IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol. 9, Issue.6, June 2021 E-ISSN: 2347-2693

Uncertainty Handling In Big Data Analytics: Survey, Opportunities and Challenges

Priya Nagargoje[1*], Monali Baviskar[2]

[1]Department of Computer science, Marathdwada Institute of Technology, Aurangabad, India
[2] Department of Computer Engineering, Marathdwada Institute of Technology, Aurangabad, India

[*]*Corresponding Author: priya.nagargoje58@gmail.com, Tel.: 9689714646*

DOI: https://doi.org/10.26438/ijcse/v9i6.5963 | Available online at: www.ijcseonline.org

Received: 15/Jun/2021, Accepted: 20/Jun/2021, Published: 30/Jun/2021

Abstract— Big Data analysis and processing is a popular tool for Artificial Intelligence and Data Science to extract applicable solution from data across a broad range of application domains. Even though Big data is in the mainstream of operations as of 2020, With the increase in data processing and storage capacity, a large amount of data is available and because of that potential issues or challenges the researchers can address, some of these issues overlap with the data science field. One of the key issue is the inevitable existence of uncertainty in stored or missing values. Any uncertainty in a source causes its disadvantageous, complexity or inapplicability to use. It is importance to ensure the reliability and a value of data source. That is why it is crucial to eliminate uncertainty or to lower uncertainty influence because data without any analysis does not have much value. In this paper we review previous work in big data analytics and Survey of many theories and techniques which have been developed to model its various forms. We have described several common techniques such as Bayesian model and fuzzy set, Shannon's entropy. We present a discussion of open challenges and future directions for handling and eliminating uncertainty in this profile.

Keywords— Big Data, Data Sciences, Data Uncertainty, Uncertainty Elimination, Machine learning, NLP, Computational Intelligence.

I. INTRODUCTION

Big Data: Big Data, Data Science, Machine Learning, Statistical Learning and Artificial Intelligence is multi-sectoral and trans-disciplinary field that uses methods, processes, techniques, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and find a solutions for various directions of human activity. In the world of internet and Smartphone huge the amount of data produced every day with the help of social platform, surveys, and in person communication. With help of technical resources people are searching and uploading data on internet every second which leads to collecting a massive data. These data needs to process to extract useful and meaningful information. And here data analysis comes in picture to segregate and process this unstructured data.

Uncertainty: Uncertainty is a anomaly present in the incomplete, inaccurate, inconsistent and unreliable data. Uncertainty in a data it affects the effectiveness and accuracy of the results because of the complexity or inapplicability to use. With Any uncertainty appears in data it is difficult or impossible to work with the resource because it reduces the value of the data, performance of the system and will not meet the expected result. Various forms of uncertainty exist in big data like heterogeneous data type because of the variance in data collection sources and concepts. Sometimes uncertainty appears due to variance in representation of data and multimodality. Unlike the data collection data uncertainty occurs in every phase of big data like data processing, data analyzing [1].
Big data is emerging field for data sciences so that extract perfect, accurate and applicable information from available data is crucial steps. In data analysis some algorithm are present to overcome uncertainty and no of techniques exist to evaluating the level of uncertainty in big data analysis. The handling of the uncertainty is a difficult step in big data because of the features of big data like volume, velocity, and variety [1].

Volume: Analysis of big data with existing technique is near to impossible due to massive volume cause uncertainty in time and accuracy. Today data is generated from IoT and AI embedded sources which capture real-time image and voice and store for every second. Data related to environment is generated every second at forecasting technology, all this leads to create uncertainty in data efficiency.

Variety in data means different forms of data type, data structure in a dataset. Data may be in structured, semi-structured and unstructured form. Heterogeneity in data sources collect variety of data in dataset. Traditional analysis algorithms are meant to process or analyze same kind of data as classification, cluster formation, regression,

parallelization with traditional environment. 'Variety' in big data is one of the most complicated characteristics which create uncertainty and overcome this issue advanced technique needs to be develop.

Velocity means the speed of data processing. As we mentioned earlier big data is collected through social media network or internet, sensors, camera and other IoT embedded devices generate at very high speed so it is very important to find greatest solution that the speed with which the data is processing must match the speed with the data is generating otherwise inconsistency in data forms which again helps to cause uncertainty in big data.

Veracity represents the quality of the data. Data can be vague, un-compatible, noisy, inconclusive or incomplete due to diversity in sources. Big data need to undergo advanced analysis process to remove uncertainty to get accuracy and establish a trust of valuable output.

Value represents the usefulness and importance of data for finding a solution for desired application. Value of the dataset is considers in the solution making process of data science. High value dataset gives a better result within less time.

II. RELATED WORK

The problems of uncertainty in data represent great challenges for big data. Many researchers had stated several methods to deal with the problem of uncertainty. In reference [1]In the research paper "A Method to Solve Uncertainty Problem for Big Data Sources, IEEE Second International Conference on Data Stream Mining & Processing" ontology-based method to solve an uncertainty problem for big data sources**.** This method works in three part uncertainty Elimination, Problem analysis and Ontology Learning and Evolution.

[2] In this researcher work authors present MetaLP, a flexible, distributed statistical modeling framework suitable for large-scale data analysis, where statistical inference meets big data computing. For experimental they form a framework which are based on three component, parallel processing for data distribution, modern non-parametric learning algorithm and meta analysis techniques. Particularly framework proposed in this model solve two basic problem in big data i.e data scalability, automation of merged data.

[3] The main motive of this work is survey of all Machine Learning algorithms for big data especially to remove anomaly. Authors have systematically described all traditional and advanced machine learning algorithm with example. They also provided the brief information about tool available for data analysis to lessen the uncertainty problem with respective scalable proof.

[4] Detail Survey on uncertainty of big data is done in this research journal. They have addressed every detail related to uncertainty in big data which is helpful for many future research work. This Journal demonstrated every key issue in proposed technique and also stated the future scope for problem statement need to cover related to big data.

[5] In this paper authors works on only machine learning algorithms. They analyze big data characteristics and according to that best suitable ML algorithm to improve scalability. We have studied all this research work and on the basis of all knowledge we collect through previous proposed algorithm and methodology we tried to put in one systematic and understandable form at one place.

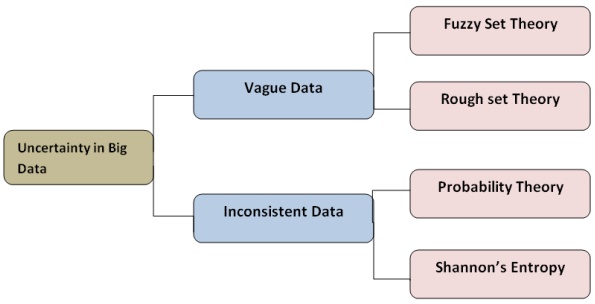

Figure 1: Uncertainty in Big Data

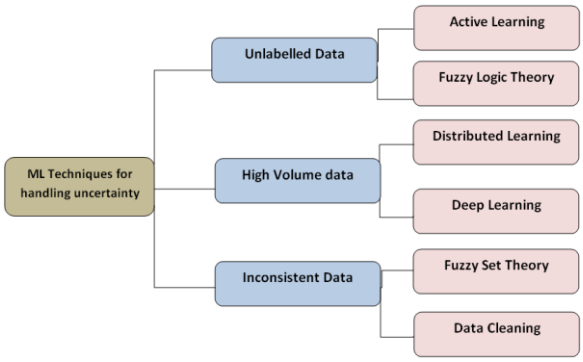

Figure 2: ML Techniques To Handle Uncertainty

III. MACHINE LERANING DATA ANALYSIS TECHNOLOGY FOR UNCERTINITY HANDELING

Machine learning techniques evolve information preprocessing, learning, and assessment stages. Data preprocessing uses data cleaning, extraction, combination and gets ready unrefined information into the frame which will be input for learning steps. The learning stage applies learning calculations and display parameters to give rise to useful preprocessed input information. The assessment stage decides the execution of the models which are educated based on learning algorithm and parameters. ML is generally used to educate models for prediction and valuable information extraction to enable data-driven decision making.

Designing an application of machine learning for big data analysis follows four design choices. a Choose the training data. b. Choose the target function. c. Choose the data representation. d. Choose the learning algorithm. Traditional ML methods are not able to handle seven v

characteristics and uncertainty of big data because they are not enough computationally efficient or scalable.

TRADITIONAL ML TECHNIQUES FOR UNCERTAINTY HANDLING IN BIG DATA.

Parametric & Non-Parametric Machine Learning Algorithms

Machine learning algorithm find a best mapping learning function which is use to target mapping of input variables as a best map to an output variable Parametric & Non-Parametric algorithm. Parametric machine learning algorithms simplify the function to a known form. This algorithm involves two steps: 1.Select a form for the function. 2. Learn the coefficients for the function from the training data. [4.2]

Parametric machine learning algorithms

- Linear Regression
- Logistic Regression.

Nonparametric machine learning algorithms do not make strong assumptions about the form of the mapping function. By not making assumptions, they are free to learn any functional form from the training data. Non-parametric methods are often more flexible, achieve better accuracy but require a lot more data and training time.

Nonparametric algorithms

- Support Vector Machines
- Neural Networks
- Decision Trees.

ADVANCED ML TECHNIQUES PROPOSED FOR BIG DATA ANALYSIS

Transfer learning, distributed learning, feature learning and active learning and deep learning are advanced ML techniques proposed for big data analysis.

- Transfer Learning :

[4]Transfer learning is the algorithm where learned information form one domain will transfer to related domain to effectively improving by applying learned knowledge in one context to new contexts.

- Distributed Learning:

Distributed learning calculations on data sets distributed among several workstations to scale up the learning process which helps to eliminate the traditional ML scalability problem.

- Deep Learning:

Deep Learning is a subfield of machine learning where algorithms are inspired by the structure and function of the ANN. In deep learning the performance of this type of model improves by training them with more examples by increasing their depth or representational capacity. Deep learning is use to eliminate uncertainty in the form of incompleteness and inconsistency.

- Active learning

Active learning are algorithms that engage processes that automatically adjust parameters to collect the most useful data as quickly as possible in order to stimulate machine learning activities and overcome labeling problems.

Challenges for ML techniques

It become challenging when data is represented without labels and this unlabeled data leads to unclear result because learning from unlabeled data is critical task. This issue can be solved by active learning which select subset of the most important instances for labeling.[5]

Learning from data with low veracity which is uncertain and incomplete data and data with low value means unrelated to the current problem are the uncertainty challenges of ML techniques. According to previous research on machine learning algorithm active learning, deep learning, and fuzzy logic theory are among the ML techniques uniquely suited to support the challenge of reducing uncertainty in big data.

Incomplete or imprecise training samples, unclear classification boundaries, and rough knowledge of the target data are leads to uncertainty, and this kind of uncertainty can be remove by data cleaning, fuzzy logic deep learning algorithm.

IV. NLP DATA ANALYSIS TECHNIQUE FOR UNCERTINITY HANDLING:

Using NLP techniques value can be retrieve from real-time dataset with the help of big data analytics. NLP tackle text data to mitigate uncertainty which enables devices to analyze, interpret and even generate text[4.3].

- Lexical acquisition.
- Word sense disambiguation.
- Part-of-speech (POS) tagging.

Lexical acquisition.

Lexical knowledge acquisition is central part of design of language processors. LA technique is used to get information about the lexical units of a language.

Word sense disambiguation.

Word sense disambiguation used for multiple meanings sentences to determine which sense of word is used in particular sentences.

[4.4]Part-of-speech (POS) tagging

Part-of-speech (POS) tagging is used to finding function of the world by labeling categories such as noun, verb and adverbs.

According to previous study in NLP-based techniques they have been applied to text mining including information extraction, topic modeling, text summarization, classification, clustering, question answering, and opinion mining. And also entity extraction and information retrieval NLP techniques used to mange and shift through massive amount of information in text form. Moreover, NLP techniques find semantic similarity among available textual articles to create or recover traceability links like broken or missing links at run-time.

CHALLENGES IN NLP TECHNIQUES

[5]NLP techniques face challenges due to native approach

for data analysis. In NLP keyword search is basic algorithm which perform efficiently in data mining terminology where keyword search method apply which collect relevant word and phrases and form related dataset. But in the case where spelling error in present then it ignore that word even though that word is relevant to a dataset [5].

NLP techniques try to find co-related data with exact string matching technique which again create problem in dataset formation because if exact string did not match it will neglect the world which will be most relevant to dataset.
In Pos tagging also some challenges are need to overcome related to context because in different language same word have different meaning which again create problems in dataset management.

V. COMPUTATIONAL INTELLIGENCE DATA ANALYSIS TECHNIQUES FOR HANDLE UNCERTINITY

In CI most of the computational techniques are nature-inspired which holds an important value in big data analysis The main advantage of using Computational intelligence for bid data analysis to remove uncertainty because it have capability to tackle complicated data processes and analytics challenges such as high complexity. CI is used any processes where traditional techniques are not sufficient. Common techniques that are currently available in CI are Artificial neural networks (ANN), evolutionary algorithms (EAs), and fuzzy logic and decision tree.

ANN: ARTIFICIAL NEURAL NETWORKS

In Artificial Neural Network system changes its structure depending on information that flows through the network.
In Artificial Neural Network system changes its structure depending on information that flows through the network. Fundamentally ANN techniques are capable of handling big amounts of uncertainty in data that's make them most suitable CI technique to deal with real-world application problem occurs in big data field.

— DECISION TREES

In decision tree model data divide into small section all section simultaneously perform logic to develop related decision tree. Decision tree model perform on traditional techniques as regression and classification model to analyze data to lessen the uncertainty in complex data. It names as decision tree because it is tree shaped where every node represents subset of large dataset [5.2].

— EVOLUTIONARY ALGORITHM

This is a multi-domain technique which can explore ample search space and manages attribute interaction very efficiently. EAs are excellent tools for analyzing datasets which includes high volume, variety, and low veracity. In EA with the help of evolutionary algorithm different type of data analysis techniques used to analyze high complexity data related to human activity.

— Fuzzy Logic

Fuzzy Logic use probability method to handle the uncertainties in data mining techniques. It uses quantifiers linguistic to provide a model for approximate reasoning and qualitative data to support CI in big data analytics to remove uncertainty. It represents uncertainty in real word and user-defined methodologies and interpretable fuzzy rules that can be used for inference and decision-making.

CHALLENGES IN CI TECHNIQUES

[5]Like any other technique CI also have many challenges in data consists of high degrees of uncertainty and outlier artifacts due to the existence of noise in data

CI-based algorithms may be impacted by motion, noise, and unexpected environments which impact on results of data analysis. Moreover algorithm function poorly when impacted by multiple factors which are developed so that can deal with one problems. In current CI techniques most of the challenges exist when dealing with the value and veracity characteristics of big data

VI. CONCLUSION

In this research we reviewed seven 'V' characteristics of big data and uncertainty occurrence due to one or more characteristics. In this paper we tried to explain fundamentals of data analysis for big data. We have present in detail description about impact of uncertainty on result of domain specific application in big data.

Further we discuss the traditional and advanced techniques which have already stated for uncertainty handling and elimination.

In this research we put an AI, ML, CI and NLP based advanced data analysis techniques because of complexity of big data. With the help of example we tried to explain challenges for available technique. In last session we put our point of view related to problem statement which are need to study in depth in the field of big data to handle uncertainty. Along with available technique, in depth analysis of advance technique in real time environment with actual parameter need to study to find out best possible solution.

Abbreviations

ML- Machine Learning, NLP- Neuro-linguisti programming, ANN- Artificial Neural Networks, EA- Evolutionary.

VII. FUTURE SCOPE

This paper reviewed uncertainty in big data and available techniques to handle or eliminate it. We found that there are number of core problems need to solve with existing technique and experimental study need to be done for each advance techniques to analyze efficiency. As we have stated in above discussion, available AI techniques also face challenges while execution so new algorithm or

terminology need to develop to overcome these challenges. According to the previous research Computational Intelligence algorithm like active learning can help in mitigate uncertainty in big data with better efficiency so these algorithm need to examine empirically with real world dataset.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of the entire researcher in Big data, machine learning and Data sciences field. We sincerely acknowledge the Reihaneh H. Hariri, Erik M. Fredericks and Kate M. Bowers to help us to understand and clarify our fundamental concept about uncertainty through the journal.

REFERENCES

[1] Andrii Berko,Vladyslav Alieksieiev, "A Method to Solve Uncertainty Problem for Big Data Sources, IEEE Second International Conference on Data Stream Mining & Processing, **August 21-25, 2018**, Lviv, Ukraine.

[2] Scott Bruce, , Zeda Li, Hsiang-Chieh Yang, Subhadeep Mukhopadhyay " Nonparametric Distributed Learning Architecture for Big Data: Algorithm and Applications" IEEE TRANSACTIONS ON BIG DATA, NOVEMBER **2017.**

[3] K. Sree Divya, P.Bhargavi, S. Jyothi," Machine Learning Algorithms in Big data Analytics" INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING · **January 2018.**

[4] Reihaneh H. Hariri, Erik M. Fredericks and Kate M. Bowers," Uncertainty in big data analytics: survey, opportunities, and challenges" Hariri et al. J Big Data Springer open source journal **2019.**

[5] Athmaja S, Hanumanthappa M, Vasantha Kavitha," A SURVEY OF MACHINE LEARNING ALGORITHMS FOR BIG DATA ANALYTICS" IEEE International Conference on Innovations in Information, Embedded and Communication Systems **2017.**

[6] M. U. Bokhari, M. Zeyauddin and M. A. Siddiqui, "An effective model for big data analytics", 3rd International Conference on Computing for Sustainable Global Development, **pp. 3980-3982, 2016.**

[7] Monali R. Baviskar, Priya N Nagargoje, Priyanka A. Deshmukh, Rina R. Baviskar, "A Survey of Data Science Techniques and Available Tools", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 **Volume: 08 Issue: 04 | Apr 2021** p-ISSN: 2395-0072

[8] R. Farias and E. Clua," Parallel Image Segmentation using Reduction-Sweeps on Multi-core Processors and GPUs IEEE" 26th Conference on Graphics, Patterns and Images (SIBGRAPI), Rio de Janeiro, Brazil.

[6] Z. Yang, Y. Zhu, and Y. Pu, Parallel image processing based on CUDA, In Proceedings of the 2008 ACM International Conference on Computer Science and Software Engineering - **Volume 03 (CSSE '08) , 2008, Vol. 3, pp. 198-201.**

[7] J. L. Berral-Garcia, "A quick view on current techniques and machine learning algorithms for big data analytics", 18th International Conf. on Transparent Optical Networks, pp.1-4, 2016. DOI: 10.1109/ICTON.2016.7550517.

[8] B. Kulis and M. I. Jordan A New k-Means algorithms via Bayesian Non parametrics Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, **2012.**

[9] N. Marz , and J. Warren, Big Data: Principles and best practices of scalable real time data systems. Manning Publications, 2015.

[10] J. Chen, D. Dosyn, V. Lytvyn, and A. Sachenko, "Smart Data Integration by Goal Driven Ontology Learning. Advances in Big Data," Advances in Intelligent Systems and Computing, Springer.

[11] Saidulu D, Sasikala R. Machine learning and statistical approaches for Big Data: issues, challenges and research directions. Int J Appl Eng Res. **2017;12(21):11691–9.**

[12] J. L. Berral-Garcia, "A quick view on current techniques and machine learning algorithms for big data analytics", 18th International Conf. on Transparent Optical Networks, pp.1-4, 2016. DOI: 10.1109/ICTON.2016.7550517.

[13] Parmar, V. & Gupta, I., 2015. Big data analytics vs Data Mining analytics. *IJITE*, **3(3), pp.258–263.**

[14] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, and A. Plaza, "On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **vol. 8, pp. 4634-4646, 2015.**

[15] Joseph D. Prusa* andTaghi M. Khoshgoftaar, "Improving deep neural network design with new text data representations" J Big Data (2017) 4:7DOI 10.1186/s40537-017-0065-8.

[16] Karl Weiss* , Taghi M. Khoshgoftaar and Ding Ding Wang Weiss," A survey of transfer learning" et al. J Big Data (2016) 3:9 DOI 10.1186/s40537-016-0043-6.

[17] Priya nagargoje, vandna jagtap, "time-efficient fracture detection using multi-core parallel processing", international journal of current engineering and scientific research (ijcesr) issn (print): 2393-8374, (online): 2394-0697, **volume-3, issue-9, 2016.**

AUTHORS PROFILE

Priya Nagargoje is an Assistant Professor, in the Department of Computer Engineering in MIT from Aurangabad University. Her research interests include Parallel Programming, High Performance Computing and Intelligent systems. She has published more than 5 research papers in conferences, workshops and reputed International Journals.

Email: priya.nagargoje@mit.asia

