International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 9, Issue 5, September - October 2020 ISSN 2278-6856

Predictive Analytics for Big Data Processing

Ketaki S. Tadkalkar

Vishwakarma Institute of Technology,
Pune

Abstract: ***With the increasing use of technology, managing large amount of data is leading problem. These big data sets are difficult to manage, query and process. Performance and efficiency of algorithms may decrease while analyzing and making predictions through big data. Data analysis, querying, searching and prediction challenges are associated with big data. These can be solved by predictive analytics. Predictive analytics is a field that deals with statistics and modeling of current and historical data in either batch or real-time manner to derive patterns from data or to predict future instances. It acts as a decision making tool that has applications in various industries. This paper discusses big data challenges, how predictive modeling solves these problems and different tools and technologies used, various approaches for applying predictive analytics and their advantages and disadvantages. This is a survey paper that covers different approaches for predictive analytics for big data such as use of R analytics tool, agile methodology, probabilistic programming.***

Keywords: Big data, predictive modeling, historical data, RF, LDA, machine learning

1. INTRODUCTION

Big data is a term widely used today. We can say, with vast amount of data and with day-to-day addition to it, analysis of big data is now causing some problems. Big data basically represents the Information characterized by much High Volume, Velocity, and Variety so it requires different Technologies and statistical Methods for its analysis and processing. Predictive modeling is one of the technologies that is used in this task. Predictive modeling is a decision-making tool that predicts values for various assets by making use of historical data. Decision making is a process that involves identifying and selecting the best option from a range of options. Predictive analytics involves a series of steps with statistical analysis, machine learning, predictive modeling, and data mining. It makes use of both past and current data in either real-time or batch format to make predictions. It may have applications in various sectors like business, predicting outcome of legal decisions, clinical decision support systems, predicting life of an equipment etc.

Different approaches and technologies might be involved in predictive analysis which involve regression as well as machine learning techniques. Regression models focus on establishing a relation between variables may be a mathematical equation that might be in linear or non-linear form.

2. LITERATURE SURVEY

In [1], Big data analytics using Agile Model by Surendra Raj Dharmapal and Dr. K. Thirunadana Sikamani, they have discussed an agile approach for solving big data problem. They have defined steps involved in big data analytics which includes approach from exactly what you are trying to achieve with big data processing to Collection of data set from right source, which means all the hardware used or parameters involved or connectivity issues are resolved before
performing
operations on data. Then next step is Filtering unwanted data, which is basically the most important step in all kinds of analyses either big data or with small datasets. Next is frequently ensuring what you are doing and what results you are getting are correct and taking timely updates of it. After that an Algorithm needs to be generated and User Predictive Analytics Tools that you are working with are checked for their correctness and accuracy.

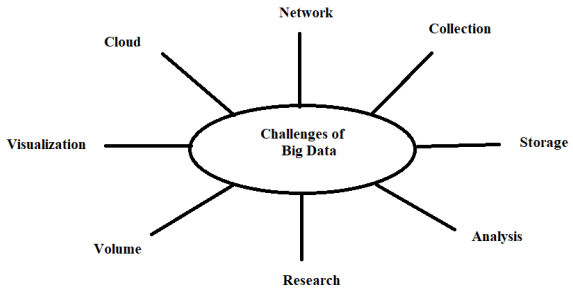
In [2], Probabilistic Programming and Big Data by Arinze Anikwue and Boniface Kabaso, they have reviewed an existing framework that is based on probabilistic programming of big data. They say probabilistic programming is based on a representation language. They selected a primary study based on a probabilistic programming framework based on Apache Spark called InferSpark which was presented by [3]. According to [3], this framework is useful in the implementation of statistical inference on big data using the distributed processing power of Apache Spark.

In [4], New Machine learning based approach for predictive modeling on spatial data by M. Gangappa, C. Kiran Mai, P. Sammulal, they have presented an approach for classification based on predictive modeling technique for spatial data sets. They have employed Advanced machine learning concepts like RST in algorithmic manner, which gives a major advantage of dimensionality reduction of data in efficient way. They have performed experiments on geospatial repository dataset and provided simulation results.

In [5], A novel approach for Big Data processing using message passing interface based on memory mapping, an approach for big data processing and management is discussed that differed from the existing ones. This method employs the memory space to read and handle data, along with the memory-mapped space extended from memory storage. Methodology of this paper is that it segments big data using memory mapping and broadcasts all segments to several different processors using parallel message passing interface. Also, the paper presents approach based on a homogenous network which works parallelly for encrypting and decrypting big data using AES algorithm. This proposed approach can be done on Windows Operating System with the use of .NET libraries.

3. PROBLEM ANALYSIS

The management of large volumes of structured and unstructured data in efficient manner is difficult using conventional relational database management systems (RDBMS). Also, each non-relational database or NoSQL tool has its own disadvantages. Hence, businesses find it challenging to select right non-relational database and best data management tool. Scaling up and down Big Data according to Current Demand is another challenge of big data. Unlike conventional software development projects, big data projects grow more and more large and complex consistently. Businesses need to keep track of infrastructures and resources for processing, storage and querying big data. Another challenge with big data is Overcoming Big Data Talent and Resource Constraints. There is a need of skilled professionals to manage and analyse huge volumes of real-time data that is being collected from various sources and in various formats. Also, big data tools enable businesses to collect real-time data from external and internal sources. collected data varies in formats and quantity.so there is a need to set up scalable data warehouses to store the incoming data in a reliable and secure way. simultaneously, there is need to invest in robust big data solutions for integrating the structured and unstructured data by eliminating disparities and inconsistencies. Maintaining Data Integrity, Security, and Privacy is another challenge. The number of studies suggest that many enterprises do not utilize big data effectively, thereby not addressing these challenges early. It is always important for the businesses to implement a well-designed strategy of managing and leveraging big data by addressing these common big data challenges. Also, big data management is a continuous process. These challenges of big data can be solved using predictive analytics by enabling reduced risks, making intelligent decisions, and creating differentiated customer experiences.



4. METHODOLOGY

In paper [6] I.e. Technologies of predictive analytics for Big Data, they have discussed standardization levels for creating analytic systems for big data. They have considered irregular structured data for processing. According to them, it is difficult to parallelly process data as they have poor space and time localization and low frequency of repeated bonding to nearby data. Standardization levels include unified interface for access by any application, common agreement for creation and exchange of data mining models, common methods for organizing analytical data. In first standardization level, SQL language is used as add-on, and it allows use of in database analytic instruments. In second, XML is base decision language while in third it is covered by CRISP-DM.

Paper [6] tells that with SQL/MM standard all data processing needs to be carried out in the place where data is stored so that computational cost for data transfer can be avoided. It is extension of SQL language and it allows formation of common SQL scripting including queries for table controls and for data mining algorithms [6].

In [7], they have performed experiments on healthcare data by R analytics tool. They describe R tool as open source software and statistical tool used for data and visual analytics. It uses predictive analytics for better decision making and for finding solutions. It has high availability of big data handling, manipulation and storage facility. So, it works best with big data.

In [7], they have implemented random forest algorithm which works best with huge datasets. Bagging is used for feature selection for assembling decision tree with variance. Another algorithm implemented by [7] is Latent Dirichlet Allocation (LDA). It is used for statistical model that allows number of observations to be expressed by unobserved groups that explain causes of similarity of data. LDA is an algorithm that helps in analysis of different kind of data that is in the form of images, audio, text files etc. It aims at finding short imagery for variables. LDA helps to find out facts by exploring data in step by step manner. LDA suggests that one way to concise the content of document quickly is by looking at the set of styles it generally uses [8].

In [9], Predictive over indebtness on batch and streaming data, they have discussed framework for inter-bank application. They have presented four common tasks over three stages. These involve preprocessing which involves feature selection and data balancing. Input to this stage is high dimensional and unbalanced data. After going through this stage, data becomes low dimensional and balanced. Second stage is training where preprocessed data is trained using supervised algorithms. Last stage is prediction. In this stage, highly unbalanced test set is passed to supervised algorithm to generate predictions.

5. RESULTS

In [7], random forest and latent Dirichlet allocation algorithms are implemented. Results found by those are random forest gives 98.75% accuracy whereas LDA gives 85.01% accuracy. It means predictions using random forest are more accurate than those with LDA so we can say, RF works very well for huge datasets. They have expressed those results in various forms like specificity, which is greater for RF than LDA, Sensitivity for LDA is smaller than RF and other parameters such as positive prediction rate, negative prediction rate etc.

In [9], For Feature Selection, they have tested 2 embedded methods and 4 filter methods. Embedded methods include rfe which uses an external estimator for recursively selecting subset of features. Another is ig that uses feature ranking from tree-based algorithms to select relevant features. They have used different tree-based classifiers

such as DT, ET, RF, GBT, and XGB. For Data Balancing, they have tested 6 under-sampling techniques and 5 over-sampling techniques.
XGB over-performs RF for batch settings. It reduces the sensitivity-specificity gap. CE and BP data have shown improvements in both specificity and sensitivity compared with the baseline [9].

6.CONCLUSION

There are different kinds of tools and techniques available for analyzing big data. Out of which predictive analysis is more suitable for working with high volumes of continuous data. Predictive analysis can be performed using various methods. This paper discusses different approaches for performing predictive modeling such as using R analytical tool, using agile method, probabilistic programming, by deriving statistical inference, by machine learning based approach, using message passing interface and memory mapping techniques. Each has its advantages and disadvantages. Different methodology is followed by them. This paper reviews all these approaches.

References

[1] Surendra Raj Dharmapal and Dr. K. Thirunadana Sikamani, “Big data analytics using Agile Model”, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016
[2] Arinze Anikwue and Boniface Kabaso, “Probabilistic Programming and Big Data”, 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems
[3] Z. Zhao, J. Pei, E. Lo, K. Q. Zhu, and C. Liu, "InferSpark: Statistical Inference at Scale," 2017.
[4] M. Gangappa, C. Kiran Mai. P. Sammulal, “New Machine Learning based approach for predictive modeling on spatial data”, 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)
[5] Saad Ahmed Dheyab, Mohammed Najm Abdullah & Buthainah Abed, “A novel approach for Big Data processing using message passing interface based on memory mapping”, Journal of Big Data, 2019, Article no.112
[6] A. Yu. Dorogov ”Technologies of predictive modeling for big data” 2015 XVIII International Conference on Soft Computing and Measurements (SCM)
[7] Priyanka P. Shinde, Kavita S. Oza, R. K. Kamat “Big data Predictive Analysis: using R Analytical Tool”, International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017)
[8] K. VijiyaKumar, V. Govindasamy and T. Esther, "An online big data take oution using latent Dirichlet allocation," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, Tamilnadu, India, 2016, pp. 2278-2283. doi: 10.1109/ICCSP.2016.7754101
[9] Jacob Montiel, Albert Bifet, Talel Abdessalem, “Predictive over indebtness on batch and streaming data”, 2017 IEEE International Conference on Big Data (BIGDATA)