IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

A Review of a Comprehensive Investigation and Analysis of Big Data Using Data Mining Techniques

[1]**P Chandra Sekhar Reddy**, Research Scholar, Department of Computer Science Engineering, University of Technology, Jaipur

[2]**Dr. Suraj V Pote**, Professor, Department of Computer Science Engineering, University of Technology, Jaipur

Abstract: Contemporary data management systems search through enormous datasets to find patterns and correlations that were previously undiscovered in addition to storing and retrieving data. The need for computer applications and data mining software is rising as a result of how quickly new technologies are being developed. To ensure that all calculations lead to the same result, the necessary software and tools need to work with remote databases. However, because of legal restrictions and the necessity for a competitive edge, distributed data mining presents privacy concerns. Experts in the fields of big data, cyber security, and data mining are therefore motivated to study more.

Researchers created Privacy-preserving Distributed Data Mining (PPDDM) to address the multi-party computation problem, in which multiple users attempt to perform a data mining task cooperatively using their respective private data sets, in order to get around these limitations and benefit from these advantages. Participants discover only the outcomes of the data mining algorithm and their own inputs after finishing the exercise. The main objective behind this research was to provide a novel Privacy Preserving Data Mining approach for creating Decision Tree Classifiers utilizing data that has been vertically partitioned. The recommended PPDM approach is utilized towards build a conclusion tree classifier in Weka, and the results are compared to those from the well-established J48 method. This analysis employs accuracy and precision as its standards. Compared to the conventional approach, the suggested PPDM algorithm offers far greater accuracy and precision.

Keywords: Big Data, Cyber Security, PPDM, PPDDM, Etc.

I. Introduction

Data analytics is the process of using unstructured facts to derive statistical insights that can be used to make informed commercial and personal decisions. Although most datasets are structured (like a CSV file) or semi-structured (like a relational database), examining unstructured data (like emails or tweets) has gained interest recently.

The following stages make up the data analytics process:

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

1. Extracting and gathering data

2. Preprocessing the data (cleaning, organizing, and mining)
3. Model development and visualization
4. Reporting and model evaluation

The gathering of data is the initial step in the data analytics process. This stage could consist of data warehousing, which includes gathering datasets from many sources, transforming them, and storing them in a repository. Using tools like import.io, webscraper.io, Beautiful Soup (a Python package), and others, data is retrieved from the web through processes like web scraping that may be included in this step.

The second stage involves prepping the data, which usually involves cleaning the data (e.g., managing missing numbers). During this stage, data wrangling techniques can also be used to alter the data, such as parsing it into a pre-defined data structure or using an algorithm like Principal Component Analysis to minimize its dimensions.

The real analytics are performed in the third phase using an appropriate machine learning technique. Using the training dataset, this phase creates a regression or classification model that can be used to a test dataset. Clustering or pattern mining may also be carried out during this phase. Rather than using a machine learning model, the data may simply be displayed to find hidden patterns.

The assessment and reporting stage of data analytics is the last one. We may need to assess whether machine learning method yields a more accurate answer for the given dataset because there are several algorithms available for the same task. The models can be compared using a variety of evaluation measures, such as the confusion matrix, specificity, sensitivity, and precision. The findings could be presented as charts, tables, or in another format.

Between 2014 and 2017, the number of people using the internet climbed from 2.4 billion to 3.8 billion. Just a few examples of the millions of images shared on Instagram and Facebook every minute, the over 656 million tweets sent daily, and the five billion Facebook "likes" posted daily are all noteworthy. Although this data does not fit into the typical organized or semi-structured data categories utilized by data analytics tools, it nonetheless has important information that, if employed, can greatly benefit enterprises. The focus of data analytics has moved to what is known as "big data analytics," which deals with extracting information from large data.

II. Challenges in Big Data Analytics

Three layers, each with a unique set of issues, make up the conceptual framework for big data analytics, according to Wu et al. (2014). Tier I addresses big data mining platforms, Tier II addresses big data semantics, and Tier III addresses big data mining methods. For the purpose of creating an effective big data processing system, the difficulties at each of the three stages must be overcome.

The main problem at Tier-I is processing data that is bigger than the RAM that a workstation can hold. The system's RAM capacity determines how much processing power the data mining software can do. Big data processing frameworks typically use parallel/cluster computing to address this, and parallel programming paradigms like Enterprise Control Language (ECL) and Message Passing Interface (MPI) can be used. Prior to creating a big data system, obstacles specific to parallel programming must be addressed.

Domain expertise and data privacy are the primary concerns at Tier-II. Developing an effective big data algorithm requires domain expertise. It aids the developer in selecting the appropriate characteristics for data modeling. Data privacy is yet another difficult big data topic. Two methods are commonly used to protect privacy: first, limiting access to the data so that only authorized individuals can see sensitive information, and second, anonymizing the data fields. It is possible to create a secured certification or access control system such that Unauthorized individuals cannot access sensitive information. Tier-III of the framework can be divided into three levels: mining of heterogeneous, incomplete, sparse, uncertain data at level two; mining of dynamic and complicated datasets at level three; and fusion of data from several sources at level one. Global optimization should be accomplished by combining datasets from many sources in a large data mining system. The machine learning model's reliability is negatively impacted by high-dimensional sparse datasets; in these situations, dimension reduction techniques are typically used. While there are other methods for handling incomplete data values, such as data imputation, the parameters of the probability distribution that produced the data are estimated when dealing with uncertain data.

III. Research Motivation

The reason for this work is that big data is a new field in data analytics that provides previously unattainable significant dataset insights. Big Data analytics is used in numerous industries, including sports, healthcare, and many more. Using tennis as an example, every shot a player hits is recorded on servers and examined for speed, racquet position, and other

variables to produce interesting statistics. Our bodies themselves are also large data sources; all we have to do is record the information they provide, such as body temperature and heart rate. If this information is collected, it can be used for study as well as to monitor our health. Thus, the field of large Data analytics has enormous potential and may be used in many ways in almost any domain.

Working with enormous data sets presents a number of issues, handling data size being only one of them. The needs of the problem determine which combination of big data tools should be employed, as there are many of them on the market, but each one has a unique set of features. These technologies can be used to quickly and efficiently analyze any kind of data in real-time (if necessary) in order to extract important information. For tasks like market segmentation and predictive analytics, typical machine learning algorithms can be transferred to big data tools. However, in the context of big data, a new method could occasionally be required to gain fresh insights from the data.

Big data is currently being used by all firms to increase their competitiveness; those who do not will undoubtedly fall behind in the marketplace. Big data insights are being used by not just commercial businesses but also non-commercial organizations (such as political parties during election campaigns) to their advantage.

IV. Problem Statement

Although there are many big data processing tools accessible, there aren't many machine learning libraries for them; that is, there aren't many standard machine learning algorithms in these libraries. These tools are limited in their ability to address all facets of the vast subject of data analytics. As a result, there is a significant discrepancy between the needed and accessible big data methods. This gap can be closed by determining which analytics domains lack enough big data algorithms, and then developing new big data-friendly algorithms and libraries specifically for those domains.

Finding data points in a dataset that deviate from typical features is the focus of the data analytics field known as anomaly detection. While many methods exist for identifying patterns in datasets, very few tools or libraries are capable of detecting patterns in large amounts of data. Thus, the topic of this thesis is anomaly detection in large datasets. In addition to being scalable, a big data solution for anomaly detection needs to be quick and extremely precise.

V. Objectives of Research

The following are the goals of this study:

- To study literature related to big data.
- To discover tools/architecture for big data analytics.
- To suggest cutting-edge algorithms for large-scale data anomaly detection.
- To put the suggested algorithm(s) into practice and test them.

VI. Conclusion and Future Scope

These days, scalable methods for large data analytics are the main emphasis of data mining. Tools like MapReduce, Apache Spark, and others should be used to create such scalable systems. Anomaly detection is one area of data mining where effective solutions are few. Because there are fewer anomalous points in a large dataset than there are in a classification or clustering operation, finding anomalies in huge data is more difficult. New techniques for big data anomaly detection have been proposed in this paper.

The density-based technique, in which the isolated data points are considered anomalies, is the most often used method for anomaly detection. Some parameters for these algorithms must be set, and using less-than-ideal values yields subpar outcomes. The performance of the existing solutions is very low because they don't offer any techniques for determining the ideal values. In order to handle large data, an optimized density-based method was presented that not only optimizes the parameters but also other aspects. The settings were optimized using swarm intelligence methods, which raises the algorithm's accuracy. Large datasets may be handled by the technique thanks to its implementation on Apache Spark, which offers distributed processing and scalability. The speed is also influenced by the distance function that is used to compute the neighbours

In order to determine the fastest distance function, the distance matrix was noted. As a result, the suggested method was quick (using the right distance function), scalable (using Spark to implement the parameters), and accurate (thanks to swarm intelligence).

Out of all the density-based techniques in the literature, the suggested solution is the most effective one for anomaly detection. But it has the same problem as all density-based solutions, which is that these algorithms are all O(n2) solutions because it has to calculate the

distance between each point and the remaining points. For a big data solution to be effective, its complexity should be Ω(n).

For anomaly detection, replicating neural networks (RNNs) provide O(n) solutions. When used with a GPU-supporting program such as TensorFlow, RNNs can provide a very quick fix for large-scale anomaly detection in data. The number of neurons in the hidden layer will, however, also influence the RNN's speed; yet, fewer hidden layers may also have an impact on performance. Instead of using gradient-descent learning, as in typical RNNs, for learning, Extreme Learning Machines (ELM) can be utilized to solve this issue. The issue of counting hidden layers is resolved by the fact that ELM has just one hidden layer. Therefore, although the learning mechanism differs, the architecture of ELM and single-layer feed-forward neural networks (SLFN) is the same. There are two sets of weights in SLFN that are to One set lies between the hidden layer and the input layer, while the second set lies between the hidden layer and the output layer. The same two sets of weights are used in ELM as well, but only the set between the hidden and output layers needs to be learned; the set between the input and hidden layers is fixed. Because both the hidden layer neurons' and the output neurons' outputs are fixed, the task reduces to simple matrix computation, or a non-iterative task as opposed to an iterative one like in SLFN. In contrast to SLFN, ELM learning is extremely quick because it is not iterative. As a result, ELM-RNN on GPU was suggested as a very quick anomaly detection approach. The quantity of neurons in the brain was the only issue that persisted the hidden layer, which has a major impact on the algorithm's accuracy. Garson's neural network pruning approach was applied to address this issue. Each hidden layer node's relative importance (RI) should be determined in accordance with Garson's algorithm, and any nodes with RI below a certain level can be eliminated. In terms of accuracy and speed, the suggested Garson-pruned ELM-RNN solution beats all other options when compared to state-of-the-art solutions.

Based on the current work, a great deal of future work can be done. Novel solutions can be proposed using methods other than density-based and neural networks. Although the density-based strategy is quite slow, its pace can be increased by utilizing various tactics. In terms of accuracy, ensemble learning is a viable strategy, but speed may have to be compromised. Inliers may also be down-sampled as a potential supervised learning strategy. The present study focuses solely on the anomaly detection field of big data mining, however there are many other fields of big data analytics where unique and efficient solutions are needed.

Recently, deep learning techniques have also been used to detect anomalies, and they may a fascinating field of research.

REFERENCE

[1]. Y.L. Teo et al.Techno-economic-environmental analysis of solar/hybrid/storage for vertical farming system: A case study, Malaysia Renewable Energy Focus (2021)
[2]. M.N. Halgamuge et al. Internet of Things and Autonomous Control for Vertical Cultivation Walls Towards Smart Food Growing: A Review Urban For. Urban Greening (2021)
[3]. G.W. Michael F.S. Tay Y.L. Then 1844 1 2021 012024...Jandl IoT and Edge Computing Technologies for Vertical Farming from Seed to Harvesting (Doctoral dissertation 2021...
[4]. S. Uphale et al.Hydroponics as an Advanced Technique for Vegetable Farming International Journal of Recent Advances in Multidisciplinary Topics (2021)
[5]. N.N. Saxena The Review on Techniques of Vertical Farming International Journal of Modern Agriculture (2021)
[6]. Z. Zhang et al. A Comprehensive Review on Sustainable Industrial Vertical Farming Using Film Farming Technology Sustainable Agriculture Research (2021)
[7]. Zohra, F. T. (2020). Design and Development of an Integrated Water System Combining Rainwater Harvesting System (RHS)...
[8]. A Mehta Study of Vertical Farming (Hydroponics) International Journal of Environmental Planning and Development(2020).
[9]. A Chatterjee et al. Implication of Urban Agriculture and Vertical Farming for Future Sustainability (2020)