can be inferred that J48 performs the best. Hence is the most accurate out of the three considered machine learning classifiers .

In this study, only three machine learning algorithms have been applied. There are various other algorithms which are needed to be explored on the big data. In the future study, machine leaning and evolutionary algorithms can be applied on the big data.

REFERENCES

[1] Eric Savitz, (2012) "Gartner: 10 Critical Tech Trends For The Next Five Years" http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years/#7f54cd124c6f [accessed on 06.01.17].

[2] Chen, M., Mao, S., & Liu, Y. (2014), "Big data: A survey". In Mobile Networks and Applications, 19(2), 171–209, Springer Science+Business Media New York 2014.

[3] M Gohar, S. H. Ahmed, M. Khan." A Big Data Analytics Architecture for the Internet of Small Things", IEEE Communications Magazine (Volume: 56, Issue: 2, Feb. 2018)

[4] X. Huang , S. Gong, "Analysis of Big-Data Based Data Mining Engine", Computational Intelligence and Security (CIS), 2017 13th International Conference, feb 2018

[5] M. Malhat, M. E.Menshawy, H. Mousa, A. E. Sisi, "Improving instance selection methods for big data classification", Computer Engineering Conference (ICENCO), 2017 13th International, Feb 2018

[6] S. Pradeep , J. S. Kallimani, "A survey on various challenges and aspects in handling big data", Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017 International Conference, Feb 2018

[7] D. M. C. Dissanayake , K. P. N. Jayasena, "A cloud platform for big IoT data analytics by combining batch and stream processing technologies", Information Technology Conference (NITC), 2017 National, Feb 2018

[8] Hettich, S. and Bay, S. D. (1999). The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science.

