RESEARCH REVIEW JOURNALS
Volume-03
Issue-09
September-2018

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Comparative study of different Big Data Analytics Techniques for Unstructured Big Data

[1]Kanwaldip Kaur & [2]Dr. Rajan Manro

[1]*Research Scholars (Deptt. of Computer Science), Desh Bhagat University, Mandi Gobindgarh, Punjab (India)*
[2]*Desh Bhagat University, Mandi Gobindgarh(Punjab)*

ARTICLE DETAILS

Article History
Published Online: 07 September 2018

Keywords
Big data, Clustering, classification, SVM, Machine learning

ABSTRACT

Big data is the current scenario requirement. Various social and business applications are producing large amount of data with great frequency. This large data processing is the most difficult task. Because extraction of the useful data will be some time cumbersome for the extraction of most suitable results. Because the results generated can enhance the future planning. Any type of suitability technique according the data nature is most difficult task. In current time for the processing of the data machine learning based tools are used. They works in two scenarios like supervised learning and other is unsupervised learning models. Supervised learning model based on building training set for the training purpose. So that machine can learn from the training and under stands the rules and regulations. So that while having testing set the learning can be applied on the big data to extract the useful aspect. Data either on health parameters of patient or on social networking related data machine learning based model is the most suitable technique. These techniques can further be enhanced to produce more efficient data.

1. Introduction

As we are moving forward large number of applications are emerging on daily basis. These applications can be business applications or the social applications. Whatever type of applications is the large amount of data is being generated on daily basis. The data generated from various applications can be structured data or may be unstructured data. Structured data is single type of sequential data. But unstructured data is non sequential and multiple types of data. Unstructured data is difficult to process and store. Because the single type of approach may not be suitable for the processing of unstructured data. But the selection of appropriate type of technique for analysis purpose can resolve the problem of inefficient processing. A efficient processing technique can process the large amount of data to extract useful pattern and facts. These patterns and facts can be used for the early warning system for both social and business events. For example disease outbreak. If the appropriate technique will be selected then the early warning system can be given. Which can reduce the effect of the disease on the persons. Various techniques which falls in the category of the machine learning and classification can be taken which can extract the useful facts from the randomly collected data[5].

1.1 Data Analytic Techniques

There are various data analytic techniques that can be used for the processing of the large bulk amount of data. So that useful patterns can be identified. Any technique cannot be said to be universal. Rather in different scenarios different techniques may be suitable. Each technique selected has its own set of advantages and disadvantages[2].

1.1.1 Classification Analysis

It is the technique used for reducing the complexity of the bulk data. It classify the data into different classes based on the characteristics of the data. For example large bulk of data regarding Airlines time schedules. The classifier will classify the data into two different classes like late flights and another class is on time flights. Each class will be having its own set of entries. For the reduction of the complexity there on late flights can be classified into different classes like late flights, highly late flights and canceled flights. That way the data will be put into hierarcal classes. Later on using any machine learning approach useful patterns can be identified[11].

1.1.2 Association Rule Learning

It is another technique used for the analytical purpose. While having big data the data is unstructured. That means the collected data hardly has any relation. Because data is collected from random sources and also being collected randomly. Association Rule learning technique identifies the relation between the random variables which otherwise looks to be different. Once the association is developed it is easy to drive the new facts and figures for the purpose of analysis. programmers use association rules to make programs capable of machine learning[9].

1.1.3 Anomaly Or Outlier Detection

While having bulk amount of data large number of entries lies into the dataset which does not suits to the rest of the entries. Such entries which does not suits to the remaining entries will be called as outlier. Selected technique identifies the outlier and remove those unsuitable entries. This in results a normalized entries. This type of technique may also be called as filtering technique. Which will filter the data based on outlier identification[6].

1.1.4 Clustering Analysis

It is another technique for the analysis. First and foremost work is to identify the clusters based on data natural behavior.

Now put each data item into appropriate cluster. This will automatically aggregate the look alike entries into single place. Later on using some machine learning process extract the useful facts for generating the reports[6].

1.1.5 Regression Analysis

It is the technique identifies the relation between different variables and drive the dependency amongst them. How the variation in one variable will be taken place as other variable do change. This type of technique is useful for the areas where prediction is required for the future course of action like disease outbreak. Driving the variables responsible for the disease. In second step device the relation between these variables that how there variation has positive or negative impact on the disease pattern[6].

1.2 Classification Techniques

1.2.1 SVM

It is the classification technique. Based on set of rules settled prior to the start of the machine the data entries are out into two different classes. Those entries which are not according to rule will falls into left hand side of the line and those data entries which are according to the rule will falls into the right hand side of the separating line[10].

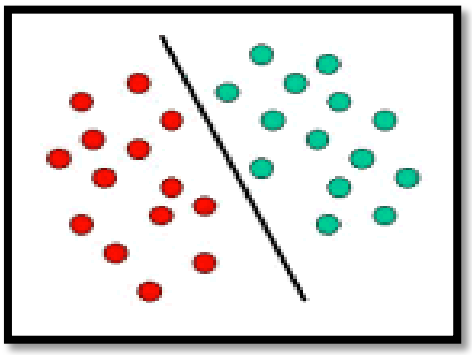

Fig. 1.1 Classifier with SVM[1]

1.2.2 DTREE

Is the decision support tool. Where tree like graph or model of choices and their attainable consequences , as well as their natural event outcome. This means whole set of graph has all the options regarding decision standard outcome and what system has evaluated. So that difference if any can be having corrective action immediately[18].

2. Literature Survey

2.1 Big Data Analytic Techniques

Modern dataset is different from the traditional dataset on three aspects like Volume, Velocity and variety. In current time large amount of data is produced at large pace in large volume. If this large data will be processed intelligently then new facts can be drawn which can be having various actionable decisions. Techniques undertaken is to take care of the gathering, storing, processing and analyzing[5].

2.1.1 Machine learning

Machine learning is the sub field of the Artificial Intelligence. It focuses on the computational model where the technique and tools will be selected by learning from the data behavior automatically. So that useful facts can be extracted. This type of technique has diverse application areas like Healthcare, science, engineering, Business and finance. It has various sub types.

i. Supervised learning

It is the type of machine learning type where the analyzed defines the training set for the given set. It train the machine that how the decision can be taken with the variation in the data items. Based on the output of the output belongs to the continuous values then it is the member of regression. But if the output belongs to the discrete set then it is the member of the classification[10].

ii. Unsupervised learning

Unsupervised learning is the process of learning with clustering. It does not has any training set settings. Rather the data will be put into different clusters based on perceived similarity. Once the association between the data items is developed then machine learning tool identify the facts and figures for the use in decision making purpose[18].

iii. Reinforcement learning

` It is the technique based on reward of punishment based learning. Each inputted entry will be processed to have its impact on the environment. The corresponding action will be either punishable or may be rewarding. If the action is right then reward is given else punishment is given[16].

iv. Deep Learning

It is the most complex type of learning technique where large amount of data inputted to the system will be having processing in multiple layers. Each layer will be small processing capability. All the processor in the respective layer will be called as neurons. These processors will works in parallel. So that multiple processors can process the inputted entries in tandem[17].

v. Association rule learning

In this learning procedure large number of entries lies into the Big data has natural relation. This natural relation will be called to be the association rule based learning. Unlike classification no extra rule is enforced on to the inputted entry.

3. Comparative Study

Author Name	Technique	Purpose	Results
Venkateswara Reddy Eluri	K-Mean clustering and canopy clustering	The aim of the research is to integrate the process of Big data and Data mining. So that useful features and facts can be identified.	K-mean produces better results in time efficient manner. Such that true facts can be represented using various levels of social and

			business applications.
R.AFadnavis and Samrudi Tabhane [13]	Hadoop architecture study has been taken for process understanding.	To explain the architecture of the Hadoop and Map reducers.	Haddop which has been taken in study has the capability to process whole architecture of the Big Data.
Fatos Xhafa, Victor Naranjo and Sanfi Caballe [14]	This paper has taken up the study of the Map reducer of Hadoop Framework.	To extract the useful and relevant facts from the Big data.	Map reducer is useful for the real time data extraction and the Decision making process.
Poonam Vashisht and Vishal Gupta([15]	Aim of the research is the process of understanding the effect of clustering on the performance of the Big data analytics	Aim is to understands the various analytic methods for the Big data. So that useful facts can be extracted.	Finally the technique has been applied on to the audio and video type of data for extraction of the useful facts.
Kyounghyun Park, Minh Chau Nguyen and HeesunWon [16]	Current research paper has taken up the study of the Hadoop Yarn.	The main of the research is to understands the Big data analytic platform.	The research has finally developed a analytic tool which can be used for the Big data Management tool.
Sruthika and N Tajunisha[17]	The technique taken up in this research paper is based on three steps. Describe, Predict, suggestions	Aim is to undertake the process of Big data analytic tools. So that better and true facts can be identified. These facts can be easy to understands and extract the useful facts.	Comparative analysis of various traditional and new age techniques. So that based on comparison appropriate technique can be selected.
Bichitra Manda ,Ramesh Kumar Saho and Srinivas Sethi[18]	This paper has researched the technique based on Hadoop architecture. Which is useful tool for the extraction of the relevant data.	Aim is to understands the architecture of the Big data and various its constraints and basic behavior.	Describes the process of basic framework of the Big data and its various inner and outer facts.
Parth Chandarana and M Vijayalakshini([19]	Aim is to identify the advantages and disadvantages of the Big Data.	Batch oriented OLTP based data processing is being used.	It will helps in processing the data based on social media. Which has bit random behavior.
Dawei Jiang et. al [20]	This research has used new technique based on epic.	Aim is to understands the random behavior of the data and deal with the normalization of the abnormal or random behaviors.	The system is developed which will be fault tolerant. Can process the random data for the purpose of the better processing.
Zoltan Prekopcsak et. a1. [21]	To Implement the new framework of the Haddop.this frame has to byepass the randomness in the data.	It is based on speedy mining of the data.	Finally it is extracted that the Hadoop is the better tool for the data processing.
Alexander Alexendrov, Rico Bergmann and Stephem even [22]	Aim is to develop the system for the parallel processing for fast incoming data.	It is based on data extraction and then later on integrate the data.	Developed a programming tool for the fast processing.
Xindong Wu, Xingquan zhU Gong Qing Wu and Wei Ding[23]	Technique is based on providing the sufficient level of security to protect the mined data to be accessed by illegal users.	Its aim is to identify the procedure which can extract the Big data analysis with suitable efficient technique	The results are based on new revolution model which is data driven.
Bama Saha and Divesh Srivastava[24]	It has implemented the statistical and logical model to data quality monitoring	Aim is to keep vigil over to the monitoring process for the quality of the information.	New rules are builded for processing of the unstructured data.
Jinsong Zhang, YanChen and Taoyiog Li [25]		Aim is to define various challenges that are being faced by Big data mining framework.	Challenges faced by the Big data while data capture, Storage, Management, Analytics are identified.

4. Conclusion

Big data is the current scenario and need for various applications like health care, Social networking, Business tec. Various applications are producing the data with great frequency and with great volume. Now problem is data mining. That means certain technique is required which can process the large data and extract various useful facts. A perfect technique does not exists which works globally. Which can process all types of the data. Suitable technique is to be selected for the processing purpose. There are various techniques which are prevalent and appropriate for the Big data processing. One such technique is machine learning tool. It has two basic way of learning based processing. One is supervised learning and other is unsupervised learning. Based

on the type of data and rate of the data at which it is coming inside to the system supervised learning is the best suited technique. It can extract the facts and figures with great sense. Because it is supervised learning based machine learning mechanism will be having training set. Which will train the system for adaption according to the current situation. Classification and clustering are another techniques which can extract the dataset and put them into different classes and put into the appropriate cluster.

5. Future Work

Currently various suitable techniques are available which can be used for processing purpose. Various types of classification and clustering based techniques are available which based on the data types clusters the data into different classes. In future various learning based machine learning tools can be adapted for the further research. So that fine tune the technique to make data extracted to be more suitable for the required situation and condition.

References

1. Stephen Kaisler, Frank Amnour, J. Alberto," Big Data: Issues and Challenges Moving Forward",46th Hawaii International Conference on System Science, IEEE,2012
2. Sam Padden, "From database to Big Data,", in IEEE Computer Society, 2012
3. Dan Garlasu, "Data Implementation Based on Grid Computing",
4. Avita Katal, Mohammad Wazid and R H Goudar, "Big Data: Issues, Challenges, Tools and good Practices", in IEEE 2013
5. Seref Sagiroglu and Duygu Sinang, "Big Data A Review",IEEE,2013
6. Yuri Demchenko, Paolo Grosso and Cees de Laat, "Addressing Big Data Issues in Scientific Data Infrastructure", in IEEE 2013
7. Parth Chandarana and M Vijayalakshrni, "Big Data Analytics Framework", in International Conference on Circuits, System, Communication and Information Technology Applications",IEEE, 2014
8. Rich Adduci, Dave Blue and Guy Chiarello, "Big Data: Big Opportunitiesto create Business value", in EMC2
9. M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, Analytics: the real-world use ofbig data: how innovative enterprises extract value from uncertain data, Executive Report, IBM Institute for Business Value and Said Business School at the University of Oxford, 2012.
10. Parth Chandarana and M Vijayalakshrni, "Big Data Analytics Framework", in International Conference on Circuits, System, Communication and Information Technology Applications",IEEE, 2014
11. Janusz Weilki, "Implementation of Big Data Concept in organizations- possibilities, impediments and challenges", proceeding of2013 Federated conference on computer science and information systems, pp985-989, IEEE, 2013
12. Dr Venkateswara Reddy & MS. Arnina Salim, "A comparative study of various clustering techniques on Big Data Sets using Apache Mahout", in 3,d MEC International Conference on Big Data SmartCity, IEEE, 2016.
13. Prof R A Fadnavis & Sannudhi Tabhane, "Big Data Processing using Hadoop", in IJCSIT, Vol I, 2015
14. Fatos Xhafa, Victor Naranjo and Sanfi Caballe, "A software Chain approach to Big Data Stream Processing and Analytics", in International Conference on Complex Intelligent and Software Intensive systems, IEEE 2015
15. Poonam Vashisht and Vishal Gupta, " Big Data Analytics Techniques: A survey", in IEEE 2015.
16. Kyounghyun Park, Minh Chau Nguyen and Heesun Won, "Web based Collaborative Big Data Analytics on Big Data as a serviceplatform", in ICACT, July 2015
17. Sruthika and Dr. N. Tajunisha, "A study on evolution of Data Analytics to Big Data Analytics & its research scope", in 2nd International Conference on Innovations in Information Embedded and communication system, IEEE, 2015
18. Bichitra Mandai, Ramesh Kumar Sahoo and Srinivas Sethi, "Architecture of efficient word processing using Hadoop for Big Data Applications", in International Conference on Man and Machine Interfacing",IEEE 2015
19. Parth Chandarana and M Vijayalakslnni, "Big Data Analytics Framework", in International Conference on Circuits, System, Communication and Information Technology Applications",IEEE, 2014
20. Dawei Jiang, Gang Chen, Beng Chin ooi and Sai Wu", "epiC: An extensible and scalable system for processing Big Data", Proceeding of VLDB Endowment, Vol7, No.
21. ZoltanPrekopcsak, Garbar Makrai and Tamas Henk, "Radoop : Analyzing Big Data with Rapidminer and Hadoop"
22. Alexander Alexendrov, Rico Bergmann and Stephem even," The stratosphere platform for Big DataAnalytics", Springer, 2014
23. Xindong Wu, Xingquan Zhu, Gong Qing Wu and Wei Ding, "Data Mining with Big Data", in IEEE transactions in knowledge and data engineering, Vol26, Number I, January 2014
24. Barna Saha and Divesh Srivastava, "Data Quality: The other face ofBig Data", in IEEE, 2014
25. Jinsong Zhang, Yan Chen and Taoying Li, "Opportunities of Innovation under challenges of Big Data", in 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2013///
26. Seref Sagiroglu and Duygu Sinang, "Big Data : A Review",IEEE,2013
27. Janusz Weilki, "Implementation of Big Data Concept in organizations- possibilities, impediments and challenges", proceeding of2013 Federated conference on computer science and information systems, pp985-989, IEEE, 2013
28. Du Zhang, " Inconsistencies in Big Data", in Proceeding of IEEE international conference on Cognitive Informatics and Cognitive Computing" IEEE, 2013
29. Katharina Ebner, Thilo Buhnen and Nils Urbach, "Think Big with Big Data: IdentifYing Suitable Big DataStrategies in Corporate Environment", 47th Hawaii International Conference on System Science", IEEE, 2014
30. YouseefMEssa, "Mobile Agent Based New Framework for improving Big Data Analysis", in International Conference on CloudComputingand Big Data, IEEE, 2013
31. Sung Hwan Kim, Nam UK Kim and Tai Myoung Chung, "Attribute Relationship Evaluation Methodology for Big Data Security", in IEEE, 2013
32. Zibin Zheng, Jiennning Zhu and Michael R Lyu, "Service generated big data and big data as a service : An Overview", in IEEE International Congress on Big Data, 2013

33. Avita Katal, Mohammad Wazid and R H Goudar, "Big Data: Issues, Challenges, Tools and good Practices", in IEEE 2013
34. Marcus R. Wigan and Roger Clarke. "Big Data's big unintended consequences", in IEEE Computer Society, 2013
35. Yuri Demchenko, Paolo Grosso andCees de Laat, "Addressing Big Data Issues in Scientific Data Infrastructure", in IEEE 2013
36. Dan Garlasu, "Data Implementation Based on Grid Computing"
37. StephenKaisler, Frank Arrmour, J. Alberto," Big Data: Issues and ChallengesMovingForward",46th Hawaii International Conference on System Science, IEEE,2012
38. Barna Saba and Divesh Srivastava, "Data Quality : The other face ofBig Data", in IEEE, 2014

