International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

Latest Trending Techniques and ML Models for Big Data Analytics - A Review

Princy Tyagi
Department of Computer Science & Engineering, Swami Rama Himalayan University

Abstract- As businesses increasingly rely on Big Data for decision-making, understanding cutting-edge methods is crucial. This review synthesizes recent advancements, highlighting the most impactful techniques and ML model algorithms are essential for analyzing and drawing insights from vast amounts of data. This research paper explores the latest techniques and ML algorithms used in big data analytics, including deep learning, neural networks, and decision trees. The challenges of big data analytics are discussed along with how these techniques can help overcome them. Real-world examples of how these techniques have been used in big data analytics are also provided.

Keywords: Big data, data analysis, artificial intelligence, information retrieval.

1. INTRODUCTION

With the continuous increase in data generated every day, big data analytics has become crucial. Examining this data to extract insights and information that can enhance decision-making, drive innovation, and increase business value is the essence of Big Data analytics. ML algorithms are the main tools used in Big Data analytics, as they can help automate the analysis process and identify patterns in large datasets.

Finding the pertinent information on big data, modelling the components of it, and turning it into usable knowledge and information are the typical steps in dealing with it. Most experts choose to employ ML techniques for knowledge discovery, Data Aggregation and Clustering, and modelling and prediction for these purposes. Data mining, which includes machine learning, offers ways to process automatically extracting detail from data when human resources are unable to do so due to either complexity or the volume of data within a limited timeframe [1].

For a while, machine learning was a discipline used in highly specialized fields like medical, earth sciences, or marketing. However, with the emergence of huge data, everyone now needs to handle it and automatically learn from it. Some focus on real-time data processing, others on anomaly detection, and some on distributed computing environments. However, all of these efforts aim to narrow the divide between artificial intelligence research and industry, which has endured until now. In addition to accuracy and algorithm complexity, these tools also address other critical aspects of applied machine learning, Examples include the capability to parallelize learning and prediction phases, the selection of a suitable coding language for modeling and data retrieval, the utilization of pre-existing high-performance libraries, and the approach for collecting and presenting results.

For those professionals who are unfamiliar with these BD (Big data) analytics has become increasingly important with rapid expansion of data generated every day. ML ideas, I will quickly describe machine learning and data mining in this fast perspective, I will list several tools and platforms multiple applications and platforms utilized by data scientists to handle extensive data and execute ML operations on it. As a result of our extensive experience with them at the BSR (Barcelona Supercomputing Centre) and our adeptness in educating scholars on data mining using these tools, please note that the tools described here are not the only ones available and are not always the ideal ones for every case. I urge the reader to use the tools I've described in their professional tasks and research, as well as to experiment with new and more advanced tools and languages.

2. DM AND ML

DM and ML are two interconnected fields that are used for extracting useful information from data. Data Mining involves uncovering patterns, relationships, and trends within extensive datasets, encompassing stages such as data collection, preprocessing, transformation, mining, and interpretation of results. ML, a subset of AI, is focused on designing and developing algorithms that learn patterns and make predictions based on data.

The science of data mining focuses on extracting knowledge and information from data. This discipline offers ideas and methodology on how to and how much we can learn from data sets in terms of meaningful information, as well as how to do so. The emergence of Big-Data has made data mining

techniques more and more significant, offering ways to extract some value from enormous databases or simply deal with larger volumes of data driving systems.

The relationship between DM and ML is that ML is one of the techniques used in DM for extracting information from data. DM utilizes ML algorithms for identifying patterns in data that can help make predictions or decisions. In contrast, ML is a component of DM that focuses on developing algorithms capable of learning from data and making predictions based on that acquired knowledge.

DM entails the utilization of diverse methodologies, including clustering, classification, association rule mining, and regression, where ML algorithms are utilized to extract knowledge from the data and generate predictions, For instance, in clustering, ML algorithms are employed to categorize data points into clusters according to their similarities, while in classification, ML algorithms are used to categorize data into different groups based on their features.

Machine Learning is a scientific discipline within Data Mining, centered on deriving models from pre-existing data in an automated manner. Consider a scenario where we aim to construct a model representing a system, with objectives ranging from enhancing our understanding of the system, producing forecasts for bespoke inputs resembling an oracle for guiding principles, to emulating its performance. In numerous instances, the intricacy of the system exceeds our capacity to manually recreate it, either due to insufficient detailed knowledge or rapid changes rendering manual model creation impractical.

ML Basics

- We can classify most ML techniques into two overarching categories depending on our modeling objectives: labeled and unlabeled learning. In Labeled learning. our objective is to retrospectively erive a model from empirical data, examples enabling the classification or estimation of new examples prospectively. Our observations, also known as instances, consist of factors and labels, resulting in a model represented as a function like f(x1...n) = ŷ1...m, where xi represents the features and ŷj denotes the produced labels. A model is considered satisfactory when the disparity between the produced labels ŷj. We can refine the modeling process by adjusting various modeling algorithm parameters (hyperparameters) until we achieve a satisfactory model. However, it is crucial to validate the chosen model using examples that were not part of the iterative training process. For instance, let's consider data gathered from a data center containing virtual machines that serve web services. Suppose our objective is to predict the workload generated by a VM by monitoring solely the number of clients accessing the service. The infrastructure provider collects data on session and packet volumes directed to a customer VM, as well as the workload associated with the VM, since direct examination of customers VM might not be permissible. The provider then employs a regression tree algorithm to create a model that, given the number of sessions and packets routed to the VM, forecasts the workload the VM will require under those conditions. This enables the provider to allocate resources to the VM in compliance with the agreed Service Level Agreement.
- In Unlabeled learning, we aim to develop a model from noticed instances that illuminates the inherent relationships within the dataset. Unlike supervised learning, we lack labeled data, and our aim is for the model to reveal appropriate labels for the data. The procedure mirrors supervised learning, albeit without explicit example outputs, as the model's inferred outputs serve as the new knowledge we seek. For instance, contemplate data retrieved from a network traffic log, wherein an Internet Service Provider (ISP) seeks to discern whether its users demonstrate a restricted range of behaviors or patterns. The internet service provider gathers instances of traffic and utilizes a clustering algo to construct a model that organizes users according to their traffic patterns, grouping them into k clusters, with each cluster containing users demonstrating analogous behavior. The ISP operator can then scrutinize the attributes of each cluster and, upon being content with the acquired insights, employ the model to classify new users by assigning them the label of the most akin cluster. Moreover, they can implement distinct policies for each cluster based on their respective attributes.

Algorithm and Approach

- There exists a society of algorithm available for modeling, prediction, and clustering, each with its own distinct advantages and limitations, ideally tailored for particular problem domains. Examples of algorithm include Tree algorithms such as CART, RPT [2], An alternative method involves utilizing metrics like information gain for each feature to pinpoint those that more effectively differentiate the examples. As a result, a decision tree is built using these feature values, assigning a target class to each leaf node. Moreover, these trees have the potential to be applied to regression tasks, where each leaf node is assigned either an average value for all examples or a regression. This leads to a regression analysis for a segmented linear function, as seen in the M5P algorithm.
- Alternatively, there are Instance-Based algorithms

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

involving the retention of examples and the comparison of new observations with those stored in memory, exemplified by k-Nearest Neighbors [3],

- Another approach involves storing the training examples within the model, whereupon the nearest k examples are utilized to determine its class through voting or to compute its expected value through averaging when a new observation is received. Additionally, Bayesian algorithms, such as Naïve-Bayes or Bayesian Networks, represent another set of techniques in this realm [4], Employing Bayes' theorem for probabilistic classification of a new instance into each class based on its features, and estimating the probabilities of each feature having a particular value for each class using existing instances. The decision between the Naïve and Network approaches hinges on evaluating the independence or interdependence of the features.
- Additionally, algorithms such as SVM construct classification models by mapping data into a hyper-dimensional space and identifying hyperplanes that effectively separate distinct classes [5]. There are also the ANN [6] Neural networks, drawing inspiration from the interconnected structure of brain neurons, can learn to classify, predict, or label data by organizing perceptrons into layers and interconnecting these layers to form networks. Additionally, clustering algorithms such as k-means aim to group data based on similarities and emphasize distinctions and commonalities among the identified groups.

In recent years, neural networks have undergone a renaissance, moved beyond mere image classification and expanding their utility to address various data types beyond images. This resurgence is attributed to advancements in neural network architectures and training methodologies, notably techniques such as DL, CRB machines, and RNN [7], which have emerged alongside the traditional Feed-Forward Neural Networks.

3. FRAMEWORKS USED FOR EXECUTION

Operators, managers, and data scientists often seek insights and expertise from vast datasets or extensive data flows in big data scenarios. To streamline this process and enable data scientists to concentrate on automation and results, numerous frameworks have emerged. Here is a summary of a selection of these frameworks, though not an exhaustive list.

Map-Reducing frameworks: Apache Hadoop and Spark

Through the establishment of individual computing processes for each input, followed by the consolidation of outputs to derive the answer, the majority of machine learning operations may be performed in parallel. A Map-Reduce approach may be used to tackle such problems; data is divided into smaller portions, each of which is calculated in parallel, and the results are combined to form the answer. A popular open-source Java framework for these uses is Apache Hadoop [8]. Users have two deployment choices: they can leverage their own clusters or data centers. Alternatively, they can opt to acquire a solution from companies offering Hadoop as a Platform as a Service, focusing exclusively on deploying applications.

Apache Spark [9], a Hadoop-based solution that is programmable in Python or Scala, focuses on specialized functions including ML, and data-streaming in an effort to boost speed. While Hadoop has been on the market for a while and now offers more features spanning a wider range of business processes, Spark develops by concentrating on and addressing a narrow range of issues, the majority of which are connected to big data processing and machine learning.

TensorFlow (Google)

One of Google's ML frameworks, TensorFlow(google) [10], built on operations and compatible with both standered hardware and GPUs devices, was recently released to the public. TensorFlow now available under the Apache license, can be acquired by users, deployed on their clusters of servers or computers featuring CPUs and GPUs, and employed in conjunction with Python scripting and its built-in libraries to designate the utilization of both CPUs and GPUs.

Azure-ML(Microsoft)

Azure-ML is a platform designed to data pipeline execution operations [11]. Users establish the data stream handled by specifying the data sources, the operations to be performed on the data, and the intended outcomes. Processes such as aggregation functions, ML algorithm, or data treatment functions may already be coded. Additionally, they are deployable on Azure services for computing and storage. Additionally, Azure-ML (Microsoft) has added native R programming capability to their platforms, enabling the user to integrate her own algorithms.

4. LIBRARIES AND PLATFORMS

Many of the algorithm mentioned above have readily available implementations across diverse tools, platforms, and libraries. Presented below are some of the well-known ones from the vast selection already in existence, accompanied by brief remarks.

Languages: R-cran and Python Sci-Kit

R, an extensively supported open-source platform, is a programming language primarily designed for statistical analysis and data exploration, accompanied by a rich array of libraries and plugging [12]. The core package is equipped with fundamental statistical operations, linear regression computations, and clustering methodologies. supplementary

packages such as kknn, e1071, nnet, and rpart, among others, provide a wide range of algorithms catering to both labeled and unlabeled learning tasks. Users can effortlessly access the R studio, libraries, and integrated development environments like RStudio. Moreover, the R language is compatible with Hadoop and Azure-ML , Empowering users to deploy and run R programs on cluster systems either as Map-Reduce applications or seamlessly integrate them into data pipelines within Azure environments.

The SciKit-Learn library [13] offers a variety of methods for classification, regression, and clustering for Python users who want to implement machine learning applications. The open-source library includes extensive documentation and was created in Python and Cython.

Prepared packages: Weka and MOA

Weka [14] a Java program developed by Waikato University, has been in operation for over a decade, catering to both novice and experienced machine learning practitioners. It encompasses a diverse array of algorithms catering to tasks such as classification, regression, clustering, and data preprocessing. Users of this package have two main options: they can either directly utilize its libraries and functions within their code or employ its graphical user interface (GUI) to conduct educational experiments, plot data, and save models. The same team's MOA [15] stream learning suite also includes utilities for continuous data processing. Weka and MOA are both freely available, extensively documented software packages.

Displaying Data: ElasticSearch + Logstash + Kibana

Before or after machine learning procedures, analysing and then showing massive data is sometimes just as crucial as developing models or generating predictions. The obtained results or recommendations hold little value if the most relevant discovered facts or calculated forecasts cannot be effectively communicated to the appropriate expert or end user to support their decision-making. Using Elastic search [16], a web search tool built on Apache License [17], a search engine capable of storing and indexing data effectively, is one technique to gather and index massive internet usage data. Users can set up associated programs like Logstash for log file processing and regular importation into Elastic search, and Kibana for presenting refreshed data via a web-Graphical user interface. Elastic search can be downloaded and deployed on a single computer or an whole cluster. Despite not incorporating machine learning, Elasticsearch excels at consolidating data and empowering users to uncover patterns and insights within their queries.

5. CONCLUSION

In conclusion, this review has explored the latest trending techniques and machine learning models utilized in big data analytics. By synthesizing current literature, it becomes evident that a multitude of advanced methodologies, including neural networks, deep learning, and ensemble techniques, are increasingly prevalent in extracting insights from vast datasets. These techniques not only enhance predictive accuracy but also facilitate deeper understanding and actionable insights. However, ongoing research and innovation in this rapidly evolving field are imperative to harness the full potential of big data analytics in various domains.

REFERENCES

[1] Hastie, T.; Tibshirani, R. & Friedman, J. (2001), The Elements of Statistical Learning, Springer New York Inc., New York, NY, USA.

[2] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). Classification and regression trees.Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

[3] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46 (3): 175–185.

[4] Russell, S.; Norvig, P. (2003). Artificial Intelligence: A Modern Approach (2nd edition). Prentice Hall.

[5] Cortes, C.; Vapnik, V. (1995). Support-vector networks. Machine Learning 20 (3): 273.

[6] Bishop, C.M. (1995) Neural Networks for Pattern Recognition, Oxford: Oxford University Press.

[7] LeCun, Y.; Bengio, Y. & Hinton G. (2015). Deep learning, Nature 521, no. 7553: 436-444.

[8] White, T. (2009). Hadoop: The Definitive Guide (1st edition). O'Reilly Media, Inc. Software available from https://hadoop.apache.org

[9] Zaharia, M.; Chowdhury, M.; Franklin, M.J.; Shenker, S. & Stoica, I. (2010). Spark: cluster computingwith working sets. In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing.

[10] USENIX Association, Berkeley, CA, USA. Software available from https://spark.apache.org

[11] Abadi, M. Et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from https://tensorflow.org

[12] Microsoft Azure Machine Learning. At https://studio.azureml.net , accessed in March 2016.

[13] R Development Core Team (2008). R: A language and environment for statistical computing. R foundationfor Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

[14] Pedregosa, F. et al. (2011) Scikit-learn: Machine Learning in Python, Journal on Machine Learning Research 12, pp. 2825-2830. Software available from

https://scikit-learn.org

[15] Hall, M; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I.H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. Software available from http://www.cs.waikato.ac.nz/ml/weka/

[16] Bifet, A.; Holmes, G.; Kirkby, R. & Pfahringer, B. (2010). MOA: Massive Online Analysis. Journal on Machine Learning Research 11, pp. 1601-1604. Software available from http://moa.cms.waikato.ac.nz

[17] ElasticSearch. Software available from https://www.elastic.co/products/elasticsearch

[18] McCandless, M.; Hatcher, E. & Gospodnetic, O. (2010). Lucene in Action, Second Edition: Covers Apache Lucene 3.0. Manning Publications Co., Greenwich, CT, USA. Software available from https://lucene.apache.org