HANDLING UNCERTAINITY IN BIG DATA PROCESSING

Dr Jyothi A P
Assistant Professor, Dept. of CSE,
Faculty of Engineering and Technology,
Ramaiah University of Applied Sciences,
Bengaluru, Karnataka, India.
jyothiarcotprashant@gmail.com

Thejaswini R
Student, Dept. of CSE,
Faculty of Engineering and Technology,
Ramaiah University of Applied Sciences,
Bengaluru, Karnataka, India.
thejaswiniraju2002@gmail.com

Sharanya N R
Student, Dept. of CSE,
Faculty of Engineering and Technology,
Ramaiah University of Applied Sciences,
Bengaluru, Karnataka, India.
sharanyaramesh999@gmail.com

ABSTRACT

Big data analytics is the matter of huge importance as the demand for understanding trends in massive datasets increases due to the boom in the collection of data from various sources. Recent developments in sensor networks, cyber-physical systems, and the popularisation Internet of Things (IoT) have created an explosion in data. This century is the age of big data and each one of us is not just the generator but also a consumer. However, because of noise, incompleteness, and inconsistency, the data gathered from sensors, social media, financial records, etc., is intrinsically questionable. Advanced analytical approaches are needed for the examination of such enormous amounts of data in order to efficiently assess and/or anticipate future courses of action with high precision. Data's intrinsic uncertainty grows along with its volume, diversity, and speed, which makes the subsequent analytics process and its conclusions less reliable. Artificial intelligence approaches, such as machine learning, natural language processing, and computational intelligence, produce outcomes in big data analytics that are more precise, swifter, and more scalable when compared to traditional data methodologies and platforms.

I. INTRODUCTION

Until 2003 all of the humankind had generated about 5 Exabyte of data.(Exabyte means 10^{18} bytes). Today we generate 5 Exabyte of data every two days. It was estimated in 2012 that all data stored in world's computer was about 2.8 zettabytes and it was projected to reach 40 zettabytes by 2020 [1]. Google currently processes 3.5 billion searches daily, or more than 40,000 queries every second. Every day, Facebook users post 293,000 status updates, 510,000 comments, and 300 million photographs. Data production is astonishing. As a result, methods are needed to evaluate and comprehend this vast amount of data to a valuable source of knowledge [2].

Using the Advanced data analysis techniques one can transform big data into smart data and obtain critical information. Smart data provides actionable evidence and advances decision-making capabilities for organizations and companies. Big data analysis possess the five V's i.e. its characteristics: high volume, low veracity, high velocity, high variety, and high value [2, 4]. Big data also has a wide range of additional qualities, including variability, viscosity, validity, and viability. Machine learning (ML), natural language processing (NLP), computational intelligence (CI), and data mining are some of the artificial intelligence (AI) techniques that were developed to provide big data analytic solutions because they are quicker, more accurate, and more precise for massive volumes of data. These sophisticated analytical methods seek to discover information, obscure patterns, and unforeseen correlations in vast datasets.

Each of the V features introduces multiple sources of uncertainty, such as unstructured, incomplete, or noisy data. The entire analytics process may contain uncertainty (e.g., collecting, organizing, and analysing big data). The majority of data mining and ML algorithms face a significant hurdle when dealing with ambiguous and

partial data. In addition, biased training data may prevent an ML system from producing the best results. Wang et al. [3] identified six major issues in big data analytics, one of which is uncertainty. They primarily concentrate on how uncertainty affects the effectiveness of learning from big data. Because uncertainty can significantly affect the accuracy of an automated technique's output, minimising uncertainty in big data analytics must be a top priority.

II. BIG DATA

Big data was declared to be the upcoming frontier for productivity, innovation, and competition in May 2011. Over 3.7 billion individuals were Internet users in 2018, an increase of 7.5% from 2016. The amount of data generated globally increased from 1 zettabyte (ZB) in 2010 to 7 ZB in 2014. Volume, Velocity, and Variety (the three Vs) were used to identify the developing characteristics of big data in 2001. The four Vs—Volume, Variety, Velocity, and Value—were also used by International Data Corporation to define big data in 2011. Veracity was added as the fifth attribute of big data in 2012.

Volume describes a dataset's breadth and scope as well as the enormous amount of data generated per second. Currently, Exabyte (EB) or ZB-sized datasets are typically regarded as big data. Such enormous data quantities may result in scalability and unpredictability issues. When attempting to scan and comprehend the data at scale using many of the currently available data analysis tools, which are not meant for large-scale databases, they can fail [5].

Variety refers to the various types of data that can be found in a dataset, such as structured, semi-structured, and unstructured data. Unstructured data, such as text and multimedia information, is random and challenging to analyse, but structured data, such as that kept in a relational database, is typically well-organized and can be quickly sorted. Tags are used to distinguish data elements in semi-structured data, such as that found in NoSQL databases. Uncertainty can appear, When expressing data of mixed data types, converting between different data types, or making changes to the dataset's underlying structure in real time, . Traditional big data analytics algorithms confront difficulties when managing multi-modal, imperfect, and noisy data from the perspective of variety [5].

The speed of processing data (expressed as batch, near-real time, real time, and streaming) is referred to as velocity, with the emphasis that the processing pace of the data must match the production speed of the data.

The data's level of quality is indicated by veracity (e.g., uncertain or imprecise data). For instance, according to IBM's estimate [5], poor data quality costs the US economy $3.1 trillion annually. Data veracity is divided into three categories: good, terrible, and undefinable since data can be inconsistent, noisy, unclear, or incomplete. Accuracy and trust are harder to establish in big data analytics because of the amount and diversity of data sources.

Value is a measure of the context and usefulness of data in making decisions. For instance, Facebook, Google, and Amazon have each used analytics to maximise the value of big data in their individual products. To provide product recommendations and boost sales and customer engagement, Amazon analyses enormous datasets of users and their purchases. To enhance location services in Google Maps, Google collects location information from Android users. Facebook keeps track of user activity to deliver relevant advertisements and friend suggestions [6].

III. UNCERTAINITY IN DATA

Big data and big data analytics are subject to a variety of uncertainties that could have a detrimental effect on the efficacy and precision of the findings. For example, the learning algorithm employing corrupted training data will probably produce wrong results if the training data is prejudiced in any way, incomplete, or obtained by inaccurate sampling. To tackle uncertainty, it is crucial to improve big data analytical approaches. The quantity of a dataset and the degree of data processing and data uncertainty are positively correlated. There are many different sorts of uncertainty, and numerous theories and methods have been created to model them.

The Bayesian theory presupposes that the probability would be subjectively interpreted depending on previous experience or knowledge. According to this view, probability is described as a measure of how strongly a rational agent believes certain uncertain assertions [7]. When faced with ambiguity, belief function theory provides a paradigm for combining incomplete facts through an information fusion process [8]. The general focus of probability theory is on the statistical properties of the incoming data [9], which includes randomness. For a confidence index when classifying, classification entropy quantifies ambiguity between classes. Entropy ranges from zero to one, with closer values to zero indicating a more comprehensive classification in a single class and closer values to one indicating membership among several different classes [10]. Fuzziness is employed to

measure uncertainty in classes [11,12]. Fuzzy logic then develops an approximate reasoning method to deal with the ambiguity inherent in human perception. The methodology was designed to mimic human reasoning in order to deal with uncertainty more effectively in the actual world. Shannon's entropy measures the information in a variable to estimate the average quantity of missing data in a random source [13]. Shannon introduced the statistical idea of entropy into the science of communication and information transfer.

In big data analytics, determining the degree of uncertainty is a crucial step. Although there are several methods for analysing big data, the accuracy of the analysis may suffer if the method itself or data uncertainty is disregarded. Big data analytical tools can be supplemented with uncertainty models, such as probability theory, fuzziness, rough set theory, etc., to produce more precise and insightful conclusions.

IV. BIG DATA ANALYTICS

Massive datasets are studied using big data analytics to find patterns, undiscovered correlations, market trends, user preferences, and other useful information that was previously inaccessible to analysis using conventional technologies. In order to make better decisions, cut costs, and enable more effective processing, a number of advanced data analysis techniques (such as ML, data mining, NLP, and CI) and potential strategies, such as parallelization, divide-and-conquer, incremental learning, sampling, granular computing, feature selection, and instance selection, can break down large problems into smaller ones.

By dividing enormous problems into smaller instances of themselves and carrying out the smaller tasks concurrently, parallelization reduces computation time with regard to big data analytics.[11]

The divide-and-conquer technique is crucial in the processing of massive data. Three steps make up the divide-and-conquer strategy: (1) breaking down a large problem into smaller ones, (2) finishing the smaller ones, where each small problem's completion aids in the larger problem's completion, and (3) combining the smaller problems' solutions into the larger one, solving the larger problem. For many years, very large databases have employed the divide-and-conquer method to change records in groups rather than all the data at once [14].

A learning algorithm called incremental learning, which is frequently employed with streaming data, trains solely with new data as opposed to only existing data. Each fresh input data is used to alter the parameters in the learning process incrementally, and each input is only used once for training. By selecting, modifying, and evaluating a portion of the data, sampling can be used as a data reduction technique for big data analytics to identify patterns in massive data sets [11].

Granular computing groups components from a big space into smaller groups, or granules [15].

Feature selection is a common strategy to dealing with huge data that involves selecting a subset of relative features for an aggregate but more accurate data representation. In data mining, feature selection is a highly helpful technique for getting high-scale data ready [16].

Instance selection is a useful component of data pre-processing in many ML or data mining jobs. Reduce training sets and runtime in the classification or training stages by using instance selection [17].

V. IMPACT OF UNCERTAINITY IN BIG DATA ANALYTICS AND ITS MITIGATION TECHNIQUES

A) MACHINE LEARNING AND BIG DATA

ML is typically used to build models for prediction and knowledge discovery to make decisions based on data. For the analysis of huge data, a number of popular advanced ML techniques have been proposed, such as feature learning, deep learning, transfer learning, distributed learning, and active learning. A system can automatically find the representations required for feature detection or classification from raw data using a set of techniques called feature learning. The choice of data format has a significant impact on how well ML algorithms perform. Deep learning algorithms are created for evaluating and extracting useful knowledge from enormous amounts of data and data acquired from multiple sources [20], but current deep learning models have a significant computational cost. By performing calculations on data sets split across multiple workstations to scale up the learning process, distributed learning can be utilised to address the scalability issue of standard ML [20]. The ability to apply knowledge acquired in one context to new situations is known as transfer learning, and it can help learners in one domain become better learners in another [21]. In order to speed up ML activities and solve labelling issues, active learning refers to algorithms that use adaptive data collection [22] (i.e., procedures that automatically alter settings to acquire the most usable data as rapidly as possible). The main causes of the uncertainty issues with ML approaches include learning from poor veracity (i.e., uncertain and partial data) and low value data.

B) NATURAL LANGUAGE PROCESSING AND BIG DATA

NLP is an ML-based method that gives machines the ability to analyse, interpret, and even produce text. Huge volumes of text data are dealt with using NLP and big data analytics, which can quickly extract value from a dataset. In order to determine which sense of a word is used in a phrase when a word has many meanings, NLP techniques like as word sense disambiguation, part-of-speech (POS) tagging, and lexical acquisition are frequently used.

Uncertainty can affect keyword searches since a document's keyword content does not guarantee that the material will be relevant. For instance, a keyword search typically only finds exact strings and disregards potentially important words with spelling mistakes. Greater flexibility is made possible by the use of fuzzy search algorithms and boolean operators, which can be used to find words that spell a given word similarly. Automatic POS taggers that have to deal with the ambiguity of some terms (such as the word "very" having distinct meanings to American and British audiences, etc.) are another instance of how uncertainty affects NLP.

Big data analytics and the integration of NLP techniques may make it possible to handle large amounts of text by using uncertainty modelling approaches like fuzzy and probabilistic sets.

C) COMPUTATIONAL INTELLIGENCE AND BIG DATA

Computational Intelligence (CI) is a collection of nature-inspired computational approaches that play a vital role in big data analysis. Complex data processes and analytics problems, such as high complexity, uncertainty, and other processes where conventional methodologies are insufficient, have been addressed with CIs. Evolutionary algorithms (EAs), artificial neural networks (ANNs), and fuzzy logic are common techniques that are currently available in CI, with examples ranging from search-based challenges like parameter optimization to optimising a robot controller.

As they are inherently capable of handling various levels of uncertainty, CI approaches are appropriate for addressing the practical difficulties presented by large data. Big data analysis can be improved by using algorithms like swarm intelligence, AI, and ML, according to [18]. These methods are employed to train computers to carry out collaborative filtering, predictive analysis tasks, and the construction of empirical statistical predictive models. By using CI-based big data analytics solutions, it is possible to reduce the complexity and uncertainty associated with processing enormous volumes of data and enhance analysis outcomes.

Table 1 summarises the above discussions.

Table 1: Summary of Mitigation Strategies.

Technique	Uncertainty	Mitigation
Machine Learning	Incomplete data Unlabelled data Scalability	Active learning, Deep learning Active learning Distributed learning, Deep learning
Natural Language Processing	Keyword search Ambiguity in POS	Fuzzy
Computational Intelligence	High Variety Low Varacity	Swarn intelligence Fuzzy

VI. CONCLUSION

This paper has discussed on how the uncertainty impacts the big data. Discussions were made on the state of the art with respect to big data analytics techniques, and how uncertainty had negatively impacted the techniques. We have discussed the issues of the five V's of big data, and also glimpse was given regarding the other V's.

V. FUTURE DIRECTION

This paper notes several prospects for future work in the field of big data analytics. It is required to have additional studies on the interactions between each big data characteristic. The scalability and efficacy of existing analytics techniques must be further examined and new techniques and algorithms must be developed in ML and

NLP to handle the real-time needs for decisions made based on enormous amounts of data. And also focus must be towards building an efficient model to represent the effect of uncertainty on big data analytics. CI algorithms offers approximate solution within a reasonable time. However, further developments are to be made to apply it to mitigating uncertainty in big data analytics.

REFERENCES

[1] Stephen Marshland(2014). Machine Learning: An Algorithmic Perspective. Chapman and Hall, CRC Press

[2] Marr B. Forbes. How much data do we create every day? 2018.

[3] Wang X, He Y. Learning from uncertainty for big data: future analytical challenges and strategies. IEEE Syst Man Cybern Mag. 2016;2(2):26–31.

[4] Jain A. The 5 Vs of big data. IBM Watson Health Perspectives. 2017.

[5] IBM big data and analytics hub. Extracting Business Value from the 4 V's of Big Data. 2016.

[6] Court D. Getting big impact from big data. McKinsey Q. 2015;1:52–60.

[7] Bernardo JM, Smith AF. Bayesian theory, vol. 405. Hoboken: Wiley; 2009.

[8] Cuzzolin F. (Ed.). Belief functions: theory and applications. Berlin: Springer International Publishing; 2014.

[9] Wang Xizhao, Huang JZ, Wang X, Huang JZ. Editorial: uncertainty in learning from big data. Fuzzy Sets Syst. 2015

[10] Brown DG. Classification and boundary vagueness in mapping presettlement forest types. Int J Geogr Inf Sci. 1998;12(2):105–29

[11] Wang X, He Y. Learning from uncertainty for big data: future analytical challenges and strategies. IEEE Syst Man Cybern Mag. 2016

[12] Wang XZ, Ashfaq RAR, Fu AM. Fuzziness based sample categorization for classifier performance improvement. J Intell Fuzzy Syst. 2015

[13] Lesne A. Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. Math Struct Comput Sci. 2014

[14] Jordan MI. Divide-and-conquer and statistical inference for big data. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; p. 4. 2012.

[15] Yager RR. Decision making under measure-based granular uncertainty. Granular Comput. 1–9. 2018.

[16] Olvera-López JA, Carrasco-Ochoa JA, Martínez-Trinidad JF, Kittler J. A review of instance selection methods. Artif Intell Rev. 2010

[17] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002

[18] Gupta A. Big data analysis using computational intelligence and Hadoop: a study. 2015

[19] Qiu J, Wu Q, Ding G, Xu Y, Feng S. A survey of machine learning for big data processing. EURASIP J Adv Signal Process. 2016

[20] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016

[21] Athmaja S, Hanumanthappa M, Kavitha V. A survey of machine learning algorithms for big data analytics. IEEE. 2017