Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

A Survey of the Computational Intelligence Techniques for Big Data

Shanthi P. , Dr. Sangeetha M.

Research Scholar, Department of Computer Science

Dr. NGP Arts and Science College Coimbatore, India.

Abstract- The term "Big Data" refers to the unprecedented growth in volume, diversity, and velocity of today's data. Despite the complexity and dynamic nature of modern data, handling massive amounts of data from various sources efficiently is a difficult undertaking. The present data will not be handled by traditional ways. Computational intelligence is used because a methodical strategy to handling current data must adapt to and learn from the changing environment. It is an area of artificial intelligence (AI) has applications in the development of intelligent computer systems. Issues and challenges across a variety of AI applications, including Natural Language Processing, Image and Video Processing, Robotics and Automation, Decision Support Systems, Bioinformatics, and Medicine, can be addressed with the help of CI approaches and techniques. Predictive analytics, real-time processing, pattern recognition, and scalability are benefits of CI. The most common technologies are Evolutionary Algorithms, Neural Networks, Fuzzy Systems, Swarm Intelligence, Expert Systems, Machine Learning, Probabilistic Reasoning and Hybrid Systems. The main methods of computational intelligence are explored in this study, along with the real-time Big Data Mining uses for them.

Keywords*-Big Data, Computational Intelligence, Applications of Computational Intelligence.*

I. Introduction

The term "big data" describes datasets wider than what can be processed and analyzed in an appropriate period of time. It incorporates structured, semi-structured, and unstructured data in a broader volume. Big data can be characterized by the five Vs: Volume, Variety, Velocity, Value, and Veracity. Volume implies massive quantities that are exceeding the limits of conventional processing, commonly expressed as terabytes and petabytes. Variety encompasses a wide range of data types and sources, including text, images, audio, video, social media and sensors. The production and handling of data in real-time are highlighted by velocity. Value reinforces how essential it is that data contain relevant information that may be used to enhance science or industry.

The significance of trustworthy data sources generating high-quality information is highlighted by veracity. The analysis of massive amounts of information creates a number of difficulties for traditional data mining approaches, which has led to a surge in the implementation of computational intelligence techniques.

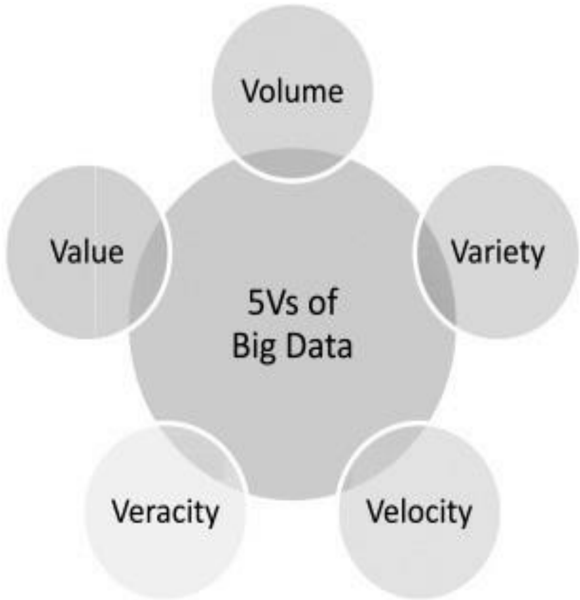

Fig 1. 5V's of Big Data

Ii. Challenges In Big Data

A. Complexity and Dimensionality

Complexity and increased dimensionality are typical traits of large data. Big data's multifarious nature, which includes enormous quantities, varied types, and dynamic velocities of data, is intrinsically linked to its complexity. This difficulty is most noticeable in big data's dimensionality, when an analysis's complexity is increased by the sheer volume of variables or features. Managing and deriving meaningful insights from the wide variety of structured, semi-structured, and unstructured data sources that are available is a challenge. When one takes into account how data changes over time, the dimensionality increases and calls for scalable infrastructure and adaptive solutions.

B. Scalability

The scalability of Big Data is its capacity to effectively manage and process ever-increasing amounts of data as the dataset expands. Scalability is essential for big data systems to maintain maximum performance and responsiveness as processing needs and data sizes increase. Scalability can be attained by either vertical or horizontal scaling. In order to spread the workload across several workstations, a distributed system can be made more horizontally flexible by adding additional hardware or nodes. On the other side, vertical scaling entails boosting the capability of already-existing hardware resources, like boosting CPU, memory, or storage. Big data solutions that are effectively scalable allow enterprises to handle and analyze enormous datasets without sacrificing performance, guaranteeing that the infrastructure can keep up with the expanding demands of needs for data processing and storage. Given the dynamic and ever-increasing quality of data in today's constantly shifting technological landscape, this ability is imperative. The vast amount of data in big data scenarios could make typical data mining algorithms inappropriate. Conventional methods find it more difficult to manage a dataset of this size since processing requires exponentially more computer resources the larger the dataset.

C. Variety of Data

The various forms and categories of data that comprise the information creating entire dataset are referred to as variety. Big data includes a variety of data kinds, such as structured, semi-structured, and unstructured data, in contrast to traditional data sources which usually comprise structured data in well-defined formats. Semi-structured data, such as JSON or XML, lacks a tight standard but still has some level of organization, whereas structured data, which is frequently found in relational databases, is ordered and adheres to a predetermined model.

On the other hand, unstructured data, which includes written documents, photos, videos, and social media posts, completely lacks a predetermined framework. Sophisticated tools and strategies, such as data integration, transformation, and processing methods that can handle the inherent complexity of many data types, are needed to manage and analyze this diversified data. Typical techniques for data mining might not be effective additionally with unstructured data, such as text, photos, or videos, since they tend to be designed for structured data.

D. Real-time processing

A key feature of Big Data is real-time processing, which meets the requirement for rapid evaluation and response to streaming data. Large volumes of data are instantly analyzed and interpreted as they are generated in real-time big data processing, enabling businesses to take prompt, well-informed decisions. Real-time processing is intended to manage continuous data streams and deliver insights in almost real-time, in contrast to traditional batch processing, which stores and processes data at predetermined intervals.

This skill is especially important in attempts to quickly extract significant details from massive data, such social media trend analysis, IoT device monitoring, and financial transactions. Many traditional data mining techniques aren't meant to be used in real time. Real-time processing in the realm of big data analytics implies managing data continuously as it is generated, with minimal delay, empowering organizations immediately glean valuable knowledge and make decisions. In sectors like e-commerce, healthcare, and transportation where prompt actions are critical, including real-time route optimization and personalized advice, this capability is especially important. Frameworks for real-time processing enable enterprises to handle the high-velocity data produced by multiple sources.

E. Imbalanced Data

Big data accumulations frequently contain imbalanced data, which is characterized by an unequal distribution of classes or categories, with some being noticeably more common than others. Traditional mining and analytics algorithms may encounter difficulties as a result of this mismatch, especially those that are intended to categorise or forecast results. Algorithms that display biases towards the dominant class due to uneven data may perform less well than ideal when it comes to seeing patterns or producing precise forecasts for minority classes. More importantly, to make sure that the model's performance is not only determined by the dominant class, sophisticated algorithms and model assessment metrics becomes crucial. In big data analytics, handling unbalanced data is essential to building trustworthy and objective models, which guarantee that conclusions and forecasts appropriately capture the intricacies seen in real-world situations.

F. Computational Resources

Big data presents significant hurdles for traditional data mining, mostly because of the size and complexity of the information involved. The scalability of traditional algorithms presents a significant challenge because they were not initially intended to manage the enormous amounts of data present in big data environments. When there is a rapid stream of data, traditional approaches may not be able to match the needs for real-time or near-real-time analytics, which makes processing speed a major barrier. Large-scale data handling inefficiencies can be made worse by these techniques, which may also be memory-intensive and lack optimization for distributed or in-memory processing. Conventional data mining methodologies encounter additional challenges due to the dynamic nature of big data environments and the requirement for adaptation to changing data distributions.

One important factor to take into account is the expense of updating computational infrastructure to cope with big data analytics. Conventional data mining approaches might not be able to fully utilize contemporary distributed and parallel computing platforms.

G. Adaptability and Self-Learning

In the field of big data analytics, adaptability and self-learning are essential qualities that enable systems to change and perform better in response to shifting data dynamics. When it comes to adaptability in the context of big data, typically mean a system's ability to survive in dynamic data environments by responding to changes in data volume, velocity, and variety. On the other side, self-learning refers to an algorithm's or system's capacity to learn from patterns and experiences found in the data over time, without the need for explicit programming. Methods often require human adaptation and periodic refinement to meet shifting circumstances. Big data analytics solutions that are adaptable may scale smoothly, meeting the demands of fluctuating data and changing workloads. This quality is especially important for real-time analytics since the volume and speed of incoming data might fluctuate quickly. Better decision-making is made possible by the system's flexibility and self-learning loop, which help it get better at deriving insightful conclusions from a variety of challenging and complicated datasets.

III. COMPUTATIONAL INTELLIGENCE

The term "computational intelligence" (CI) denotes a set of dynamic computational techniques and strategies which enable machines to think, acquire knowledge, and resolve intricate issues in a manner that is similar to that of human intelligence. In order to solve complicated issues that are either unsolvable or extremely difficult for traditional computing methods, it attempts to imitate natural processes. Without the need for explicit programming, CI is a generic term for a wide range of methods and algorithms that facilitate intelligent learning, decision-making, and problem-solving. There are many different fields and industries that employ computational intelligence. It drives e-commerce recommendation systems, enables self-driving vehicles, facilitates with medical diagnostics, streamlines supply chain operations, supports financial forecasts, and much more. It has practically infinite applications. Businesses and organizations may achieve a competitive edge and produce superior outcomes by exploiting computational intelligence's capabilities.

Key Components of Computational intelligence include

A. Evolutionary Algorithms

Natural selection is the conceptual framework used by genetic algorithms, genetic programming, and numerous other Evolutionary Algorithms (EAs). In order to determine the optimum or nearly optimal responses to challenging situations, they incorporate iteratively constructing practical alternatives in multiple eras. The objective of evolutionary algorithms is to find the best course of action to accomplish the intended results by a process of natural selection comparable to that of Darwinism.

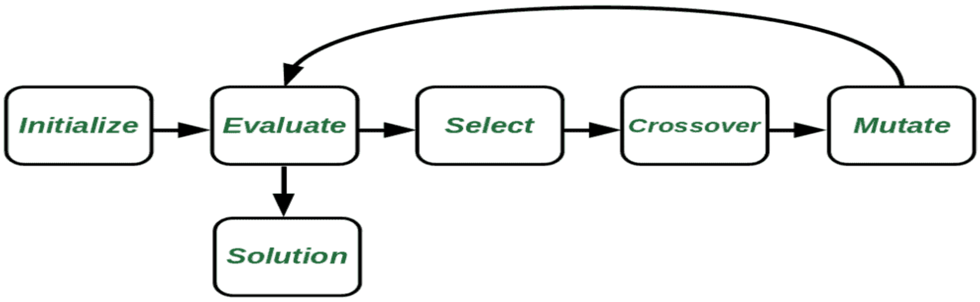

Fig 2. Overview of Evolutionary Algorithm

The weakest solutions are eliminated, and stronger, more viable prospects are preserved and revisited in subsequent evolution. There are numerous advantages that evolutionary algorithms (EAs) offers in the field of optimization. The option to do global optimization is one of their main advantages, which renders them appropriate for intricate search spaces with several optimal responses. Their resilience diminishes sensitivity to random fluctuations by enabling them to efficiently negotiate noisy or unclear objective functions. EAs parallel architecture makes computing more efficient and allows for the simultaneous evaluation of population members, which speeds up optimization procedures. Their capacity to put their expertise to a wide range of problems without necessitating in-depth knowledge of certain fields demonstrates their adaptability.

EAs are competent at solving constrained optimization problems using a variety of strategies. Moreover, they are appropriate in situations where optimization landscapes change over time because to their flexibility in dynamic environments. Effective exploration and exploitation of the search space is how EAs move across it. They also have the advantage of not being dependent on derivative information in scenarios when gradients are difficult to get. Evolutionary algorithms are based on biological principles and are inspired by nature. They offer a simple framework for resolving practical issues.

B. Neural Networks

Computational models patterned after the architecture and functions of the human brain are called Neural Networks. Since they are composed of interconnected nodes, or neurons, arranged in layers, they are able to discern complex patterns and correlations within the data. Multiple hidden layer topologies are used by deep learning, a subtype of neural networks.

An input layer, one or more hidden layers, and an output layer constitute the layers of nodes, or artificial neurons that collectively make up every neural network. Every node has a weight and threshold associated with it and is linked to other nodes. A node is activated and sends data to the subsequent layer of the network if its output surpasses the specified threshold value. If not, no data is transferred to the network's successive levels. Training data is vital for neural networks to learn and acquire more precise over time. They are effective tools in computer science and artificial intelligence that empowers to quickly classify and cluster

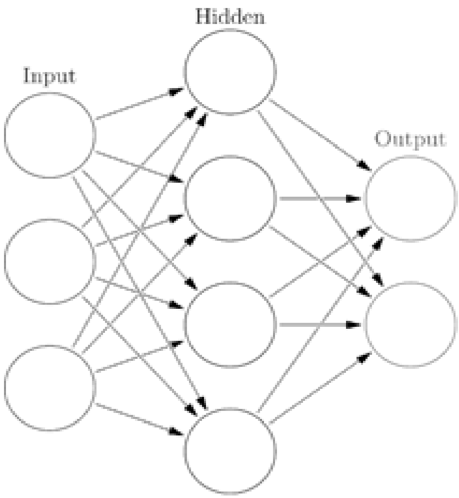

Fig.3 Neural Networks

data once they have been refined for accuracy. Comparing speech and image recognition tasks to manual identification by human experts, the latter can take hours, while the former can be accomplished in minutes. A widely recognized example of a Neural Network is the search algorithm used by Google. Artificial neural networks (ANNs) and simulated neural networks (SNNs) are other acronyms for neural networks. They are a cornerstone of deep learning models and a subset of machine learning.

C. Fuzzy Systems

The principles of fuzzy logic are extremely adjacent to the essence of human thought. Relying on how the programmer has configured the system for data mining, fuzzy logic may use approximation to forecast the next piece of data or it may be intuitive. The main cause why fuzzy logic is employed in data mining is that it enables programmers to input

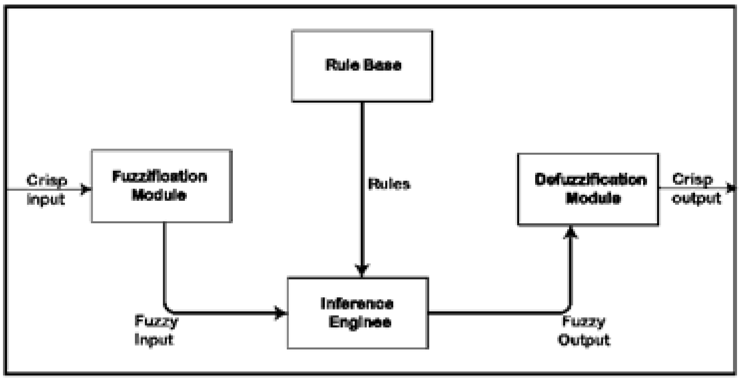

Fig 4. Fuzzy Systems

values without needing to know the precise amount. This is totally distinct from conventional reasoning. Due to its ability to account for degrees of truth, fuzzy logic addresses imprecision and uncertainty. Machines can manage partial or subjective data by using fuzzy systems, which are utilized for decision-making in scenarios when information is uncertain. Either "yes" or "no" is the return value in traditional logic. The return produced by technical processing involving a computer or programming design is either digital "1" or "0". For instance, a Programme can choose between two speeds: fast and slow. In conventional logic, there is no such thing as too fast or too slow. There are clusters for the speeds in fuzzy logic. Fuzzy logic distinguishes between four speeds: too quick, fast, medium speed, slow, and too slow.

D. Swarm Intelligence

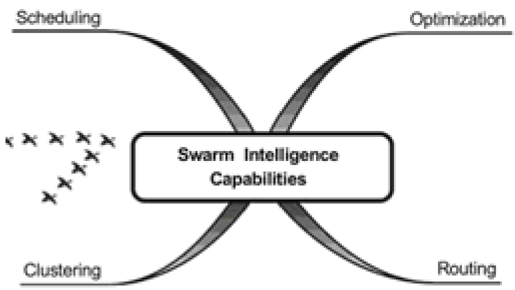

Fig 5. Swarm Intelligence

The collective activity of social insects like termites, ants, and bees served as the model for the concept of Swarm Intelligence (SI). In artificial systems, especially in the realm of artificial intelligence, the goal is to resemble the decentralized, self-organized structure of these natural systems. SI, the "bio-inspired computation method," was created by analyzing the intricate interactions between individuals operating in unsupervised groups to understand how swarms behave collectively in the natural world. SI is an extensively utilized and potent instrument for optimization tasks; algorithms based on SI are exceptionally flexible and may be applied to nearly any scientific field. They are also particularly effective at addressing tricky problems for which there is no viable classical solution. When using SI techniques, a fixed-size population of individuals is investigated across generations, and without conducting any individual selection operations, the findings of each generation's search are evaluated to adapt the search technique for subsequent generation;

E. Expert Systems

Expert Systems (ESs) replicate a human expert's method for making choices in a particular domain by employing knowledge representations and inference engines. These rule-based systems are very effective for activities

Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

requiring a lot of knowledge.

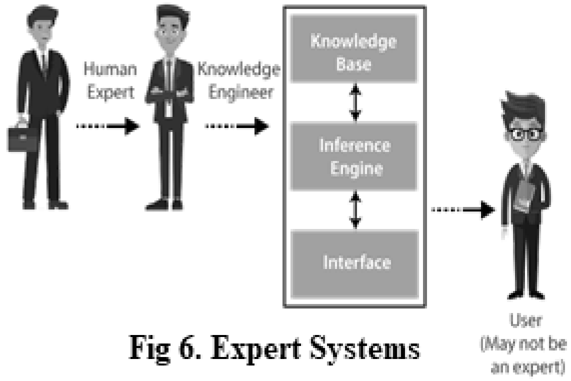

Fig 6. Expert Systems

ESs are a way to disseminate knowledge that has been obtained, either directly or indirectly, from subject matter experts in various scientific fields. They assist users who lack specialized knowledge as well as serve as backups, if not explicit substitutes for human experts. Consequently, an ES's effectiveness ought to emanate from its capacity to replicate the human decision-making process. Despite the broad spectrum of application areas that an expert system could potentially be used in, each model has distinctive characteristics that are compatible with the problem that the system is intended to address.

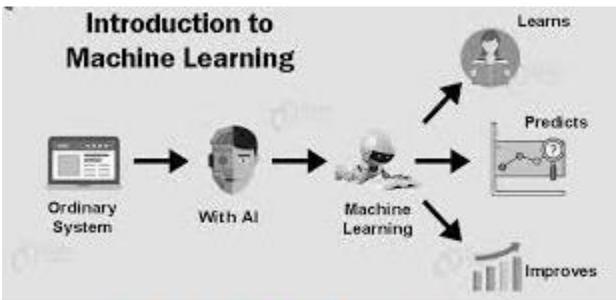

Fig 7. Machine Learning

F. Machine Learning

In the realm of computer science, one of the most significant subcategories of artificial intelligence is machine learning. It is defined as the study of automatically upgrading decision-making or automated data processing algorithms based on prior experience. It empowers systems to automatically learn from their interactions and get superior over time without necessitating explicit programming. Designing computer programs that can access data and use it for learning is the primary objective of a machine learning model. Machine learning is a more general term that covers a range of CI methods. It entails creating algorithms that, without explicit programming, assist computers to recognize patterns in data. In machine learning, supervised learning, unsupervised learning, and reinforcement learning are popular forms. Machine learning tools examine data sets, discover patterns, and then extrapolate inferences or predictions from those patterns utilizing data-driven algorithms and statistical models. Rather than implementing explicit instructions like traditional rules-based analytics systems, the algorithms learn from the data as they interact against it.

G. Probabilistic reasoning

The application of probability theory to describe uncertainty and base decisions on probabilities is known as probabilistic reasoning. Reasoning about probabilities or likelihoods of occurrences or outcomes is the essential component of probabilistic reasoning. It provides systems to consider advantage of the inherent uncertainty in real-world data and evidence and to model and reason about unreliable situations in a systematic and quantitative manner. A mathematical framework for resolving uncertainty and arriving to rational assessments in unpredictable situations is offered by probabilistic reasoning. Need of probabilistic reasoning is when there are unpredictable outcomes, when specifications or possibilities of predicates becomes too large to handle and when an unknown error occurs during an experiment. In probabilistic reasoning, there are two ways to solve problems with uncertain knowledge: Bayes' rule and Bayesian Statistics.

H. Hybrid Systems

A sophisticated and versatile method of structuring and deriving profound knowledge from massive databases is conveyed by hybrid systems which employ computational intelligence for big data. These systems optimize data analysis accuracy and efficiency by integrating the powers of computational intelligence approaches, such machine learning algorithms, with conventional big data processing frameworks. Neural networks, evolutionary algorithms, fuzzy logic, and other machine learning algorithms are incorporated into data processing pipelines to facilitate adaptive learning, pattern identification, and decision-making. The system can self-adapt constantly boost its performance by leveraging the incorporation of big data and computational intelligence technological advances. Hybrid systems blend the most beneficial aspects of both paradigms: the application of computational intelligence's cognitive abilities for intricate pattern recognition and prediction tasks, and big data frameworks' processing capacity for scalability. By virtue of this synergy, organizations are more equipped to cope with the obstacles caused by the dynamic and voluminous nature of big data, make knowledgeable choices, and find pertinent insights which culminate in more intelligent and effective data-driven solutions. This can improve the robustness and overall performance of intelligent systems.

Iv. Computational Intelligence For Big Data

As big data environments are dynamic, the properties of the data may alter over time. The intrinsic complexity, diversity, and vastness of modern datasets render computational intelligence essential for managing enormous quantities of data. Regarding the field of big data analytics, conventional methods frequently encounter difficulties in effectively handling and deriving significant insights from huge amounts of data. A range of adaptive and self-learning abilities are elevated by computational intelligence, encompassing machine learning, artificial intelligence, and other complex algorithms. These methods are particularly effective at spotting patterns, trends, and correlations in large datasets that could be too complex or dynamic for traditional approaches to comprehend entirely. Moreover, predictive modelling and decision-making made achievable by computational intelligence facilitates businesses to predict future patterns based on historical data. These systems are essential to extracting substantial information from big data owing to their capacity to modify and learn from dynamic information environments. This knowledge eventually assists with making intelligent choices, recognizing patterns, and revealing hidden relationships within the extensive and convoluted web of contemporary datasets. As a consequence, computational intelligence is crucial for reaping the most of big data since it supplies the methods and instruments required to identify meaningful patterns in a constantly changing and data-rich environment.

Significance of CI for big data

A. Machine Learning Algorithms

CI by employing machine learning algorithms to analyze, recognize patterns, and forecast outcomes is essential to fully exploiting the potential of large datasets. These algorithms are scalable and effective. Traditional analytical techniques may not perform well in the context of big data analytics, since datasets are frequently enormous, diversified, and changing quickly. With the aid of supervised and unsupervised learning techniques, machine learning algorithms can automatically spot patterns and trends in large datasets, enabling insights that might not be feasible with manual examination. While clustering algorithms like k-means and hierarchical clustering aid in revealing latent structures within the data, algorithms for classification and regression such as decision trees, random forests, support vector machines, and neural networks perform exceptionally well in these tasks. Machine learning algorithms enable enterprises to fully utilize big data across a range of industries, including finance, healthcare, and business intelligence, by providing actionable insights and enabling data-driven decision-making.

B. Deep Learning for Feature Extraction

In the field of big data analytics, Deep Learning has become a potent method for feature extraction because of its unmatched capacity to extract detailed patterns and representations. Automatically learning hierarchical features from raw data is a strong suit for deep learning architectures, especially deep Neural Networks (Deep Neural Networks with several layers). This skill comes in situations when the sheer volume and diversity of big data might make traditional feature engineering difficult. Through the iterative learning process, deep learning models like Recurrent Neural Networks (RNNs) for sequential data or convolutional neural networks (CNNs) for visual data can automatically uncover significant features and representations, obviating the requirement for manual feature extraction. These networks' depth and complexity allow them to detect complex associations and intangible characteristics that can be crucial to precise analysis and forecasting. Big data analytics has made enormous progress attributable to deep learning's ability to automatically extract key traits. This has heightened the accuracy and efficiency of tasks including image recognition, natural language processing, and other complicated data-driven applications.

C. Clustering and Segmentation

Techniques for grouping and segmenting data are essential parts of big data analytics because they reveal important patterns and structures in vast, heterogeneous datasets. While segmentation entails splitting a dataset into discrete parts, clustering groups related data points according to their intrinsic commonalities. In the huge and complicated world of big data, these techniques are essential for revealing hidden interactions, spotting spontaneous groups, and identifying meaningful distinctions. Organizations can apply algorithms like k-means, hierarchical clustering, or DBSCAN to group data points into clusters, which facilitates a more detailed comprehension of the underlying patterns. On the other hand, segmentation helps to divide big datasets into digestible and insightful chunks, enabling focused study. These approaches are utilized in an assortment of fields, such as cyber security anomaly detection and consumer segmentation in marketing. Organizations can extract actionable insights and optimize their strategies in response to the complex nature of large and diverse datasets by using clustering and segmentation, which not only improves the interpretability of big data but also lays the groundwork for more precise and effective decision-making processes.

D. Stream Processing for Real time Analytics

Real-time analytics requires stream processing since it allows data to be analyzed and responded to in real-time, as opposed to in batches. When dealing with large amounts of data that are continuously flowing in at a rapid pace from a variety of sources, batch processing techniques might not be able to provide timely insights. Stream processing technologies provide for real-time, continuous, parallel data stream analysis. As events develop, these technologies enable organizations to spot patterns, recognize trends, and get useful information. Applications for real-time analytics driven by stream processing are found in different fields, such as IoT for control and monitoring, healthcare for patient monitoring, and finance for fraud detection. When organizations are flexible enough to process and analyse data fast, they are better able to respond swiftly to changing conditions. Leveraging a flexible and responsive approach to data-driven decision-making requires stream processing.

E. Natural Language Processing

In Big Data Analytics, Natural Language Processing (NLP) is necessary for extracting insightful conclusions from an immense quantity of unstructured textual data. NLP is an interdisciplinary study that combines computer science and linguistics with the goal of enabling robots to produce, comprehend, and interpret human language. Natural language processing (NLP) techniques are useful for sentiment analysis, entity recognition, and topic modelling in the context of big data, which involves a large amount of text data from sources like social media, customer reviews, and documents. Textual content can be organized and analyzed with the use of methods like named entity recognition, tokenization, and part-of-speech tagging. Recurrent neural networks and transformer designs, such as BERT, are examples of machine learning models that improve language-based task accuracy and contextual awareness. Organizations can make better decisions by using natural language processing (NLP) to extract important information from vast volumes of text, including sentiments, trends, and important information. Natural language processing (NLP) plays a key role in utilizing the potential of unstructured textual data in the big data landscape, with applications ranging from social media monitoring to consumer feedback analysis.

F. Evolutionary Algorithms for Optimization

The sheer complexity and size of datasets in the field of big data analytics frequently present difficulties for conventional optimization techniques, which is where evolutionary algorithms (EAs) shine as powerful tools for optimization tasks. EAs iteratively develop a population of candidate solutions over generations, electing and recombining the fittest individuals to arrive at progressively optimal solutions. This process is inspired by the principles of natural selection. These methods are used for big data optimization tasks like model optimization, parameter adjustment, and feature selection. They are highly adapted to managing the complexity and variability present in big datasets because of their capacity to explore enormous solution areas and adjust to changing conditions. EAs can efficiently navigate and converge to near-optimal solutions in situations when the search space is high-dimensional and poorly understood. Organizations may improve the efficacy of machine learning models, optimize data processing operations, and get deeper insights from the massive amounts of data that comprise the big data landscape by utilizing evolutionary algorithms in big data optimization.

G. Distributed and Parallel Computing

CI frameworks have been optimized to operate adequately in circumstances with parallel and distributed computing. The key building blocks for overcoming the difficulties presented by the enormous amount and complexity of big data are distributed and parallel computing. Large-scale data processing could render it difficult for standard computing infrastructures to produce findings quickly enough. Tasks that require a lot of data are divided into smaller, more manageable pieces and distributed over several nodes or clusters using distributed computing platforms like Apache Hadoop and Apache Spark. In contrast, parallel computing speeds up processing time by carrying out these subtasks concurrently. By working together, this cooperative strategy greatly expedites data processing, allowing businesses to fully use their big data assets. In addition to increasing processing speed, distributing workloads among several nodes also guarantees fault tolerance, scalability, and dependability. Organizations are able to enable real-time analytics, machine learning, and other data-intensive applications by analyzing enormous information more effectively. The integration of distributed and parallel computing plays a crucial role in surmounting the computational impediments linked to large data, hence enabling the extraction of meaningful insights and promoting innovation across several fields.

H. Ensemble Learning for Improved Accuracy

CI ensemble learning approaches, such boosting algorithms and random forests, enhance the prediction accuracy, precision and resilience of prediction models. Combining predictions from multiple separate models to produce a final forecast that is more accurate and dependable is the fundamental idea behind ensemble learning. Ensemble techniques like bagging and boosting reduce the drawbacks of individual models and utilize performance of variety of algorithms. A more thorough and precise prediction is produced by the ensemble's ability to capture many aspects of the underlying data patterns due to the diversity of the constituent models. In order to decrease overfitting and increase generalisation, Random Forest, a well- liked ensemble approach, builds an ensemble of decision trees, each trained on a random subset of the data. Gradient Boosting is another popular method that produces a sequence of weak models one after the other, fixing the mistakes of the previous model. When dealing with complicated, noisy, or heterogeneous data, ensemble learning proves to be especially advantageous. When applied to big data analytics, it improves predicted accuracy, builds resilience against outliers, and eventually produces machine learning models that are more reliable and strong in a variety of contexts.

I. AutoML for Automated Model Selection

Big data analytics is being revolutionized by AutoML (Automated Machine Learning), which offers an organized and efficient method to identify, train, and refine machine learning models without necessitating a great deal of laborious work. AutoML streamlines the model selection process in the configuration of significant and complicated datasets, where the search space for optimal models and hyper parameters might be very large. It simplifies and automates every aspect of the machine learning process, encompassing feature engineering, model selection, data preprocessing, and hyper parameter modifications. AutoML systems leverage algorithms and heuristics to investigate diverse model and configuration combinations, with the ultimate objective of determining the most effective approaches for a specific issue. This approach promotes machine learning by enabling non-experts to gain insight into the power of complex algorithms while simultaneously conserving both resources and time.

V. Conclusion

CI in the context of big data entails applying certain machine learning algorithms, deep learning strategies, clustering techniques, and optimization algorithms to tackle the particular difficulties presented by sizable, varied, and dynamic datasets. These particular CI applications support efficient big data analysis, interpretation, and utilization across a range of industries. CI enables frequent integration of code changes, which promotes quick development and iteration in the setting of large scale data processing. In Big Data analytics, it is critical to identify problems early on. Scalability testing, a crucial component of Big Data analytics is effectively automated, guaranteeing that the analytics infrastructure can manage growing volumes of data. The blend of agility, prompt issue identification, collaboration, scalability testing, version control integration, and improved code quality promotes CI as a perfect complement for the intricate and constantly shifting demands of Big Data analytics initiatives.

References:

[1] Machine Learning for Big Data Processing: A Literature Review: Bharat Kumar Padhi, Dr. S S Nayak, and Dr. B N Biswal,December 2018 | IJIRT | Volume 5 Issue 7 | ISSN: 2349-6002.

[2] A Fuzzy Logic Approach for Data Mining On Big Data. Available from: https://www.researchgate.net/publication/326468814_A_Fuzzy_Logic_Approach_For_Data_Mining_On_Big_Data, June 2018.

[3] A Comparative Study of Business Intelligence and Artificial Intelligence with Big Data Analytics, Jasmin Praful Bharadiya, American Journal of Artificial Intelligence, Vol. 7, No. 1, 2023, pp. 24-30. DOI: 10.11648/j.ajai.20230701.14

[4] Expert systems: Definitions, Advantages and Issues in Medical Field Applications, Saibene A, Assale M, Giltri M - Expert Systems with Applications, 2021 – Elsevier

[5] Improved big data stock index prediction using deep learning with CNN and GRU, Jithin Eapen; Abhishek Verma; Doina Bein, DOI: 10.1504/IJBDI.2020.113868

[6] A Review on Big Data theoretical and Application approach: Past and Present: Rabel Guharoy, Sumit Kumar, Upasak Pal; 2017 IEEE

[7] A Very Brief Introduction to Machine Learning With Applications to Communication Systems,King's College London, United Kingdom ,Osvaldo Simeone IEEE Dec 2018

[8] A Survey on Machine Learning: Concept, Algorithms and Applications, Kajaree Das, Rabi Narayan Behera, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017.

[9] Machine Learning Algorithms: A Review Ayon Dey; International Journal of Computer Science and Information Technologies, Vol. 7 (3), 2016, 1174-1179.

[10] The Role of Fuzzy Logic in Decision Making Process; Katarína Valaskova1 A, Tomas KliestikB and Maria Misankova; 2nd International Conference on Management Innovation and Business Innovation, Jan 2014

[11] Fuzzy Based Machine Learning: A Promising Approach; Sujamol S., Sreeja Ashok and U Krishna Kumar, CSI Communications November 2017

[12] "Data Mining with Big Data", Xin dong Wu, Fellow, IEEE, Xing quan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE. IEEE transactions on knowledge and data engineering, vol. 26, no. 1, January 2014.

[13] "A survey of machine learning for big data Processing" Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng. Qiu et al. EURASIP Journal on Advances in Signal Processing (2016) 2016:67 DOI 10.1186/s13634- 016-0355-x.

[14] "Data intensive applications, challenges, techniques and technologies: A survey on Big Data". Philip Chen, C. L., & Zhang, C.-Y. Information Sciences, (2014).275, 314–347. DOI: 10.1016 /j.ins.2014.01.015.

[15] Big Data Provenance: Challenges, State of the Art and Opportunities, Jianwu Wang, Daniel Crawl, Shweta Purawat, Mai Nguyen, Ilkay, 2015 IEEE International Conference on Big Data.

[16] Big Data Analytics : challenges and solution, Puneet Singh Duggel, Sanchita Pual, International conference on Cloud, Big data and Trust 2013, Nov 13-15 RGPV.

[17] A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, D. P. Acharjya , Kauser Ahmed P, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[18] A formal definition of Big Data based on its essential features, Andrea De Mauro, Marco Greco and Michele Grimaldi, March 2016 DOI: 10.1108/LR-06- 2015-0061.

[19] Big Data Analytics: A Literature Review Paper Nada Elgendy and Ahmed Elragal, Department of Business Informatics & Operations ICDM 2014, LNAI 8557, pp. 214–227, 2014.© Springer International Publishing Switzerland 2014

[20] Algorithmic challenges in Big Data analytics, Veronica Bolon-Canedo, Beatriz Remeseiro, Konstantinos Sechidis et al., ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26- 28 April 2017, i6doc.com publ., ISBN 978-287587039-1. http://www.i6doc.com/en/