Journal of
Data Mining in Genomics & Proteomics

ISSN: 2153-0602

AL-Masri and Nasir, Journal of Data Mining in Genomics & Proteomics 2016, 7:2
DOI: 10.4172/2153-0602.1000190

Short Communication

Open Access

Learning Machine Implementation for Big Data Analytics, Challenges and Solutions

Ahmed N AL-Masri* **and Manal M Nasir**

American University in the Emirates, DIAC, P.O. Box: 31624, United Arab Emirates

***Corresponding author:** Ahmed N AL-Masri, American University in the Emirates, DIAC, P.O. Box: 31624, United Arab Emirates, Tel: +971-4-449-9000; E-mail: ahmedn82@hotmail.com

Received date: December 28, 2015; **Accepted date:** February 16, 2016; **Published date:** February 22, 2016

Abstract

Big Data analytics is one of the great challenges for Learning Machine (LM) algorithms because most real-life applications involve a massive information or big data knowledge base. By contrast, an Artificial Intelligent (AI) system with a data knowledge base should be able to compute the result in an accurate and fast manner. This study focused on the challenges and solutions of using with Big Data. Data processing is a mandatory step to transform unstructured Big Data into a meaningful and optimized data set in any LM module. However, an optimized data set must be deployed to support a distributed processing and real-time application. This work also reviewed the technologies currently used in Big Data analysis and LM computation and emphasized that the viability of using different solutions for certain applications could increase LM performance. The new development, especially in cloud computing and data transaction speed, offers significant advantages to the practical use of AI applications.

Keywords: Big data; Learning machine; Hadoop ecosystem; Map reduce; Cloud computing

Introduction

The market demand for LM and AI applications, especially in robotics, image processing, object detection and recognition, classification, and so on, has shown significant growth in the past years. The performance of LM depends on the applied algorithm and data set preparation, which involve time-consuming processes. In fact, no particular system can guarantee superior performance without module adjustment. Big Data solutions offer a new scientific innovation that can be integrated with LM and AI systems to promise high performance in a short time.

Any organization or institution can improve their performance by analyzing the current situation (based on the available information) and taking the right decision (based on the predicted result). Developing such analytic tools is expensive because LM requires a large computing infrastructure, especially when dealing with a big database. The cloud platform limits the difficulties of processing and memory size in analyzing a big data set within a sensible timeframe. For example, Google developed a machine-learning prediction Application Programming Interface (API) by providing numerous libraries or scripts to access the API using different programming languages. The prediction API can be used to solve a complete pattern, such as spam detection and movie-ranking prediction. Data are simulated across several data centers using the cloud storage, where most prediction outputs take less than 0.2 s [1].

The Google API cloud platform has two access methods: (1) the API explorer, which is an interactive tool that is simply accessed from a web browser, and (2) the Google plugin for eclipse. Google also developed an app script for third-party services and the Google prediction client library for R-language. The Amazon machine-learning service is another example supported by the Amazon Web Services and is designed to enable organizations to deploy their Big Data analytics solution on a cloud platform [2]. Microsoft [3] and many other cloud service providers have started offering an LM service because of the high market demand in prediction application.

Assuncao et al. [4] reviewed the development methodologies and environments for performing Big Data analytics on a cloud platform. They categorized the Big Data analytic solutions of an organization into descriptive (i.e., model of past customer activities), predictive (i.e., predicts customer needs based on the available data), and prescriptive (i.e., assist companies with the decision-making process). The challenges and state-of-the-art methods that are currently involved Big Data analysis were overviewed by Chen and Zahang [5]. Jin et al. [6] addressed the importance and opportunities of the concept of Big Data. They also presented the challenges encountered in terms of data, system, and computational complexity, as well as possible solutions to these challenges.

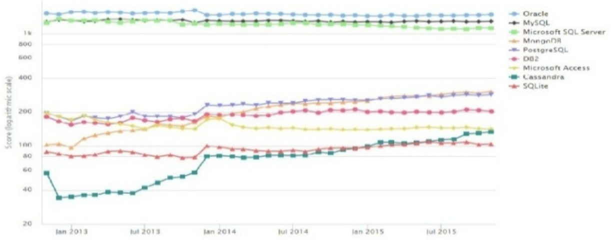

Figure 1: Most popular DB management systems [8].

Numerous big data analytic platforms, such as SQL and Cloudera Impala, have a simple, reliable, efficient, and scalable processing mechanism, which is integrated across many systems. On the contrary, NoSQL supports a mechanism for the storage and retrieval of unstructured non-relational DBs. NoSQL data modeling and description can be found in [7]. The Oracle relational DB management system is the most widely used DB engine as of November 2015 [8]. Figure 1 shows other DB engine ranking for the last two years.

The present study focuses on the implementation of different LM approaches to big data scale and proposes an optimal solution to bridge the gap between the data preparation process and performance of the intelligent system.

LM System

LM is used to solve most non-linear problems in many fields because of its capability to extract important information from training samples. Most of the methods used in worldwide research are based on statistical techniques for analyzing organization data sets. The LM designed by scientific developers has recently offered a high-performance and scalable learning system for data-driven discovery. Supervised learning is implemented when sample patterns are trained with their expected output. This type of training should have integer inputs/outputs; thus, the trained system is able to predict the unknown output based on the given inputs. However, the first challenge in implementing an LM system is the data pre-processing, such as normalization, transformation, handling of missing data, feature extraction and selection, training, optimization, and generalization model. Figure 2 illustrates the implementation of an LM system to predict organization behavior. The second challenge is summarized by the data size increment and Big Data term, where the LM system deals with a high number of inputs/outputs. Data clustering provides advantages to the LM implementation of big data analytics, but designing and developing such system require considerable effort.

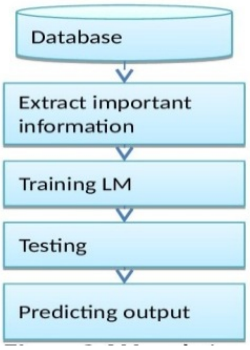

Figure 2: LM analytic system.

The design of an LM algorithm depends on analytic purpose of the organization (customer requirements), that is, whether it is descriptive, predictive, or prescriptive. Numerous computing developers exist, such as the Hadoop ecosystem [9], which is an open-source implementation that is applied as management, processing, and storage analytic solutions. MapReduce is also an associated scheme that solves large data processing with parallel and distributed algorithms [10]. The LM and parallel training approaches provide potential advantages to big data implementation [11].

Best Practices of LM Modules

Information technology applications are developing and growing exponentially, emphasizing the need for engineers and researchers to address the increasing market demands. Any business or government transforms the existing DBs to a knowledge base, which involves the data-processing stage as part of their enterprise system. In reference to a recent review [12], LM is implemented worldwide in diverse fields, such as climate and environment, bio, medicine, and health, galaxies and universe, and economics and finance [13]. For a sustainable LM module, the said contexts should be considered to ensure an efficient and optimal LM computational solution that can be integrated with any knowledge base. Challenges should also be considered before the application of LM in any business environment (Figure 3).

Data measurement

Statistic analytics is based on the data type, that is, whether it is nominal, ordinal, interval, or ratio. An LM model can solve non-linear problems based on its level of measurement, but it should be designed, classified, and identified. The same data measurement must be implemented for each individual LM model to evaluate the performance of the predicted result [14]. Data classification, preparation, and processing significantly affect the simulation results, which also reflect the final evaluation of the entire LM model. LM is also implemented for big data experiments [7,15,16]. In some studies, LM is combined with the feature selection and data processing of a data knowledge base [17-19].

Security, privacy, and integrity

The terms security and privacy have different meanings in computer science. Analyses of data security are applied to security and surveillance practices to encourage experts to use big data security assemblages [20]. Nevertheless, the security ambition in big data applications focuses on privacy, integrity, authenticity, and access control [21]. In cloud-computing implementation, sharing data resources using different data formats by multiple-user access presents new challenges to big data security. The Cloud Security Alliance has covered such difficulties and provided solutions based on four categories: infrastructure security, data privacy, data management, and integrity and reactive security [22]. Another factor is data availability, which is important in LM algorithms, especially when data are stored in various locations and distributed processing is deployed. Other security issues are discussed in Reference [23]. The Institute of Electrical and Electronics Engineers (IEEE) Standards Association has initiated numerous standards and protocols [24] allied with big data management (structured and unstructured, formats, size of data), security, integrity, analytics, and so on.

Scalability

The information in a practical DB is unstructured by nature, even when it comes to LM implementation and regardless of the data platform. Modifying the number of inputs/outputs for an intelligence system without retraining data patterns remains a difficult task. Scalability challenge is an important aspect for any business because it may diversify their intention any time. The solution is to provide online and offline training services, which allow deploying any changes and test them using offline services before switching to online or real-time demonstration. A scalable distributed system is implemented for big data analytics [25] and results in a high performance in the operational time.

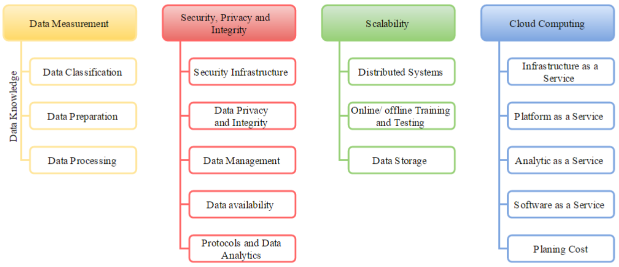

Figure 3: Classification of LM challenges.

Cloud computing

The most striking attribute of cloud computing in LM practice is its elasticity and scalability. Cloud computing service providers offer full or partial infrastructure, Platform, Analytics and/or Software as a Service to clients based on their requirements. Customers may take benefits of the cloud by demonstrating and validating the proposed LM algorithm (i.e., finalizing system inputs/outputs, determining the number of neurons in each layer, and identifying the significance of the proposed module) before they implement it in a real-time operation.

The only disadvantage of renting a cloud service is the processing time, which presents a dramatically non-linear relation between the cost of leasing the service for analytic purposes and the cost of targeting payback retention for the gained knowledge. Chang and Wills [26] designed an architecture module to compare the non-cloud and cloud storage of big data for organizational sustainability modeling. Their results illustrate high consistency between the actual and expected execution times using the cloud platform but low consistency on the non-cloud platform. Thus, the cloud as a platform can reliably be implemented for big data storage and LM execution.

Big data present challenges to further research in conventional computing science, thereby promoting the search for modern analytic utilities. Good planning with proper selection of needed analytic tools can result in a promising LM module to be used in the field of intelligent computing science.

LM Algorithms

AI programming is rapidly growing in various applications. In fact, LM algorithms can be applied in any field of study that involves the ability to learn from data and to predict the behaviour of a particular output. The most widely used LM algorithms were listed by Brownlee [27]. This study attempted to identify the LM algorithms that are applicable to and can deal with a big data knowledge base. Table 1 illustrates the available LM algorithms used for big data analytics.

Algorithm	Method	Advantage	Challenge
Extreme Learning Machine (ELM) [28]	Reviews the theoretical model of different ELM algorithms	High stability, speed, and accuracy under general or specific conditions Supports many learning applications, such as regression, classification, feature selection, clustering, and representational learning	Improves a data-dependent generalization mechanism for generating hidden-layer parameters
Artificial Neural Networks (ANNs) [29]	Demonstrate LM–ANN models to analyze and predict the output given by a data set	Compare different learning strategies	Generalization and optimization capabilities of the learning system
Online machine learning [30]	Online NN, Online Support Vector Machine (SVM), online Kernel Principal Component Analysis (KPCA)	The classifier can adapt or retrain the changes in the input data for prediction. The prediction and online classification processes are sometimes integrated for big data analytics.	An extensive demonstration in practical applications remains a significant challenge for online-learning methods.
ELM Clustering (ELMC) [31]	KMeans algorithm in ELM, non-negative matrix factorization (NMF) algorithm in ELM	An ELM feature space is supported by KMeans clustering. It can handle a large number of input parameters. NMF is tested for finding the low-dimensional representation of non-negative high-dimensional data, which can provide initiative data mapping to simplify the process. The overall performance has less effect with the changes in the hidden-layer nodes.	Further testing needs to be conducted, especially with other ELM feature-mapping techniques.
Unsupervised Discriminative Extreme Learning Machine (UDELM) [32]	Handles learning tasks with only unlabeled data	Merges the local manifold learning with global discriminative learning Gives better data representation than the ordinary unsupervised ELM, which conserves only the local structure of data	Generalization and optimization factors need to be enhanced.

Table 1: Recent development in the implementation of LM algorithms for big data analytics.

Many other LM algorisms are applied to different applications, such as parallel processing, AI techniques, data processing, and elastic ELM algorithm [33].

Conclusion

The LM concept is being increasingly adopted in current and future trends in big data implementation. This paper presents the challenges that are faced by various LM tools to provide an adaptable framework that fits in the big data analytic domain. Analytic modules can be integrated with the LM engine to overcome the circumstances of data processing. This technique is practical to prepare sample sets in LM analysis. Big data analytics and LM implementation support each other and can be powerful tools to understand and predict business behaviour based on customer information input. Any organization can improve its performance by analyzing the current situation and taking the right decision. However, current hardware utilities are encouraging new technologies to integrate AI algorithms in many applications, especially in solving classification problems, such as image processing and detection. Making a deal with a third-party service may help an organization arrive at a better decision. The future of LM focuses on unsupervised learning (i.e., dealing with unable or unstructured data), in which clustering and density estimation are applied.

References

1. https://cloud.google.com/prediction/.
2. https://aws.amazon.com/machine-learning/.
3. https://azure.microsoft.com/en-us/services/machine-learning/.
4. Assuncao MD, Calheiros RN, Bianchi S, Netto MAS, Buyya R (2015) Big Data Computing and Clouds: Trends and Future Directions. J Parallel Distrib Comput 79: 3-15.
5. Chen PCL, Zhang CY (2014) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci (Ny) 275: 314-347.
6. Jin X, Wah BW, Cheng X, Wang Y (2015) Significance and Challenges of Big Data Research. Big Data Res 2: 59-64.
7. Gudivada VN, Rao D, Raghavan VV((2014) No SQL Systems for Big Data Management. in 2014 IEEE World Congress on Services 190-197.
8. Solid IT (2015) DB-Engines Ranking - Trend Popularity.
9. Landset S, Khoshgoftaar TM, Richter AN, Hasanin T (2015) A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J Big Data 2: 24.
10. Wang G, Zomaya A, Perez GM, Li K, Eds (2015) Algorithms and Architectures for Parallel Processing. Cham: Springer International Publishing 9529.
11. Wang B, Huang S, Qiu J, Liu Y, Wang G (2015) Parallel online sequential extreme learning machine based on MapReduce. Neurocomputing 149: 224-232.
12. Al-Jarrah OY, Yoo PD, Muhaidat M, Karagiannidis GK, Taha K (2015) Efficient Machine Learning for Big Data: A Review. Big Data Res 2: 87-93.
13. Fan J, Han F, Liu H (2014) Challenges of Big Data analysis. Natl Sci Rev 1-38.
14. Jian CY, Li GS, Hu JQ (2013) A Modular Prediction Mechanism Based on Sequential Extreme Learning Machine with Application to Real-Time Tidal Prediction in Extreme Learning Machines 2013: Algorithms and Applications. Springer 2014: 35-53.
15. Schulz, Karolus J, Janssen F, Schweizer I (2015) Accurate pollutant modeling and mapping: Applying machine learning to participatory sensing and urban topology data in Proceedings - International Conference on Networked Systems. NetSys.
16. Lv Y, Duan Y, Kang W, Li Z, Wang FY (2015) Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE Trans Intell Transp Syst 16: 2.
17. Zhao M, Ding X, Shi Z, Yao Q, Yuan Y, et al. (2016) An efficient active set method for optimization extreme learning machines. Neurocomputing 174: 187-193.
18. Taormina R, Chau KW (2015) Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. J Hydrol 529: 1617-1632.
19. Benoit F, Heeswijk MV, Miche Y, Verleysen M, Lendasse A (2013) Feature selection for nonlinear models with extreme learning machines. Neurocomputing 102:111-124.
20. Aradau C, Blanke K (2015) The (Big) Data-security assemblage: Knowledge and critique. Big Data Soc 2: 1-12.
21. Almutairi M, Abdullah R, Albukhary T, Kar J (2015) Security and Privacy of Big Data in Various Applications. 2: 19-24.
22. Chen H, Bhargava B, Zhongchuan F (2014) Multilabels-Based Scalable Access Control for Big Data Applications. IEEE Cloud Comput 1: 65-71.
23. Yu Y, Mu Y, Ateniese G (2015) Recent Advances in Security and Privacy in Big Data J.UCS Special Issue. J Univers Comput Sci 21: 365-368.
24. Rozenfeld M (2014) Standards That Support Big Data. The IEEE news source.
25. Cheng Y, Qin C, Rusu F (2012) GLADE: Big data analytics made easy. Proc ACM SIGMOD Int Conf Manag Data 3.
26. Chang V, Wills (2016) A model to compare cloud and non-cloud storage of Big Data Futur Gener Comput Syst 57: 56-76.
27. Brownlee J (2013) A Tour of Machine Learning Algorithms. Machine Learning Mastery.
28. Huang G, Bin Huang G, Song S, You K (2015) Trends in extreme learning machines: A review. Neural Networks 61: 32-48.
29. Sharma C (2014) Big Data Analytics Using Neural networks. San Jose State University.
30. Motai Y (2015) Kernel Association for Classification and Prediction: A Survey," IEEE Trans. Neural Networks Learn Syst 26: 2.
31. He Q, Jin X, Du, Zhuang F, Shi Z (2014) Clustering in extreme learning machine feature space. Neurocomputing 128: 88-95.
32. Peng Y, Zheng WL, Lu BL (2015) An unsupervised discriminative extreme learning machine and its applications to data clustering. Neurocomputing 174: 250-264.
33. Xin J, Wang Z, Qu L, Wang G (2015) Elastic extreme learning machine for big data classification. Neurocomputing 149: 464-471.