THINK INDIA JOURNAL

ISSN:0971-1260
Vol-22- Issue-14-December-2019

Machine Learning Algorithms on Big Data Analytics

Sharon Susan Jacob[1], **R.Vijayakumar**[2]

School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
[1] sharonsusanjacob@gmail.com, [2]vijayakumar@mgu.ac.in

ABSTRACT

Big Data is used for a collection of data sets that are large and complex, which is difficult to store and process using available database management tools or traditional data processing applications. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big data is typically broken alone by these characteristics: volume, velocity and variety. *Volume* refers to the amount of data that is getting generated. *Velocity* point out the speed at which data is getting generated. *Variety* indicates the different types of data that is getting generated. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. Emerging companies needed to find new technologies that would allow them to store, access, and analyze huge amounts of data in near real time so that they could monetize the benefits of owning this much data about participants in their networks. In particular, the innovations *Map Reduce*, *Big Table*, and *Hadoop* proved to be the sparks that led to a new generation of data management. Big Data could be of three types: Structured, Semi-Structured and Unstructured. Twitter is one of the famous social networks worldwide and Twitter tweets and other social media posts are some of the central source of unstructured data. With the help of machine learning algorithms, it is possible to change unstructured data into organized form. In this work, twitter data was taken as the dataset. For training data, supervised machine learning algorithms like Linear Support vector and Naïve Bayes are used. For analyzing the data, Python programming was used. The result shows that Naïve Bayes classifier yielded more classification accuracy than Linear Support vector classifier.

Keywords—*Big Data, Hadoop, Linear Support vector, Map Reduce, Naive Bayes*

1. INTRODUCTION

Big Data refers to the large amounts of data which is pouring in from various data sources and has different formats. Even previously there was huge data which were being stored in databases, but because of the varied nature of this Data, the traditional relational database systems are incapable of handling this Data. Big Data is much more than a collection of datasets with different formats, it is an important asset which can be used to obtain enumerable benefits.

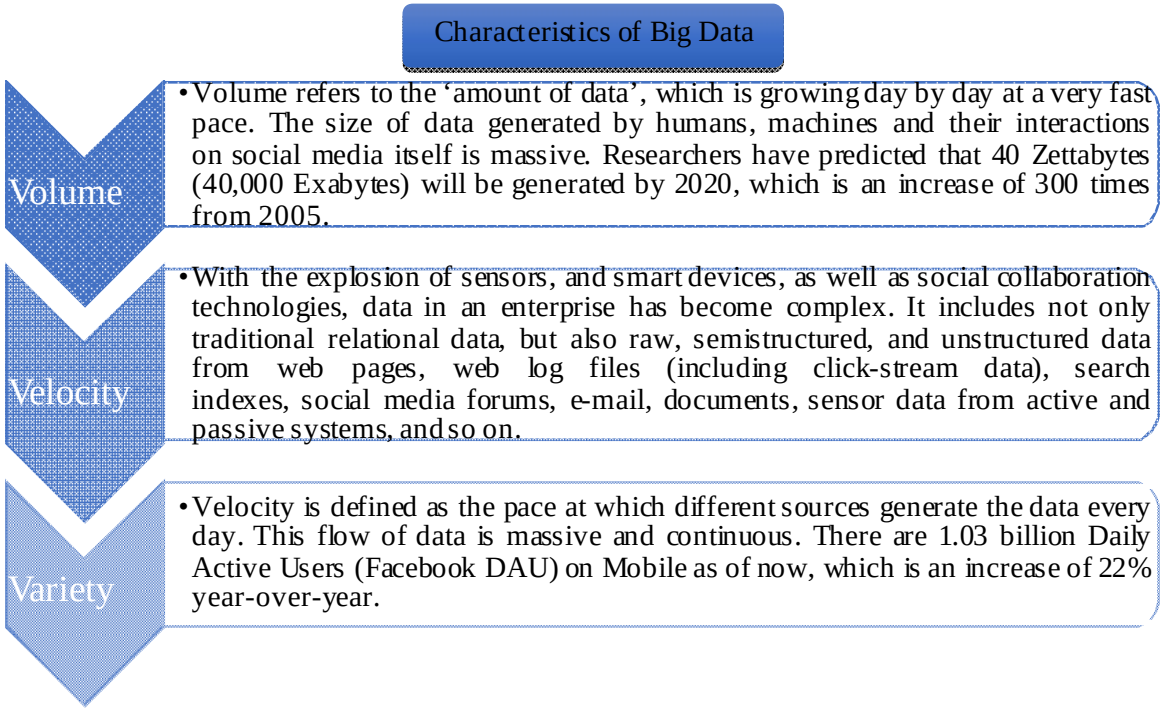
 Copyright © 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

As companies begin to evaluate new types of Big data solutions, it unfold many new opportunities in the field. Manufacturing companies may be able to monitor data coming from machine sensors to determine how processes need to be modified before a catastrophic event happens. It will be possible for retailers to monitor data in real time to upsell customers related products as they are executing a transaction. Big data solutions can be used in healthcare to determine the cause of illness and provide a medical person with guidance on treatment options.

1.1 Characteristics of Big Data

Big data is being generated by everything around us and is not a single technology but a combination of old and new technologies that helps companies gain actionable insight. Big data is typically broken down by three characteristics:



Additional characteristics recently introduced about Big data:

- *Validity*: correctness of data
- *Variability*: dynamic behaviour
- *Volatility*: tendency to change in time
- *Vulnerability*: vulnerable to breach or attacks

Copyright © 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

- *Visualization*: visualizing meaningful usage of data

2. INNOVATIONS TO DATA MANAGEMENT

Emerging companies needed to find new technologies that would allow them to store, access, and analyze huge amounts of data in near real time so that they could monetize the benefits of owning this much data about participants in their networks. Their resulting solutions are transforming the data management market. In particular, the innovations MapReduce, Big Table, and Hadoop proved to be the sparks that led to a new generation of data management. These technologies address one of the most fundamental problems — the capability to process massive amounts of data efficiently, cost effectively, and in a timely fashion.

- **MapReduce**

MapReduce was designed by Google as a way of efficiently executing a set of functions against a large amount of data in batch mode. The "map" component distributes the programming problem or tasks across a large number of systems and handles the placement of the tasks in a way that balances the load and manages recovery from failures. After the distributed computation is completed, another function called "reduce" aggregates all the elements back together to provide a result.

- **Big Table**

Big Table was developed by Google to be a distributed storage system intended to manage highly scalable structured data. Data is organized into tables with rows and columns. Unlike a traditional relational database model, Big Table is a sparse, distributed, persistent multidimensional sorted map. It is intended to store huge volumes of data across commodity servers.

- **Hadoop**

Hadoop is an Apache-managed software framework derived from MapReduce and Big Table. Hadoop allows applications based on MapReduce to run on large clusters of commodity hardware. Hadoop is designed to parallelize data processing across computing nodes to speed computations and hide latency. Two major components of Hadoop exist: a massively scalable distributed file system that can support petabytes of data and a massively scalable MapReduce engine that computes results in batch.

 Copyright © 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

3. BIG DATA ANALYTICS

Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights. Basically, Big Data Analytics is largely used by companies to facilitate their growth and development. This majorly involves applying various data mining algorithms on the given set of data, which will then aid them in better decision making.The possible kinds of big data analysis are given below:

Table 1. Kinds of BigData Analysis

Analysis Type	*Description*
Basic analytics for insight	Slicing and dicing of data, reporting, simple visualizations, basic monitoring.
Advanced analytics for insight	More complex analysis such as predictive modeling and other pattern-matching techniques.
Operationalized analytics	Analytics become part of the business process.
Monetized analytics	Analytics are utilized to directly drive revenue.

A comparison was done regarding the performance of machine learning algorithms on Big data.

Data Set

Large amount of data is increasing every day because of social media and Twitter is one of the social media site which gives an opportunity to people to express their ideas and opinions about a particular topic. Twitter tweets and other social media posts are some of the central source of unstructured data. So twitter data is taken as the Data set.

The data set after pre-processing and feature extraction, the data is given to machine learning model. The *pre-processing* involves removal of unimportant features from the data. The steps involved aim for making the data more machine readable in order to reduce ambiguity in feature extraction. Pre-processing involves removal of retweets, stemming, lemmatization etc. Selection of useful words from tweets is *feature extraction*.

4. MACHINE LEARNING ALGORITHMS APPLIED FOR TRAINING THE DATA SET

Linear Support vector and Naïve Bayes classifier are the algorithms implemented for preparing and training the data. In ***Linear Support vector*** algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. It tries to find a hyper-plane that

 Copyright © 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

maximizes the margin of training data. The training examples that closest to hyperplane are called support vectors since they are supporting the margin (as the margin is only a function of support vectors).***Naïve Bayes*** is a probabilistic technique for constructing classifiers. The characteristic assumption of the naïve Bayes classifier is to consider that the value of a particular feature is independent of the value of any other feature, given the class variable. An advantage of naïve Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification and that the classifier can be trained incrementally. It is a classification technique based on Bayes' theorem. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

P(c/x) = P(x/c)P(c)/P(x)

where,

P(c|x) is the posterior probability of class (target) given predictor (attribute), P(c) is the prior probability of class, P(x|c) is the likelihood which is the probability of predictor given class and P(x) is the prior probability of predictor.

5. ANALYSIS OF THE DATA

Twitter sentiment analysis was done using the language Python which provides high quality implementation of numerous data analysis and visualization method, from regression to statistics, text analysis and much more. In using the Language Python, the packages SCIKIT-LEARN, NumPy and Pandas are used. The ***Scikit-learn*** is a powerful library that provides many machine learning classification algorithms, efficient tools for data mining and data analysis. ***NumPy*** provides a high-performance multidimensional array object, and tools for working with these arrays. ***Pandas*** is a software library written for the Python programming language and it offers data structures and operations for manipulating numerical tables and time series.

6. PERFORMANCE EVALUATION

The performance analysis of Linear Support vector and Naïve Bayes classifier on Big data is done. The algorithms were employed with the collected samples of datasets containing 1 lakh samples of twitter information. The metrics used here for evaluating the performance are accuracy, precision, recall and F1-score.

$$Precision = \frac{tp}{tp+fp} (1)$$

 Copyright © 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

$$Recall = \frac{tp}{tp+fn} \quad (2)$$

$$F1\,Score = 2 * \frac{(recall * prcision)}{(recall + prcision)} \quad (3)$$

where, t_p is true positive which correctly predicted positive values, t_n is true negative which correctly predicted negative values, f_p is false positive which is falsely predicted positive class and f_n is falsely predicted negative class. The metric representation is demonstrated in the table below:

Table 2. Metric Representation of Big Data (Twitter data)

Technique	Accuracy	Precision	Recall	F1 score
Linear Support Vector	66%	.71	.66	.62
Naïve Bayes	73%	.75	.74	.73

The accuracy obtained with Naïve Bayes classification technique on Bigdata is 73% whereas Linear Support vector produces 66% accuracy. The precision and recall measures obtained by the Naïve Bayes method are 0.75 and 0.74 where as Linear Support vector produced 0.71 precision and 0.66 recall value. The F1 score obtained by the Naïve Bayes method is 0.73 whereas Linear Support vector produces 0.62 score. Thus it is evident from the table that Naïve Bayes Classifier yielded more classification accuracy than Linear Support vector classifier.

7. CONCLUSION

Big data mining is an emerging trend in the field of data science. It is the arising need for various engineering domains. Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy. The classification accuracies of Naïve Bayes and Linear Support vector classifier on the twitter data is compared and the result shows that Naïve Bayes yielded more classification accuracy than Linear Support vector classifier.

REFERENCES

[1] S. R. Alty; S. C. Millasseau; P. J. Chowienczyc; A. Jakobsson "Cardiovascular disease prediction using support vector machines"*2003 46th Midwest Symposium on Circuits and Systems. DOI: 10.1109/MWSCAS.2003.1562297*

[2] Qiaowei Jiang; Wen Wang; Xu Han; Shasha Zhang; Xinyan Wang; Cong Wang. "Deepfeature weighting in Naive Bayes for Chinese text classification" *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS).DOI:10.1109/CCIS.2016.7790245.*

[3] Jiawei Han, Micheline Kamber and Jian Pei, "DATA MINING Concepts and Techniques",

 Copyright © 2019Authors

ISSN:0971-1260
Vol-22- Issue-14-December-2019

Morgan Kaufman Publishers, USA, (2002).
[4] B. Giraldo, A. Garde, C. Arizmendi, R. Jan,S. Benito, I. Diaz, D. Ballesteros. "Support Vector Machine Classification Applied on Weaning Trials Patients."*2016 IEEE*.
[5] Geetika Gautam, Divakar Yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 2014 Seventh International Conference on Contemporary Computing (IC3), August 2014, pg.437-442.

 Copyright © 2019Authors