Utilization of Dataset for Data Prediction with Machine Learning a Review

Meenakshi Kondal and Virender Singh

October 3, 2022

Utilization of Dataset for data prediction with Machine Learning A Review

Meenakshi Kondal[1]
Dr. Virender Singh[2]

Faculty of Computer Science,
Goswami Ganesh Dutta Sanatan Dharma college,
Chandigarh, India.
meenakshi.kondal@ggdsd.ac.in

Assistant Professor, Department of Information Technology,
Goswami Ganesh Dutta Sanatan Dharma College, Chandigarh, India
virender@ggdsd.ac.in

Abstract- With the convergence of computing and communication, we feed on information in its raw form of data. There is a huge of information locked up in databases that have not yet been discovered or articulated. Data plays a vital role for applications in assigning a category, object localization, behavior analysis, and image retrieval. Nowadays, data accomplish with complex mathematical tasks like lines of code, memory, and speed limitations. Organizations now focus on analyzing data that are getting accumulated and involve acquiring knowledge from reliable data sources, rapidity in processing information in deploying analytics to forthcoming challenges. The prediction in this research paper is applied to interpret the data set with the use of machine learning techniques/models to analyze the forecast value from the comparative study. Machine learning provides the technique to extract information from raw data that is expressed in a comprehensible form and can be used for a variety of purposes.

Keywords: Resilient Distributed Datasets, Hadoop Distributed File System, Regression, Job-tracker.

INTRODUCTION

In computers, data is altered in a form for processing and analyzing but in the real world; data is immense and has a complex structure with the existing architecture. Data emerges from infinite sources which are classified in various forms such as structured, unstructured, and semi-structured [1][2]. The rapid growth of data provides opportunities for data analytics. Data Analytics is the scientific and statistical tool for analyzing raw data to renovate information for acquiring knowledge.

The role of analytics is to assemble, store, process, and analyze data to address empirical methods in the real world for decision-making. The analytics on data reveals hidden patterns, unfound correlations, market trends, consumer requirements, and future recommendations, which assist in the critical decision-making process[1]. Machine Learning is the existing and current automated technology involved in handling and evaluating data structures. It is a technical tool of data science that creates logic from data by transforming data into knowledge[6]. Many powerful algorithms of machine learning are developed to learn patterns, acquire insights, and forecasting events. Machine Learning techniques are classified into two types supervised and unsupervised learning[2][6]. Supervised learning techniques emphasize accurate predictions whereas unsupervised learning work on compact descriptions of the data. Machine Learning techniques learn to understand the complex datasets for critical decisions and tune the features for extraction of high performance[18].

DATA ANALYTICS MODELS

Machine Learning provides efficient analysis models for capturing knowledge by improving prediction for data-driven decisions. It plays an eminent role in the field of computer science that paves its way in analyzing robust emails, spam filters, convenient text, voice recognition, web search engines, and game developments [6]. Learning is an activity in which a model is tuned to solve various problems by understanding the characteristics of parallel distributed data. Learning models are developed by the input data feed into the system. Since the data is complex and massive it is essential to perform computations on an environment by using machine techniques and to adapt to new complexities [15].

In this modern era, a large assortment of data is immersed that starves for knowledge. This abundant data can be in structured or unstructured forms. Structured data is the data that is arranged in tables and unstructured is the data that is in irregular forms such as images, documents, text, audio, video[5]. Human intervention is reduced by automated machines to build models for processing huge amounts of unstructured data. Machine Learning is an application of science that creates logic from data by transforming them into knowledge[9]. Machine Learning consists of many powerful algorithms to learn patterns, and acquire insight, and predictions. Artificial Intelligence(AI) evolved as a subfield for the development of self-learning algorithms for insight Machine Learning provides efficient analysis models to capture knowledge by improving prediction and data-driven decisions[10]. Learning process involves following procedures for processing massive data sets.

The steps involved in processing the datasets
Step1: Collection of data.
Step 2: Partition the data as trained datasets and test by identifying labels.
Step 3: Deploy the Machine Learning algorithms on models. Step 4: Return output as per the requirement specified.

Machine Learning methods are classified into two techniques. These are the Supervised Learning Method and Unsupervised Learning Method[1][2]. Each learning method has its significance and dimensionality. In the supervised learning method, a model from labeled training data is interpreted to predict test data[11]. The term supervised learning method refers to labels that are known as a set of samples to attain the desired outcome. Supervised learning is categorized into two tasks: Classification and Regression. Consider an example of e-mail spam filtering; a model using an e-mail corpus is trained with supervised learning methods to identify the spam[15].This model is tested to detect the occurrence of spam[3][4]. A learning task with discrete class labels that use binary sorting to predict a new instance from past observations is known as classification. An expected task with continuous unordered class labels to forecast new instances and to find the relationship between those variables is known as Regression. It is an associated method for prediction analysis.

The unsupervised Learning method deals with unlabeled data or unknown structures. The data is explored for information extraction to analyze the outcome variable[12]. For feature selection, different algorithms are used to reduce the dimensionality of a dataset. Another approach for dimensionality reduction is feature extraction. Summarization and Understanding are the two techniques for N-

Dimensional space in predefined structures [13]. This technique automatically returns amazing outcomes by implementing grouping and learning methods on data points. Learning is uncertain and unpredictable as they produce random results for a different choice of features. Clustering is an unsupervised learning technique to cluster similar features, a technique to group similar objects on different metrics. The procedure for clustering is to collect information from different consent [14]. Each cluster is sorted by crafting the datasets that influence the model. Finally, these crafted clusters campaign for metric computations. It is an exploratory data analysis methodology, which organizes the data in meaningful and structural blocks devoid of any experience in group membership. Each block in a cluster analyzes the groups and shares the degree of similarity between objects for classification[15]. This technique derives significant relationships among datasets.

SPARK MODEL

Spark has a unique library for Machine Learning called MLlib. It solves a wide range of data problems attributed to streaming, graph computation, and real-time interactive query processing. MLlib assists by leveraging the scale and speed that builds specialized use cases for varied analytical models[11][12]. A spark is a well-developed tool, which manages the execution of tasks across a cluster of computers. A cluster or group of machines utilized by the Spark framework pools the resources of machines by allowing the usage of cumulative resources at a single point in time, to accomplish these goals Spark uses RDDs (Resilient Distributed Datasets) for partitioning the datasets on cluster computers by avoiding data loss. Spark is the earliest system that allows an efficient general-purpose programming language for an interactive process of large datasets on a cluster[13].

Spark provides a unified engine provisioned with ease of usage and a faster attitude for large-scale data processing. It implements iterative Machine Learning workloads by interacting and scanning datasets parallel to sub-second latency[2]. Today’s organizations create dissimilar data from diverse sources, which are user-centric. Developing personalization, recommendation, and predictive insights are major goals for advancement in organizations. Spark is designed with advanced features that incorporate various features like simplicity, scalability, easy integration, compatibility, and speed. These are employed to solve challenges faced by organizations[15]. Diverse use cases are applied that adapt quickly with the iterative models to meet the requirements of organizations. The recent popularity of Spark models invokes significant interest in implementing scalable versions of Machine Learning algorithms. Machine Learning is the existing and current automated technology involved in handling and evaluating massive data structures. The process of developing a mixture of information/models contributes to the quality of the final system. These models are designed for guiding the experimental study in the analytics field of research[11][10].

SPARK FRAMEWORK

The main concern of the systems is the amount of computational time. Speed is a key component for large datasets, which are efficient for query processing and computation. The limitations faced by the programming model, which are high disk rate, low throughput, and diminishing performance of a cluster are addressed by Spark[3]. Spark framework was introduced by Apache Foundations for parallel distributed Systems. It is an independent technology. Spark has its cluster management architecture for better storage[4]. The significant feature of Spark is its in-memory cluster computing for enhancing the processing speed of an application. The additional feature of Spark is it has a powerful processing unit for fast computations. Advanced methods like interactive query analysis and streaming are extended features of the Spark framework model. Spark covers a wide range of workloads such as Batch applications, Iterative algorithms, and Interactive queries[12].

HADOOP FRAMEWORK

The modern technological world faces a tremendous increase in data, which is complex to handle on a single machine. Therefore, it is essential to distribute the partitioned data blocks on multiple machines. In a nutshell, Hadoop is an effective analysis and reliable data storage system[5]. In today’s scenario, every organization faces lots of challenges especially in analyzing the stream of data. But a large amount of data when to read and written through multiple disks in parallel mode consumes maximum time in data analytics[15].A file system distributed across a network of machines is referred to as distributed file system. Data are scattered over various locations on the network. It is typical to manage and control the flow of data with a wide number of complications in network programming. For an instance, the biggest challenge is node failure in a distributed environment[18]. The file system used in the Hadoop ecosystem is HDFS (Hadoop Distributed File System) [5]. HDFS key components are Name Node and Data Nodes. Hadoop Distributed File System uses master-slave architecture. Data is distributed on racks in Hadoop Cluster. Job-tracker and task-tracker are two housekeeping components, which are scattered on the cluster [6].

Hadoop Distributed File System of Data stores a massive amount of information that scales up exponentially and survives the failure of storage infrastructure without data loss. The problem raised during the usage of these systems is the network bandwidth[18]. In certain cases, when the data is too large for transmission the employed bandwidth and entire system collapse resulting in performance issues because the computing nodes need to be idle for a long time. Hadoop Distributed File System (HDFS) system creates triplet instances of the same data scattering them on different servers[15][16]. A general way of avoiding data loss is through replication. Any failure or data loss can be handled by utilizing the redundant copies. The distributed and extracted data from many sources needs to be combined and this is currently the most formidable challenge. Hadoop is an open-source distributed file system implementation. It provides a heterogeneous storage system for data loading with fine-grain fault tolerance and execution of more complicated functions through simple SQL queries[17]. The implementation relies on an in-house cluster management system for the distribution and execution of user tasks on shared machines. Apache Hadoop provides a versatile ecosystem with data processing tools featuring automation, scaling, and built-in error torrents [7].

MAP REDUCE MODEL

MapReduce is a data flow paradigm for data-centric applications. It is a simple explicit data flow programming model and preferential over the traditionally high-level database approaches [8]. MapReduce paradigm parallelizes huge data sets using clusters or grids. Data is a large data processing system developed to solve the issues of various traditional methods. As it has large voluminous data, privacy turns out to be a major challenging issue. To provide acute security the critical modeling techniques require analytical processing, effective storage, and successful retrieval techniques. Therefore, a massive parallel programming model termed MapReduce is employed on a distributed framework that provides a high computational environment for Data Analytics[1][2]. This model has two significant functions Map and Reduce to process the huge data structures. A program comprises two functions, a Map Procedure and a Reduce Procedure. These functions are used for processing large-scale datasets on computer clusters. Each map task processes the data input and produces key-value pairs. The output produced by each map task is further taken as an input for reducing the task. Finally, passing through many phases of reduce tasks such as shuffle, merge and sort function the MapReduce model emerges with a final summation output. The process of constructing such a fusion of information/models has a tremendous impact on the quality of the final system[8]. In concise this model is designed for guiding the experimental study in the analytics field of research. Despite the significant features, MapReduce has many limitations. It suffers from high disk reads and writes, and low throughput, which results in low performance of a cluster, low latency, and poor reading structure.

MapReduce divides the input into fixed partitions called input splits, which run in the user-defined map function on each record. The time required to process the map function on splits depends on the number of maps used on a cluster[8]. It takes a minimum amount of time for the higher number of map counts with increased performance. Similarly, the performance decreases with a lesser map count. The splits that occur minute are parallel processed for improving better load balancing. A swift system processes more split into one course of action. Machines existing in the cluster environment are prone to failure[7]. If some process execution fails, then the job running on a map task is transferred to another map task to maintain load balancing in which the number of splits is increased for fine grain. In other cases, if splits are too small it increases the overhead problem consequently increasing the disk reads/writes, which results in low performance. This increases the overhead by escalating the total execution time

SUMMARIZATION OF PROCESSING TOOLS

Tools Utilized	Purpose
Spark	In Built distributed frame work
Hadoop	Effectual analysis and reliable data storage
MapReduce	Framework for data distribution
Machine Learning	Techniques Developing an analytical model
Supervised Learning Methods	Prediction
Unsupervised Learning Methods	Uncertain and unpredictable

Table 1 gives a detailed description of the processing tools used in this research work.

CONCLUSION

This paper provides a description of Machine Learning techniques for structural data backgrounds. In addition, it also stated the role and the importance of Spark MLlib library construction, supervised learning techniques, features of Data analytics, and advantages of Spark framework and Hadoop structures have been emphasized for Machine Learning applications.

REFERENCES

[1] Douglas T., Todd T., and Miron L. (2004) 'Distributed computing in practice: The condor experience' in Concurrency and Computation Conference.

[2] Milos Hauskrecht (2017),'Ensemble Learning' in Book Chapter 12 Data Mining pp: 479-501

[3] Andreas K., Nikolaos N., Giannis T. and et.al. (2017) 'Large Scale implementation for Twitter Sentiment Classification' in MDPI Journal Vol 10 pp: 1-21.

[4] https://DataBricks Import – How to guide/documentation/

[5] Abouzeid K., Bajda P., Rasin.A, and et.al (2009), 'Hadoopdb: An Architectural Hybrid of MapReduce And DBMS Technologies For Analytical Workloads', PVLDB.

[6] S. Ananthi, (2015) 'A Theoretical Model for Big data Analytics using Machine Learning Algorithm' International Conference (WCI/ICACCI 2015), India, pp.632-636.

[7] Rabi Prasad P., (2013) 'Big Data Processing with Hadoop-MapReduce in Cloud Systems', International Journal of Cloud Computing and Services Science, Vol.2, No.1, February 2013, pp: 16-27.

[8] Hadoop Map/Reduce tutorial http://hadoop.apache.org/common/docs/r0.20.0/mapred tutorial.html.

[9] https://jaxenter.com/Hadoop-core-architecture-components/

[10] https://jaxenter.com/machine-learning-an-introduction-for-programmers-122135.html

[11] Caruana, Rich, Nikos K., and Ainur Y. (2008) ‘An Empirical Evaluation of Supervised Learning in High Dimensions’ Conference on Machine Learning, ACM.

[12] http://blog.aureusanalytics.com/unsupervised-learning/

[13] Davidov, D., Rappoport, (2006) ‘A. Efficient Unsupervised Discovery of Word Categories Using Symmetric Patterns and High Frequency Words’. In Proceedings of the International Conference on Computational Linguistics, Sydney, Australia, 17–21 July 2006; pp. 297–304.

[14] http://blog.aureusanalytics.com/unsupervised-learning/

[15] Latchoumi T.P., Jayakumar L., Janakiraman S. et.al (2016) ‘OFS method for selecting active features using clustering techniques’ in Informatics and Analytics conference. ISBN: 978-1-4503-4756-3.

[16] Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big data. 2016;

[17] Sarker IH, Kayes ASM. Abc-ruleminer: user behavioral rule-based machine learning method for context-aware intelligent services. J Netw Comput Appl. 2020; page 102762

[18] Sarker IH, Alqahtani H, Alsolami F, Khan A, Abushark YB, Siddiqui MK. Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling. J Big Data. 2020;

[19] Rokach L. A survey of clustering algorithms. In: Data mining and knowledge discovery handbook, pages 269–298. Springer, 2010.

[20] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl Syst. 2015;89:14–46.

[21]Kondal, M. and Singh, V., Comparative Analysis of Tineye and Google Reverse Image Search Engines.

International Journal of Innovative Science and Research Technology. Volume 7, Issue 3, March – 2022.

[22] Otter DW, Medina JR , Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst. 2020.

[23]Mohammed M, Khan MB, Bashier Mohammed BE. Machine learning: algorithms and applications. CRC Press; 2016.

[24]McCallum A. Information extraction: distilling structured data from unstructured text. Queue. 2005;3(9):48–57.