ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 11, November 2018

Application of Machine Learning with Big Data

Umang

Assistant Professor, Department of Information Technology, KU, SSJ Campus Almora, Uttarakhand, India

ABSTRACT: Advancements in Big Data harvesting technologies make it realistic to store and interpret large amounts of data generated from diverse sources. These sources produce both unstructured and semi-structured real time data. Organizations are getting benefitted from deriving the business intelligence from these data stores. Predictive analytics, a branch of analytical models help in deriving foresights based on the current variable inputs, has brought lot of interest among researchers. Predictive analytics and machine learning usually work in tandem as predictive analytical method usually include a machine learning algorithm. Using machine learning and artificial intelligence algorithms, organizations can uncover complex statistical patterns and optimize their business processes. But the traditional machine learning algorithms were developed in out of context with Big Data with few assumptions. This study explores the opportunities and challenges of predictive analytics with Big Data and also reviews the issues which machine learning algorithms have to deal with Big Data for predictive analysis.

KEYWORDS: Big Data Analytics, Big Data Predictive Analysis, Machine Learning

I. INTRODUCTION

Big Data deals with massive data sets derived from various data sources constituting both structured and unstructured data. The data is being generated at rates faster than most enterprise databases can handle. IDC [1] estimates that over 50 billion IoT sensors will be in operation by 2020, and more than 200 billion devices will be networked by 2030. The growth of the devices with Internet of Things (IoT) would provide substantial potential benefits for businesses in general and society in particular. The ability to extract value from these Big Data stores for knowledge discovery and better decision-making comes from Big Data analytics [2] [3]. Big Data analytics has become an emerging area for data researchers and is the core of Big Data [4].

Predictive analytics [5] involves using sophisticated technologies which help organizations to use both the data stored in the data repositories and the real-time data to derive business intelligence of what lies ahead for the organization. This brand of analytics involves simulations with large processing computers with advanced database technology. In addition, predictive analytics uses several mathematical techniques that probe data, derive valuable patterns, and make accurate predictions. These predictive models make it possible to support business decisions with more effectiveness and involve less cost.

Most business processes would benefit from these predictive models. Still, they can be beneficial when much digital data is available, and the business process involves many similar decisions. They can also be interested in the process where the business outcomes significantly impact the profits or efficiency. As a result, they are used in vast business areas such as healthcare, retail, insurance, financial and government services. Machine learning focuses on developing fast and efficient algorithms and models that enable real-time data processing. These algorithms provide improved accuracy and performance compared to conservative algorithms. These algorithms also get smarter with use and experience [6]. Machine learning is best suitable for exploiting the prospects of Big Data. It provides value from large and distinct data sources with less dependence on the human direction. It is well suited for dealing with complexity and various data sources involving several variables. Unlike traditional algorithms, machine learning systems' accuracy and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

efficiency improve with an increase in the size of data sets. The more data is fed into the machine learning system, the more it can learn from it, and the quality of the results improves. The rest of the study is organized as follows. Section 2 introduces the methods, challenges and opportunities of predictive analysis with Big Data. Section 3 describes the issues which machine algorithms have to deal while. They are performing predictive analytics with Big Data. Finally, section 4 presents the challenges and scope of future research opportunities in this area.

II. PREDICTIVE ANALYTICS WITH BIG DATA

Analytical solutions have been a boon to many organizations in improving decision-making. Companies are trying to analyze historical and real-time data to forecast future elements. Evans et al. [7] define advanced analytical solutions to be broadly categorized into three models

- *Descriptive analytics is used when data* is driven by analyzing in-store data to understand past and current decisions.
- *Predictive analytics* is a more advanced type of analytics that looks at past data and predicts future events.
- *Perspective analytics* is an advanced of the three, which involves optimizing the results of predictions. It is deciding on the actions to be performed. It is a combination of data, mathematical models and business rules.

Predictive analytics [5] constitute a set of statistical methods involving machine learning, artificial intelligence, regression techniques and data mining that predicts future events and behaviours. It is a systematic process where an algorithm finds out the patterns and relationships among the variables. Predictive analytics and predictive models, along with data mining techniques, involve using multi-variable analysis methods such as time series and advanced regression models. They help to discover meaningful patterns that enable organizations to make intelligent, effective and fast decisions.

Today business organizations are collecting vast amounts of data such as customers, markets, social networking, cloud, real-time and performance data. So Big Data comes into the picture for storage and analysis of these vast and variant data sets. Predictive analytics helps to gain insights from these massive amounts of data to optimize business processes. There are numerous case studies of the use of Big Data predictive analytics for practical use as A.R. Reddy et al. [8] present the role of Big Data analytics in general and predictive analytics in particular in healthcare to face healthcare issues. Krumeich et al. [9] detail the opportunities of predictive analytics in mining Big Data for event-based predictions, which provide proactive control over business processes. Gulwani et al. [10] review various algorithms and methods to be used in E-Learning systems for predicting student performance with high accuracy and ease of interpretation. Egebjerg et al. [11] present a model to predict the number of spectators in a football match based on spectators' online and offline behavior. These extracted results can be helpful for companies to understand the customer base and improve their marketing strategies.
Predictive analytics with Big Data involves systematic procedure and involves the following steps:

1. Data generated from various sources has to be gathered, and only data relevant to the business goals is to be identified and stored.
2. The stored data is to be prepared by performing various data-cleaning processes to derive appropriate data for analysis.
3. A predictive model must be designed using statistical or machine learning algorithms depending on the type of data available and the level of prediction needed as a part of the system.
4. Evaluate the prediction model for effectiveness and accuracy with a data sample.
5. Apply the model in applications and derive the foresight of business.
6. Monitor and adjust various algorithm parameters and improve the model's efficiency and accuracy.

III. CHALLENGES

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Predictive analytics provide opportunities to organizations with few challenges. Some of these challenges are as follows:

- ***Data challenges*:** The various challenges the companies have to deal with when concerning Big Data are significant [12]. Among them, coping with the numerous data formats, maintaining the data quality, and deriving value from data sources are the main challenges which predictive analytics have to be concerned about.
- Choosing the business intelligence tools and pre-processing tools to identify particular relevant information.
- Determining the variables necessary for the prediction process. It also involves how traditional data elements are related to one another and which factors have better influences on the business outcomes.
- Choosing a proper predictive model to mine the data and discover patterns and iterate through this process and improve the efficiency and accuracy of the results.
- Simplifying the analytical process and automating the important and necessary actions of the process.
- Minimizing the data movements while performing the analysis to conserve the computing resources.
- Enabling decision-making based on predictive modelling and business rules.

Providing techniques and solutions to these challenges and easing the process will be crucial to the organization's growth and profits.

IV. MACHINE LEARNING WITH BIG DATA

Machine language is a fundamental component in data analytics. It is considered a critical driver in the Big Data evolution. The reason for this is the ability of machine learning algorithms to learn from the data and provide data insights and predictions [13]. Predictive analytics uses machine learning algorithms to develop predictive models, which give foresight [14]. Several algorithms involving neural and naïve Bayesian networks have been proposed.

A common assumption of machine learning algorithms is that the algorithms get better as data values increase, providing accurate results [15]. Still, Big Data imposes various challenges to these machine learning algorithms due to massive data sets. The challenges that machine learning algorithms face due to Big Data are as follows.

- ***Processing:*** One of the main challenges of Big Data computations is that computational complexity increases as the data size grows. It also affects the time and memory needed to train the algorithm. In some situations, as the data size grows, the performance of algorithms depends on the architecture used to store and move data.
- ***Storage:*** Most machine learning algorithms assume that the data being processed is in a single file on a disk [16]. Due to the size of data sets, the data items not only do not fit in the memory but are also distributed over many files in different physical locations. But as the size of the data grows, machine learning algorithms which depend on this assumption tend to fail. One of the methods to provide a solution for this problem is Map Reduce. Grolinger et al. [15] have discussed the challenges of Map Reduce in Big Data.

Machine learning algorithms would require data to reside at a single location for processing. As a result, it would require the transfer of data elements from different backgrounds. This transfer of data elements would cause processing delay and consume network bandwidth. Due to this, storage and data locality are challenges to address in any Big Data system.

- Data Samples

With the size growth, data is not uniformly distributed [17]. This uneven distribution of data items severely affects the performance of a machine-learning algorithm. Japkowicz et al. [18] show that traditional machine learning mechanisms, such as decision trees and neural networks, are susceptible to these uneven data distributions in data samples.

- Data attributes

A machine learning algorithm's effectiveness and predictive ability reduce with an increase in the number of attributes for a data item [19]. As the volume of Big Data increases, there is a potential for an increase in the number of points.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Another attribute issue is feature selection, which helps select the relevant features and allows the machine learning algorithm to perform better. Also, identifying the relationship between the data items is a massive challenge due to the size of the data sets.

- Data Quality

A data sample usually needs more quality data. This data might contain outliers, missing values, errors and false positives, which might not have any meaning within the data. However, these data elements seriously affect the performance and results of the algorithms. Therefore providing a means to exclude these outliers and false positives is crucial in machine learning with Big Data.

- Data Heterogeneity

Big Data analytics involves integrating data from various diverse sources. They can vary in the form of type, formats and implementations. Moreover, they have different designs and meanings in other contexts, and interpreting these data elements is another machine-learning challenge.

- Real-Time Analytics

Many machine learning algorithms assume that learning starts once the data is available in the store. But with the concept of real-time streaming data, such an assumption is not valid. Real-time analytics adds data to the existing stores. Therefore machine learning algorithms must support learning with these incremental data sets [20]. With these data sets, real-time processing is also a significant challenge to get instant business insights.

V. CONCLUSION

With Big Data, analytics is fast moving from the conventional business intelligence methods that utilize basic information to a more predictive and prescriptive approach to discover patterns and interrelationships for better organizational decision-making. This work presents the various issues ad challenges of predictive analytics when dealing with Big Data and the implementation and application of machine algorithms for analysis. So adopting new machine learning techniques to solve the existing difficulties and combining existing solutions to provide performance improvements is needed for developing machine learning with Big Data.

REFERENCES

[1]. *IDC future scope: worldwide Internet of Things. 2015* https://www.idc.com/research/forecasts.jsp
[2]. V Mayer Schönberger and KCukier, *Big Data: A Revolution that Will Transform how We Live, Work and Think.* Houghton Mifflin Harcourt, 2013.
[3]. Dubey, Gunasekaran , Childe, Wamba, & Papadopoulos. "*The impact of big data on world-class sustainable manufacturing"*. The International Journal of Advanced Manufacturing Technology, 1-15.2015.
[4]. H V Jagadish, J Gehrke, A Labrinidis, Y Papakonstantinou, J M Patel, R Ramakrishnan, and C Shahabi, "*Big Data and its Technical Challenges*," Communications of the ACM, vol. 57, no. 7, pp. 86–94, 2014.
[5]. Abbott, D. *"Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst"*.John Wiley & Sons.2014.
[6]. M James, C Michael, B Brad, and B Jacques, "*Big Data: The Next Frontier for Innovation, Competition, and Productivity*".The McKinsey Global Institute, 2011.
[7]. Evans JR. *"Business Analytics – methods, models, and decisions".* Pearson. 2013;
[8]. A R Reddy and P S Kumar, *"Predictive Big Data Analytics in Healthcare"* , Second International Conference on Computational Intelligence& Communication Technology (CICT), Ghaziabad, 2016, pp. 623-626.
[9]. J Krumeich, B Weis, D Werth, and P Loos, *" Event-driven business process management: Where are we now?- A comprehensive synthesis and analysis of literature,"* Business Process Management Journal, vol. 20, no. 4, 2014.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

[10]. M S Vyas and RGulwani, *"Predictive analytics for E learning system,"* International Conference on Inventive Systems and Control (ICISC),Coimbatore, 2017.
[11]. N H Egebjerg, N Hedegaard, G Kuum, R RMukkamala and RVatrapu, *"Big Social Data Analytics in Football: Predicting Spectators and TV Ratings from Facebook Data,"* 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, 2017.
[12]. LalithaBalla, Chavva Ravi Kishore Reddy, A V L N Sujith. *"BigData Analytical Challenges with IOT".* International Journal of Distributedand Cloud Computing ", Volume 5 Issue 1, June 2017.
[13]. M Rouse, "*Machine Learning Definition*" 2011. http://whatis.techtarget.com/definition/machine-learning.
[14]. M Rouse, "*Predictive Analytics Definition*" 2009. http://searchcrm.techtarget.com/definition/predictive-analytics.
[15]. K Grolinger, M Hayes, W AHigashino, AL'Heureux, D S Allison, and M A MCapretz, "*Challenges for MapReduce in Big Data*" in Proceedings of the 2014 IEEE World Congress on Services (SERVICES), 2014.
[16]. K A Kumar, J Gluck, A Deshpande, and J Lin, "*Hone: 'Scaling Down' Hadoop on Shared-Memory Systems*" Proceedings of the VLDB Endowment, vol. 6, no. 12, pp. 1354–1357, 2013.
[17]. M Ghanavati, R K Wong, F Chen, Y Wang, and C S Perng, "*An Effective Integrated Method for Learning Big Imbalanced Data*" inProceedings of the 2014 IEEE International Congress on Big Data, 2014.
[18]. N Japkowicz and S Stephen, "*The Class Imbalance Problem: a Systematic Study*" Intelligent Data Analysis, vol. 6, no. 5, pp. 429–449, 2002. [19]. G Hughes, "*On the Mean Accuracy of Statistical Pattern Recognizers*" IEEE Transactions on Information Theory, vol. 14, no. 1, pp. 55–63,
1968.
[20]. X Geng and K Smith-Miles, "*Incremental Learning*," in Encyclopedia of Biometrics SE 304, Springer US, 2009,pp.731–732