ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 5, May 2018

Machine Learning Approaches for Data analytics and Modeling: A Comprehensive Study

[1] Dr.M.Vinoth Kumar, [2] N. Girish, [3] S. Babu Kumar
[1] Associate Professor, Department of ISE, Dayananda Sagar Academy of Technology and Management, Bangalore
[2][3] Assistant Professor, Department of ISE, Dayananda Sagar Academy of Technology and Management, Bangalore,

***Abstract*: - Recently Machine learning is a growing technology in various applications of academia and Industry which includes healthcare, social media, agriculture, economy and finance. It plays an essential role in data mining to handle the huge data generated and maintained by the different machine. Big data analytics has been geared as a driving force to reinvent how machine learning techniques are used for data analytics with high dimensional features in heterogeneous format. This paper presented a comparative study of machine learning techniques for big data analytics and modeling used by data scientist. We apply supervised learning for perfect and imperfect domain knowledge to fulfill the vision of machine learning data analytics and modeling.**

Keywords— Machine learning, big data, data analysis, data modeling.

I. INTRODUCTION

Machine Learning has become one of the mainstays of information technology over the past two decades. Machine learning provides example data and past experience as a data to optimize the performance of a computer system. It is evolved from Computational Learning and pattern recognition Theory. A learning algorithm takes a set of sample data as an input named as a training set. In general, there are mainly three categories of learning: supervised, un-supervised, and reinforcement. In supervised learning, the training set consists of samples of input vectors together with their corresponding appropriate target vectors, also known as labels. In unsupervised learning, no labels are required for the training set. Reinforcement learning deals with the problem of learning the appropriate action or sequence of actions to be taken for a given situation in order to maximize pay off. With the ever increasing amounts of data becoming available there is a smart data analysis process will become even more pervasive as an important ingredient for technological growth. Basically Machine learning involves many tasks such as recognition, diagnosis, planning, robot control, prediction, etc.

Supervised Learning
Supervised machine learning designs a model for predictions based on evidence in the presence of uncertainty. A learning algorithm takes a known set of input trained data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new generated data. Supervised learning is used for known data for the output which can able to predict. It uses classification and regression techniques to develop predictive models.

Unsupervised Learning
Unsupervised learning finds hidden patterns in data. It is used to draw inferences from input datasets consisting of data without labeled responses. One of the most common unsupervised learning techniques is clustering. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include object recognition, market research and gene sequence analysis.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 5, May 2018

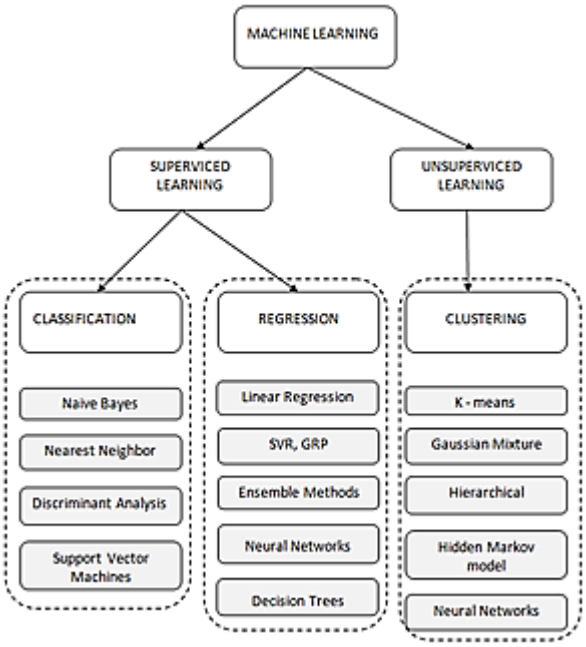

Figure 1. Machine learning techniques

II. METHODOLOGY

Developed by specialized analytics experts systems and software, big data analytics can explore the way to various business applications, including more effective marketing, new revenue opportunities, better customer service, and competitive advantages over rivals. Big data analytics applications often include data from both internal and external sources, such as weather data or demographic data compiled by third-party information services providers. Once the data is ready, it can be analyzed by advanced analytics processes; predictive analytics, machine learning and deep learning

We propose machine learning algorithm for data analytics with six stages

(1)Data acquisition
(2) Extracting and loading
(3) Aggregation
(4) Data analytics,
(5) Verification and actions
(6) Intrusions detection.

Data acquisition is the process of gathering, filtering, and cleaning data before the data is stored in a storage solution. The acquisition of big data is managed volume, velocity, variety, and value of the data. Most data acquisition scenarios assume high-variety, high-volume, high-velocity but low-value data, making it important to have adaptable and time-efficient gathering, filtering, and cleaning algorithms that ensure that only the high-value fragments of the data are actually processed.

Extracting and loading allows raw data to be loaded directly into the target and transformed there itself. The processing capability of ETL is built into a data warehousing infrastructure that reduces the time that data spends in transit and is more cost-effective. Data aggregation gathered data and expressed it in a summary form, for purposes such as data statistical analysis to get more information about particular groups based on specific variables. Data analytics is examining large and varied data sets to uncover hidden patterns, unknown correlations, and other useful information. It provides a means of analyzing data sets to help organizations make informed business decisions. It involves in predictive models, what-if analyses and statistical algorithms powered by high-performance analytics systems.

Data modeling is designs data structures at various levels of abstraction from conceptual to physical model. It describes logical design of the system, where as design describes physical implementation of a data and database. An intrusion detection monitors network traffic and monitors for suspicious activity and alerts the system.. It may also responds to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the networks. It is functioned by determining whether a set of actions can be deemed as intrusion on a basis of one or more models of intrusion. The growth of data mining methods has consequently brought forth a wide range of algorithms drawn from areas as pattern recognition, machine learning and database analysis. There are many types of algorithms that may be used to mine audit data. Lazar identifies data algorithms as a set of heuristics and designs between data mining models.. In effect, the author suggests that for a model to be formulated, the algorithms must start by analysing the data type provided, in order to find particular trends and patterns. These results of analysis are later used by the algorithm for defining optimal parameters to create the selected mining model. The parameters are applied across the dataset, together with selected patterns and detailed statistics.

III. DATA ANALYTICS AND MODELING

Data analytics and modeling are reviewed in terms of the three representative approaches; decision support databases, operational databases and Big Data technologies. Relational model and ER model for operational databases; star schema model and OLAP cube model for decision support databases; and key-value store, document-oriented database, wide-column store and graph database for Big Data-based technologies. Regarding data analytics, it observes that operational databases are more suitable for OLTP applications and decision support databases are more suited for OLAP applications. Big Data technologies are more appropriate for scenarios like batch-oriented processing, stream processing, OLTP and interactive ad-hoc queries and analysis.

IV. CONCLUSION

The supervised learning technique is useful for big data analytics and modeling on the intelligent security. The structured domain knowledge is modeled and approximate queries provide a type-labeled training dataset for a machine learning algorithm to evaluate its training and testing error rates of the data. In perfect domain knowledge learning, a security expert has all type-labeled instances, so this provides an optimal performance benchmark. Therefore, we can apply the decision tree algorithm to enforce the supervised learning.

REFERENCES

[1] Yuh-Jong Hu et al.,"Structured Machine Learning for Data Analytics and Modeling: Intelligent Security as An Example," IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2015

[2] Athmaja S. et al .," A survey of machine learning algorithms for big data analytics," International Conference on Innovations in Information, Embedded and Communication Systems, 2017.

[3] Suresh Kumar P and S. Pranavi, "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics" International Conference on Innovations in Information, Embedded and Communication Systems, 2017.

[4] Ananthi Sheshasaayee and J V N Lakshmi "An insight into tree based machine learning techniques for big data Analytics using Apache Spark" International Conference on Intelligent Computing,Instrumentation and Control Technologies, 2017.

[5] Anna L. Buczak and Erhan Guven, "A Survey of Data Mining and Machine Learning

[6] Methods for Cyber Security Intrusion Detection", IEEE Communications surveys & tutorials, vol. 18, no. 2, 2016

[7] Michael Mayhew et al., "Use of Machine Learning in Big Data Analytics for Insider Threat Detection," Cyber security and trusted computing, 2015

[8] Ananthi Sheshasaayee and J V N Lakshmi, "Machine Learning approaches on Map Reduce for

[9] Big Data Analytics ", IEEE Communications surveys & tutorials, vol. 11, no. 2, 2015

[10] Rayner Alfred, "The Rise of Machine Learning for Big Data Analytics" 2nd International Conference on Science in Information Technology, 2016.

[11] Yu-Xuan Wang et al .," Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization" Proceedings of the IEEE International Conference on Applied System Innovation, 2017.

[12] Shimei Jin et al .," Graph-Based Machine Learning Algorithm with Application in Data Mining" Proceedings of the IEEE Third International Conference on Research in Computational Intelligence and Communication Networks, 2017.

[13] Alex Kaplunovich and Yelena Yesha, "Cloud Big Data Decision Support System for Machine Learning on AWS" IEEE International Conference on Big Data, 2017