Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 07, No. 1, July : 2022

Predictive Analytics using Machine Learning with BigData

Y.HARINATH
Assistant Professor
harinath5404@gmail.com

SHAIK MAHABOOB SUBHANI

Assistant professor
smsubhanialts@gmail.com

J. RAVI KISHORE
Assistant Professor
ravialts2022@gmail.com

Abstract: The development of BigData technology has made it feasible to store and analyze vast volumes of data supplied by several sources. These sources generate both unstructured and semi-structured data in real time. Deriving business insight from these data repositories is advantageous for organizations. Researchers have shown a great deal of interest in predictive analytics, a branch of analytical models that aid in obtaining foresights based on the present variable inputs. Often, predictive analytics and machine learning operate together, since predictive analytical methods typically use a machine learning algorithm. Using machine learning and artificial intelligence algorithms, businesses may discover and improve intricate statistical trends and business processes. However, typical machine learning algorithms were designed with minimal assumptions and out of context with BigData. This study investigates the advantages and disadvantages of predictive analytics using BigData. This paper also examines the challenges machine learning algorithms have when analyzing Big Data for prediction purposes. Finally, we discuss the future research directions in predictive analytics employing machine learning and BigData.

Keywords: BigData, Predictive Analysis, Machine Learning, BigData Analytics.

1. Introduction

BigData is about dealing with huge data sets derived from wide variety of data sources constituting both data that is structured and unstructured. 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 networked by 2030. The growth of Internet of Things (IOT) would provide huge potential benefits for business in general and society in particular. The ability of extracting value from these BigData stores for knowledge discovery and better decision making comes from

BigData analytics [2] [3]. BigData analytics has become an emerging area for data researchers and is the core of BigData[4] .

Predictive analytics [5] involve using sophisticated technologies which helps the organizations to use both the data stored in the data repositories and also 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. Predictive analytics uses a number of mathematical techniques that probe data and derive useful patterns and make accurate predictions. These predictive models make it possible to support the business decisions with more effectiveness and involving less cost.

Most business processes would benefit from these predictive models, but they can be beneficial in situations where there is abundance of digital data available and where the business process involves large number of similar decisions. They can also be involved in process where the business outcomes have huge impact on the profits or efficiency. As a result they are used in wide areas of business such as healthcare, retail, insurance, financial and government services.

Machine learning focuses on developing fast and efficient algorithms and models that enables real time processing of data. These algorithms provide improved accuracy and performance compared to the conservative algorithms. These algorithms also get smarter with use and experience [6]. Machine learning is best suitable for exploiting the prospects BigData. It provides value from large and distinct data sources with less dependence on human direction. It is well suited for dealing with complexity and variety of data sources, where number of variables are involved. Unlike traditional algorithms, machine learning systems accuracy and efficiency improves with increase in size of data sets. As more data is fed into the machine learning system the more it can learn from the data and quality of the results improves.

The rest of the paper is organized as follows. Section 2 introduces the methods, challenges and opportunities of predictive analysis with BigData. Section 3 describes the issues which machine algorithms have to deal while performing predictive analytics with BigData. Section 4 presents the challenges and scope of future research opportunities in this area.

2. Predictive analytics with BigData

Analytical solutions have been a boon to a large number of organizations for improving the decision making process. Companies are trying to analyze the historical and real time data to forecast the future elements. Evans et al. [7] defines advanced analytical solutions to be broadly categorized into three models

- *Descriptive analytics*, which are driven by analyzing in-store data to understand the past and current decisions.
- *Predictive analytics* is somewhat a more advanced type of analytics which looks at the past data and predicts the future events.
- *Perspective analytics* is 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 behaviors. 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 use of 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 huge amounts of data such as customer, markets, social networking, cloud, real time and performance data. So BigData comes into picture for storage and analysis of these huge 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 use of BigData predictive analytics for effective use:

1. A.R.Reddy et al. [8] presents role of BigData analytics in general and predictive analytics in particular in the field of healthcare to face the issues in healthcare.
2. Krumeich et al. [9] details the opportunities of predictive analytics on mining BigData for event-based predictions which provide proactive control over business processes.
3. Gulwani et al. [10] reviews various algorithms and methods which are to be used in E-Learning systems for predicting student’s performance with high accuracy and ease of interpretation.
4. Egebjerg et al. [11] presents a model to predict the number of spectators in a football match based on online and offline behavior of spectators. These extracted results can be helpful for companies to understand the customer base and improve their marketing strategies.

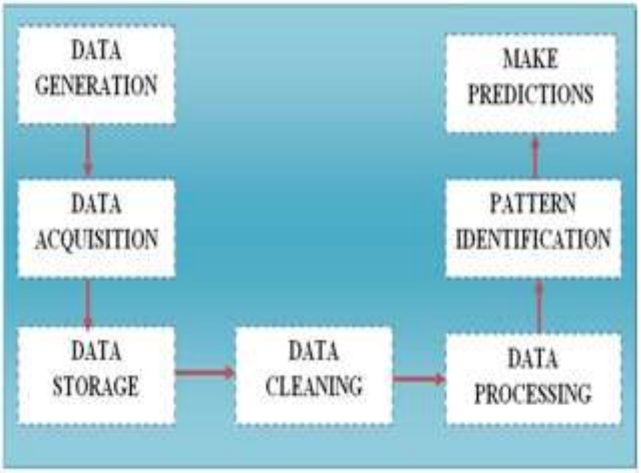

Fig 1: Predictive Analysis Procedure

Predictive analytics with BigData involves systematic procedure as depicted in Fig.1. Predictive analysis 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 to perform analysis.
3. A predictive model has to be designed using statistical model 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 foresights of business.
6. Monitor and adjust various parameters of the algorithm and improve the models efficiency and accuracy.

2.1. Challenges

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

- ***Data challenges*****:** The various challenges which the companies have to deal with, when concerning BigData is large [12]. Among them dealing with the numerous data formats, maintain the data quality and deriving value form 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 special relevant information.
- Determining the variables important for the prediction process. It also involves how traditional data elements are related to one another and which elements have better influences in 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 modeling and business rules.

Providing techniques and solutions to these challenges and easing the process will crucial to the organizations growth and profits.

3. Machine learning with BigData

Machine language is a fundamental component in data analytics. It is considered as key drivers in the BigData evolution. The reason for this is the ability of the 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 provider foresights [14]. A Number of algorithms involving neural networks and naïve Bayesian networks have proposed for this purpose.

A common assumption of machine learning algorithms is that the algorithms get better as data values increase, which in turn provide accurate results [15], but BigData imposes a variety of challenges to these machine learning algorithms due to massive data sets. The challenges which the machine learning algorithms have to face due to BigData are as follows

- ***Processing:***One of the main challenges of BigData computations is that as the data size grows, computational complexity increases. It also effects the time and memory needed to train the algorithm. In some situations, as the data size grows the performance of algorithms depend on the architecture used to store and move data.
- ***Storage:***Most of the machine learning algorithms assumes that the data being processed is present in a single file on a disk [16]. Due to the size of data sets, the data items not only do they not fit in the memory but also items are distributed over a large number of files, residing 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 challenges of Map Reduce in BigData.

Machine learning algorithms would require data to be residing at a single location for processing. As a result it would require transfer of data elements from different locations. This transfer of data elements would cause processing delay and consume network bandwidth. Due to this storage and data locality is a challenge to be addressed in any BigData system.

- **Data Samples**

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

- **Data attributes**

The effectiveness and predictive ability of a machine learning algorithm reduces with increase in the number of attributes for a data item [19]. As the volume of BigData increases, there is a potential in increase in the number of attributes. Another issue with attributes is the feature selection, which helps to select the relevant features, which will help the machine learning algorithm to perform better. Also identifying the relationship between the data items is huge challenge due to the size of data sets.

- **Data Quality**

A data sample may usually include quality less data. This data might contain outliers, missing values, errors and false positives, which might not have any meaning within the data. 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 the context of machine learning with BigData.

- **Data Heterogeneity**

BigData analytics involves integrating data from various diverse sources. They can vary in the form of type, formats and implementations. Not only they have different formats but they have different meanings in different contexts and interpreting these data elements is another challenge of machine learning.

- **Real-Time Analytics**

Many machine learning algorithms assume that learning starts once the entire 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 an important challenge to get instant business insights.

4. Conclusion& Future work

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

5. 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.

[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 Distributed and 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*" in Proceedings 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. 731–735.