IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANALYSIS ON EXTRACTING COMPLEX PATTERNS AND THE APPLICATION OF DEEP LEARNING ALGORITHMS AND ARCHITECTURES FOR BIG DATA ANALYTICS

Mairaj Mohsin Mohammed ,Student ,Civil Engineering ,Osmania University ,Hyderabad

ABSTRACT

In today's world, data science methods are widely applied in a variety of fields. Open source pre-print services provide scient metric research of abstracts for processing data on subject areas and directions, and this work focuses on developing new ways, in particular – the development of a unique model. The study's goal is to determine the extent to which data science is being used and how important it is in various disciplines of study. Deep Learning's ability to mine and learn from large amounts of unsupervised data makes it a valuable resource for Big Data Analytics. Researchers are looking into whether or whether Deep Learning can be utilized to address some of the most pressing Big Data Analytics issues, such as extracting complex patterns from vast amounts of data, semantic indexing, labeling, and rapid information retrieval. Streaming data, high-dimensional data, model scalability, and distributed computing are all introduced by Big Data Analytics, and they all need to be examined further in Deep Learning research.

Key words: Data science, Deep learning, Big data.

1. INTRODUCTION

The problem of capturing, analyzing, and storing data is one that affects every industry today [1]. Using Big Data has become critical to an organization's day-to-day health and to predicting weather, traffic, and natural calamities using satellite data Big data and data science have an impact on business and sales. A company's stock fluctuation can be predicted using big data, which could be useful for political organizations and financial institutions [3, 4]. Smart Farming is a way of farming that use sensors and gadgets to collect and analyze large amounts of data, allowing for unprecedented decision-making capabilities [5]. It draws more customers who are interested in new developments in the manufacturing and delivery of goods and services. We now have the ability to use driverless smart transportation, smart homes, the Internet of Things, and Cloud Computing, making life easier as data science advances [6]. Despite the growth of data, there are still a number of hurdles to overcome. These include data challenges (such as large volumes of inconsistent data), process challenges, and management challenges (including privacy, security, governance, and ethical issues) [7].

Deep Learning algorithms are one option for automating the extraction of complicated data representations (features) at high abstraction levels. Since these algorithms establish a hierarchical structure of layers for learning and describing information, they can express higher-level (abstract) qualities in terms of lower-level (less abstract) ones. As with Deep Learning algorithms, the neocortex of the human brain has a layered learning process that mimics the hierarchical learning architecture of the algorithms. Deep learning techniques are highly useful for dealing with massive amounts of unsupervised data because they learn data representations greedily layer-by-layer. Data representations formed via stacking non-linear feature extractors have been shown to improve classification modeling, improve the quality of generated samples using generative probabilistic models, and maintain the invariant property of data representations in empirical investigations, for example (such as Deep Learning).

Big Data encompasses a wide range of issues and strategies used in a wide range of application areas that generate and preserve large volumes of unstructured raw data for the sake of performing domain-specific analyses. Recent developments in data-intensive technology and increasing processing and storage capacity have greatly aided the growth of Big Data science. Exabytes of data have been gathered and kept by technology-based corporations including Google, Yahoo, Microsoft, and Amazon. There are billions of people who use social media sites like Facebook, YouTube, and Twitter on a daily basis, generating enormous amounts of data.

1.1 Overview

We currently have more data than ever before since big data is becoming well-known and a rapidly-moving subject of concentration in our present technological state and society. It's still a problem since we have to keep track of, access, manage, and deal with it. Big data is characterized by three features: Large amounts of data, data that cannot be classified as standard relational databases, and data that is rapidly generated, captured, and processed are all examples of big data. With regard to health care, science, engineering, finance, and business, big data is having a profound impact. Big data has a wide range of implications for our society, and both technical and nontechnical professionals will continue to pay attention to them. According to the amount of data available and the rate at which it is growing, it appears like we are living in a data-flood period. The rising processing power over the last two decades has resulted in an overwhelming volume of data streams. As the amount of available data grows, it's becoming increasingly difficult to manage large amounts of data.

It is possible to employ machine learning to handle the issues associated with big data, which has recently become popular. Deep learning algorithms can classify data and generate hierarchical layer abstractions, and the commercialization of machine learning frameworks has made it easier for academics to implement and deploy machine learning solutions fast for large data. Because of deep learning's relevant methodologies and algorithms for large data analytics, it has become an important part of machine learning. Machine learning and pattern recognition researchers are now working on deep learning. In a wide range of fields, including as speech recognition and computer vision as well as in natural language processing, it has had great success. Deep learning applied to big data will enable enterprises to generate amazing and significant results since large data gives huge opportunities and potential for change.

2. LITERATURE REVIEW

Equation Class Clustering and bottomup Lattice Traversal is a method presented by Zaki et al. ECLAT finds frequent itemsets by performing a depth-first search. It represents data vertically rather than horizontally like the Apriori algorithm does. As a result, the ECLAT algorithm is more effective and scalable when it comes to learning association rules. In comparison to the Apriori algorithm, this one is superior for small and medium datasets.

For those who are unfamiliar with Frequent Pattern Growth (FP-Growth), it's an approach to learning association rules based on the Han et al (FP-tree). There is no candidate generation with the FP-growth method [2], although there are frequent candidate itemsets generated while building rules with the Apriori algorithm [8]. This results in the successful strategy of 'divide and conquer,' which results in the production of a tree.

Sarker et al. [3] A rule-based machine learning technique was recently proposed in their previous publication to uncover intriguing non-redundant rules for providing real-world intelligent services. With the help of this technique, the duplication in associations can be efficiently identified, and a set of nonredundant association rules may be discovered. The top-down technique of this algorithm creates an association generation tree (AGT), from which the association rules can be extracted by traversing the tree. The ABC-RuleMiner approach is better to typical rule-based approaches in terms of rule generation and intelligent decisionmaking, especially in a context-aware smart computing environment that considers human or user choice.

The definitions of big data and deep learning presented by many authors have varied, but they have all been straightforward, consistent, and easy to understand. There are two types of big data, according to Hu et al. : big data research and big data structures (or data lakes). However, big data science is concerned with "techniques relating to the acquisition of large amounts data and the conditioning and evaluation of this data," whereas a large data framework is comprised of "software libraries and associated algorithms that enable the distributed processing and analysis of large data problems across clusters of computer units" [4].

The "big data problem" is a term coined to describe the relationship between big data and the 4Vs [5]. The term "big data" refers to data that is difficult to analyze and extract using normal data management tools and techniques. Large datasets comprised of enormous volumes of data, a variety of information, and a great diversity include big data. These datasets include both structured and unstructured data and come more quickly (in terms of velocity) than traditional datasets.

Artificial intelligence researchers are increasingly focusing on deep learning as a machine learning technique [6]. To "learn hierarchical features for the tasks of classification and pattern recognition," deep learning applies the supervised or unsupervised method.

As stated by [7], Deep learning's goal is to get machine learning one step closer to its initial vision of being able to think for itself. Medical imaging, botany, picture identification and food processing are just a few of the industries where deep learning has been put to use. It's also been used in research projects spanning a wide range of fields, such as computer science and engineering.

3. METHODOLOGY

Deep learning techniques rely on the automatic extraction of representations (also known as abstractions) from data. Massive amounts of unsupervised data are used by deep learning algorithms to automatically extract complex representations. There are various ways in which these algorithms are impacted by artificial intelligence, which tries to duplicate the ability of the human brain to observe and evaluate situations, learn from experience and make decisions in the face of high complexity in the human mind. Deep Learning algorithms are being developed in part to better replicate the hierarchical learning strategy of the human brain. Using models based on simple or shallow learning architectures, such as decision trees, support vector machines, and case-based reasoning, to extract relevant information from input corpora may be ineffective.

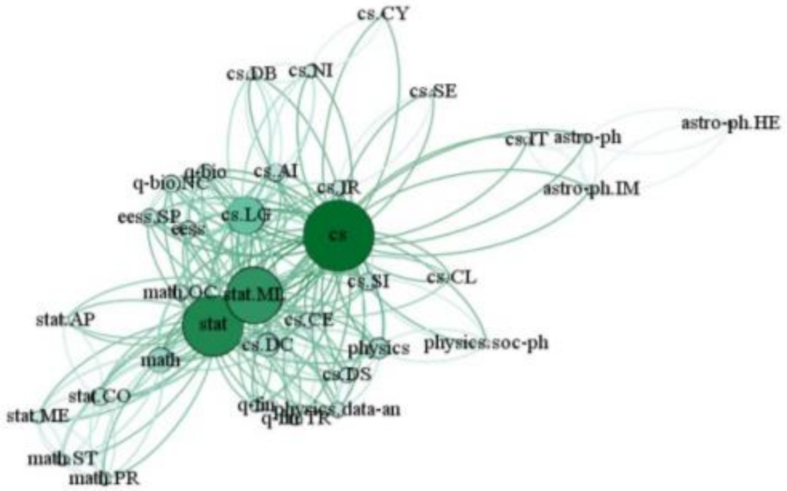

Fig. 1. Subject domain network for Big data concept.

Table 1 displays key network parameters. The concept "neural network" has the most nodes and the lowest density of all the networks offered. It demonstrates the method's broad applicability. The Gephi software was used to create the graphics (gephi.org). A few publications in mathematics (math) and physics deal with "big data," although most of the literature on the topic is focused on computer science and statistics (fig.1.).

When used with non-local and global generalization, Deep Learning architectures can create new, non-local learning patterns and connections across a wider area [4]. [5] In reality, deep learning is a critical step on the road to machine intelligence. Furthermore, it renders robots independent of human knowledge, which is the ultimate goal of artificial intelligence in addition to providing sophisticated data representations ideal for AI tasks. It uses unsupervised data to create representations directly without the need for human intervention.

TABLE I. MAIN PARAMETERS OF THE SUBJECT DOMAIN NETWORKS

Concept	Density of nodes	Number of nodes	Number of edges
Neural network	0.136	90	1090
Deep learning	0.159	54	454
Big data	0.213	35	254
Complex network	0.234	31	218

Data distribution, which allows for many different ways to represent abstract characteristics like labels and numeric values within the input data, is a crucial idea in Deep Learning methods. This makes each sample easier to describe, and it leads to better generalization. In direct proportion to the amount of abstract features extracted, there are an infinite number of configurations. When we consider that the observed data was formed as a result of the interactions between numerous known and unknown factors, we can infer that more (unknown) data patterns can be explained by new configurations of the learned factors and patterns. Using a distributed representation yields more patterns as the number of learned components increases, as opposed to learning from local generalizations.

3.1 Linear Discriminant Analysis (LDA)

Class conditional densities and Bayes' rule are used to produce a linear decision boundary classifier called linear discriminant analysis (LDA). 'Generalized Fisher's linear discriminant,' another term for this method, means It reduces model complexity or computing costs by projecting a given dataset onto a lower-dimensional space. A Gaussian density is used in the basic LDA model assuming that all classes have the same covariance matrix. ANOVA and regression analysis, both of which try to express one dependent variable as a linear mixture of other features or data, are strongly linked to LDA.

3.1.1External evaluation

To make sure we are on the correct route, we talked to the university library, PhD students who had already done SLRs in related disciplines, and a university professor. For example, we asked the university librarian for help and enlisted her to show us how to find relevant resources in various digital libraries and how to evaluate the accuracy of our search phrases. We narrowed the study's focus based on the comments we received and adjusted the review's scope, search technique, and inclusion and exclusion criteria.

3.1.2Conducting phase

We demonstrate how we identified our studies, selected the pilots, and selected them in this section to show how the SLR was carried out.

CONCLUSION

For the purpose of delivering scientometric research using open access preprint archives to isolate and process data related to publication domains and acceptable research objectives, this study focuses on developing new methodologies, such as new unique models. Analyzing data science's use in various research domains is the goal of this study. We demonstrated how data science may be integrated into a variety of fields of research utilizing terminology such as "big data," "deep learning," "neural networks," and "complex networks" from one of the largest open access archives. Deep Learning's lack of maturity need a great deal more research. Deep Learning models need to be scaled up, data abstractions need to be improved, distributed computing needs to be used, semantic indexing and data tagging need to be used, information retrieval needs to be improved, and domain adaptation needs to be done. More work must be done on these topics in particular. Future research should target one or more of these Big Data issues. This will add to the growing body of work in Deep Learning and Big Data Analytics theory and application.

REFERENCES

1. Zaki MJ. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng. 2000;12(3):372–90.
2. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: ACM Sigmod Record, ACM. 2000;29: 1–12.
3. **3.** Sarker IH, Kayes ASM. Abc-ruleminer: user behavioralrulebased machine learning method for context-aware intelligent services. J NetwComput Appl. 2020; page 102762
4. H Hu, Y Wen, TS Chua, et al. (2014) Toward scalable systems for big data analytics: A technology tutorial. IEEE Access 2: 652-687.
5. A Fernandez, S del Rio, V Lopez, et al. (2014) Big data with cloud computing: An insight on the computing environment, mapreduce, and programming frameworks. Wiley Interdiscip Rev Data Min KnowlDiscov 4: 380-409.
6. X Su, D Zhang, W Li, et al. (2016) A deep learning approach to android malware feature learning and detection. IEEE Trust 244-251.
7. CaglarGulcehre (2015) Welcome to deep learning.
8. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Fast algorithms for mining association rules. In: Proceedings of the International Joint Conference on Very Large Data Bases, Santiago Chile. 1994; 1215: 487–499.

9. Aha DW, Kibler D, Albert M. Instance-based learning algorithms. Mach Learn. 1991;6(1):37–66.
10. Alakus TB, Turkoglu I. Comparison of deep learning approaches to predict covid-19 infection. Chaos SolitFract. 2020;140:
11. Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Comput. 1997;9(7):1545–88.
12. Ankerst M, Breunig MM, Kriegel H-P, Sander J. Optics: ordering points to identify the clustering structure. ACM Sigmod Record. 1999;28(2):49–60.
13. Anzai Y. Pattern recognition and machine learning. Elsevier; 2012.
14. Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Algorithms. 2020;13(10):249.
15. Baldi P. Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning, 2012; 37–49 .
16. Balducci F, Impedovo D, Pirlo G. Machine learning applications on agricultural datasets for smart farm enhancement. Machines. 2018;6(3):38.
17. Boukerche A, Wang J. Machine learning-based trafc prediction models for intelligent transportation systems. ComputNetw. 2020;181