International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Study on Machine Learning with BigData

Jithin Sebastian, Rooben CJ, Johnsymol Joy

Bca Final Year, Saintgits College Of Applied Science, Kottayam, 686532, India

A B S T R A C T

Machine learning is used to take sharp decisions. Big data helps to identify new opportunities. By including machine learning algorithm to big data, we could get analyzed and defined results. Using machine learning computers get into a self learning mode without explicit programming. This paper gives an idea about association of big data with machine learning, what main technologies are associated with big data and also the application of machine learning in big data.

Keywords: Machine Learning, Big data, Information Technology

1. Introduction

Machine learning is a crucial area of AI . The objective of machine learning is to get knowledge and make intelligent decisions. Machine learning algorithms can be divided into supervised, unsupervised, and semi-supervised . Another division of machine learning based on the output of a machine learning system includes classification, regression, clustering, and density estimation, etc. Machine learning approaches comprise decision tree learning, association rule learning, artificial neural network , support vector machines (SVM), clustering, Bayesian networks, and genetic algorithms, etc. Machine learning is used widely in big data and big data is a combination of both structured and unstructured data which is so huge that it is very difficult to process it with old techniques and database. This paper will introduce machine learning & its applications in big data and the challenges faced and the technology advancement of machine learning in big data. .

1.1. Techniques of Machine learning with Bigdata

Supervised learning are often categorized into classification and regression. When the category attribute is discrete, it’s classification; when the category attribute is continuous, it’s called regression. Decision tree learning, naïve Bayes classifier, and k-nearest neighbor (k-NN), etc. are classification methods. Linear regression and logistic regression are regression methods. Unsupervised learning categorize instances into groups of similar objects.

Clustering can be grouped into 3:-

*SUPERVISED CLUSTERING

*SEMI-SUPERVISED CLUSTERING

* *Corresponding author.*
E-mail address: jithinsebastian789@gmail.com

*UNSUPERVISED CLUSTERING

Decision trees classify data supported their feature values. Decision trees are constructed recursively from training data employing a top-down greedy approach during which features are sequentially selected. Decision tree classifiers arrange the training data into a tree-structure plan.

Support vector machine (SVM) is a binary classifier that finds linear classifier in higher dimensional feature space to which original data space is mapped. SVM shows excellent performance for data sets during a moderate size. It has inherent limitations to big data applications.

Deep machine learning has become a search frontier in AI . It is a machine learning technique, where many layers of data processing stages are exploited in hierarchical architectures. It enumerates hierarchical attributes or representations of the observational data, where the higher-level features are explained from lower-level ones.

1.2. 5 V's of Bigdata

- Volume: The solemnity that the term big data owns is due to its volume.
- Velocity: the prevalence of processing data with high accuracy and speed.
- Variety: the various sorts of data, Structured, Unstructured and Semi-Structured.
- Veracity: the standard and consistency of knowledge .
- Value: The end-stage is to withdraw useful data.

1.3. Machine learning applications with Bigdata examples

- To test the connection between crime and social characteristics. This approach decreases the crime data's dimensionality.
- Use unsupervised machine learning to break crime data into groups; use the clustering algorithm k-means to cluster crime data into risky, average, and stable regions.
- Use supervised machine learning approaches to predict whether a specific area is dangerous or safe.

1.4. Challenges of Machine-learning applications with Bigdata

General machine learning challenges are:

(I)the design of scalable and versatile computational frameworks for machine learning;
(ii) the ability to understand data characteristics before implementing algorithms and tools for machine learning;
(iii) the ability to build, learn and infer with increasing sample size, dimensionality and label categories.

Big Data Analytics present specific challenges for machine learning and data analysis in addition to analyzing large data volumes, including format variance of raw data, fast-moving streaming data, data analysis trustworthiness, widely distributed input sources, noisy and low quality data, high dimensionality, algorithm scalability, imbalanced input data, unsupervised and un-categorized information, minimal supervised/labeled information, etc.

Other main issues in Big Data Analytics are sufficient data storage, data indexing /tagging, and fast retrieval of information. Consequently, when dealing with Big Data, creative data processing and data management solutions are warranted.

1.5. Technologies associated with Machine learning

Big Data's future is transparent and unshakeable. Technologies such as IOT , machine learning, artificial intelligence, and more, if you have heard, are finding their way into our daily lives. Big Data sits squarely in an authoritative role behind all of these. Some of the technologies associated with big data are listed below:-

ARTIFICIAL INTELLIGENCE

Artificial intelligence through learning will alter data analytics (using results from previous analytical tasks) so that results will arrive faster and much more precision over time. Artificial intelligence would be able to make detailed future forecasts based on current events with vast quantities of data to draw from, along with empirical findings from past queries.

EDGE-COMPUTING

The benefit of an edge-computing system is that it decreases the amount of data to be transmitted across the network, thus minimizing network traffic and associated costs. It also lowers requirements for data centers or cloud storage services, frees up room for other workloads and removes a single possible failure point.

IN MEMORY-DATABASES

As organizations pursue fast and simple access to data and analytics to inform many business decisions, the adoption of in-memory computing is growing. In-memory computing provides the insights they need to improve operational, financial, marketing, and sales productivity. As in-memory computing developments occur, it is becoming more accessible and simpler to introduce, making it widely available.

NO-SQL DATABASES

Structured data contained in relational database management systems is queried, manipulated and handled by database administrators (RDMS).

NoSQL databases, on the other hand, store unstructured data and provide rapid output. This implies that it provides versatility when managing a wide range of large quantities of data forms. MongoDB , Redis , and Cassandra provide several

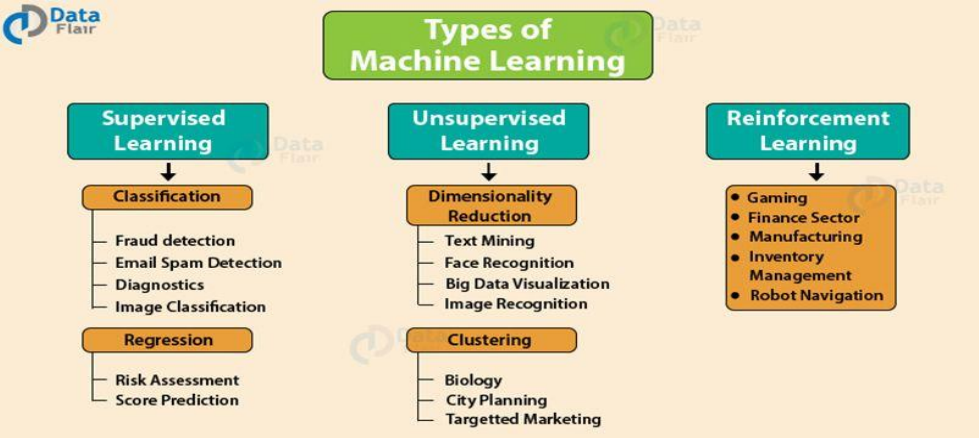

Fig. 1 - (a) Types of machine learning

2. Illustrations

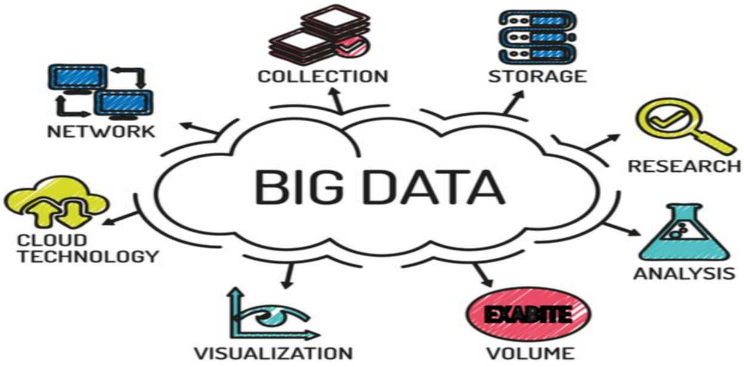

Fig. 2- (b) Challenges of Machine learning with big data

3. Conclusions

Based on certain quality metrics, the splitting criteria of decision trees are selected, which requires handling the entire data set of each expanding nodes. This makes it hard for decision-making trees to be used in software involving big data. SVM demonstrates very good efficiency at a modest size for data sets. Big data systems have inherent limits. Deep learning is suitable for solving problems related to the amount and variety of big data. However, since it needs a lot of preparation time, it has certain limitations on big data.

Big data machine learning applications have faced challenges such as memory limitation, difficulty dealing with big data due to speed, volume, and variety, etc., and learning training restricted to a number of class types or a specific labeled dataset, etc. The technologies associated and the challenges faced in machine learning in big data can also be used for further research topics.

References

- Researchgate.net/publication (2016),Machine learning in big data.
- Li. L ,(2015) Experimental comparison of multiclass classifiers,Informatica,39,71-85
- Qian . H (2014).Pivotal R:A package for machine learning on big data
- Sciencedirect.com,(2019)”Challenges of Big data Technologies”.
- An Introduction to Computational Learning Theory (M. Kearns and U. Vaziran)
- Online Learning (by N. Cesa-Bianchi)