Journal of Big Data

SURVEY PAPER Open Access

Uncertainty in big data analytics: survey, opportunities, and challenges

Reihaneh H. Hariri*, Erik M. Fredericks and Kate M. Bowers

*Correspondence:
rhosseinzadehha@oakland.edu
Oakland University,
Rochester, MI, USA

Abstract

Big data analytics has gained wide attention from both academia and industry as the demand for understanding trends in massive datasets increases. Recent developments in sensor networks, cyber-physical systems, and the ubiquity of the Internet of Things (IoT) have increased the collection of data (including health care, social media, smart cities, agriculture, finance, education, and more) to an enormous scale. However, the data collected from sensors, social media, financial records, etc. is inherently uncertain due to noise, incompleteness, and inconsistency. The analysis of such massive amounts of data requires advanced analytical techniques for efficiently reviewing and/or predicting future courses of action with high precision and advanced decision-making strategies. As the amount, variety, and speed of data increases, so too does the uncertainty inherent within, leading to a lack of confidence in the resulting analytics process and decisions made thereof. In comparison to traditional data techniques and platforms, artificial intelligence techniques (including machine learning, natural language processing, and computational intelligence) provide more accurate, faster, and scalable results in big data analytics. Previous research and surveys conducted on big data analytics tend to focus on one or two techniques or specific application domains. However, little work has been done in the field of uncertainty when applied to big data analytics as well as in the artificial intelligence techniques applied to the datasets. This article reviews previous work in big data analytics and presents a discussion of open challenges and future directions for recognizing and mitigating uncertainty in this domain.

Keywords: Big data, Uncertainty, Big data analytics, Artificial intelligence

Introduction

According to the National Security Agency, the Internet processes 1826 petabytes (PB) of data per day [1]. In 2018, the amount of data produced every day was 2.5 quintillion bytes [2]. Previously, the International Data Corporation (IDC) estimated that the amount of generated data will double every 2 years [3], however 90% of all data in the world was generated over the last 2 years, and moreover Google now processes more than 40,000 searches every second or 3.5 billion searches per day [2]. Facebook users upload 300 million photos, 510,000 comments, and 293,000 status updates per day [2, 4]. Needless to say, the amount of data generated on a daily basis is staggering. As a result, techniques are required to analyze and understand this massive amount of data, as it is a great source from which to derive useful information.

Advanced data analysis techniques can be used to transform big data into smart data for the purposes of obtaining critical information regarding large datasets [5, 6]. As such, smart data provides actionable information and improves decision-making capabilities for organizations and companies. For example, in the field of health care, analytics performed upon big datasets (provided by applications such as Electronic Health Records and Clinical Decision Systems) may enable health care practitioners to deliver effective and affordable solutions for patients by examining trends in the overall history of the patient, in comparison to relying on evidence provided with strictly localized or current data. Big data analysis is difficult to perform using traditional data analytics [7] as they can lose effectiveness due to the five V's characteristics of big data: high volume, low veracity, high velocity, high variety, and high value [7–9]. Moreover, many other characteristics exist for big data, such as variability, viscosity, validity, and viability [10]. Several artificial intelligence (AI) techniques, such as machine learning (ML), natural language processing (NLP), computational intelligence (CI), and data mining were designed to provide big data analytic solutions as they can be faster, more accurate, and more precise for massive volumes of data [8]. The aim of these advanced analytic techniques is to discover information, hidden patterns, and unknown correlations in massive datasets [7]. For instance, a detailed analysis of historical patient data could lead to the detection of destructive disease at an early stage, thereby enabling either a cure or more optimal treatment plan [11, 12]. Additionally, risky business decisions (e.g., entering a new market or launching a new product) can profit from simulations that have better decision-making skills [13].

While big data analytics using AI holds a lot of promise, a wide range of challenges are introduced when such techniques are subjected to uncertainty. For instance, each of the V characteristics introduce numerous sources of uncertainty, such as unstructured, incomplete, or noisy data. Furthermore, uncertainty can be embedded in the entire analytics process (e.g., collecting, organizing, and analyzing big data). For example, dealing with incomplete and imprecise information is a critical challenge for most data mining and ML techniques. In addition, an ML algorithm may not obtain the optimal result if the training data is biased in any way [14, 15]. Wang et al. [16] introduced six main challenges in big data analytics, including uncertainty. They focus mainly on how uncertainty impacts the performance of learning from big data, whereas a separate concern lies in mitigating uncertainty inherent within a massive dataset. These challenges normally present in data mining and ML techniques. Scaling these concerns up to the big data level will effectively compound any errors or shortcomings of the entire analytics process. Therefore, mitigating uncertainty in big data analytics must be at the forefront of any automated technique, as uncertainty can have a significant influence on the accuracy of its results.

Based on our examination of existing research, little work has been done in terms of how uncertainty significantly impacts the confluence of big data and the analytics techniques in use. To address this shortcoming, this article presents an overview of the existing AI techniques for big data analytics, including ML, NLP, and CI from the perspective of uncertainty challenges, as well as suitable directions for future research in these domains. The contributions of this work are as follows. First, we consider uncertainty challenges in each of the 5 V's big data characteristics. Second, we review several techniques on big data analytics with impact of uncertainty for each technique, and also review the impact of uncertainty on several big data analytic techniques. Third, we discuss available strategies to handle each challenge presented by uncertainty.

To the best of our knowledge, this is the first article surveying uncertainty in big data analytics. The remainder of the paper is organized as follows. "Background" section presents background information on big data, uncertainty, and big data analytics. "Uncertainty perspective of big data analytics" section considers challenges and opportunities regarding uncertainty in different AI techniques for big data analytics. "Summary of mitigation strategies" section correlates the surveyed works with their respective uncertainties. Lastly, "Discussion" section summarizes this paper and presents future directions of research.

Background

This section reviews background information on the main characteristics of big data, uncertainty, and the analytics processes that address the uncertainty inherent in big data.

Big data

In May 2011, big data was announced as the next frontier for productivity, innovation, and competition [11]. In 2018, the number of Internet users grew 7.5% from 2016 to over 3.7 billion people [2]. In 2010, over 1 zettabyte (ZB) of data was generated worldwide and rose to 7 ZB by 2014 [17]. In 2001, the emerging characteristics of big data were defined with three V's (Volume, Velocity, and Variety) [18]. Similarly, IDC defined big data using four V's (Volume, Variety, Velocity, and Value) in 2011 [19]. In 2012, Veracity was introduced as a fifth characteristic of big data [20–22]. While many other V's exist [10], we focus on the five most common characteristics of big data, as next illustrated in Fig. 1.

Volume refers to the massive amount of data generated every second and applies to the size and scale of a dataset. It is impractical to define a universal threshold for big data volume (i.e., what constitutes a 'big dataset') because the time and type of data can influence its definition [23]. Currently, datasets that reside in the exabyte (EB) or ZB ranges are generally considered as big data [8, 24], however challenges still exist for datasets in smaller size ranges. For example, Walmart collects 2.5 PB from over a million customers every hour [25]. Such huge volumes of data can introduce scalability and uncertainty problems (e.g., a database tool may not be able to accommodate infinitely large datasets). Many existing data analysis techniques are not designed for large-scale databases and can fall short when trying to scan and understand the data at scale [8, 15].

Variety refers to the different forms of data in a dataset including structured data, semi-structured data, and unstructured data. Structured data (e.g., stored in a relational database) is mostly well-organized and easily sorted, but unstructured data (e.g., text and multimedia content) is random and difficult to analyze. Semi-structured data (e.g., NoSQL databases) contains tags to separate data elements [23, 26], but enforcing this structure is left to the database user. Uncertainty can manifest when converting between different data types (e.g., from unstructured to structured data), in representing data of mixed data types, and in changes to the underlying structure of the dataset at run time. From the point of view of variety, traditional big data

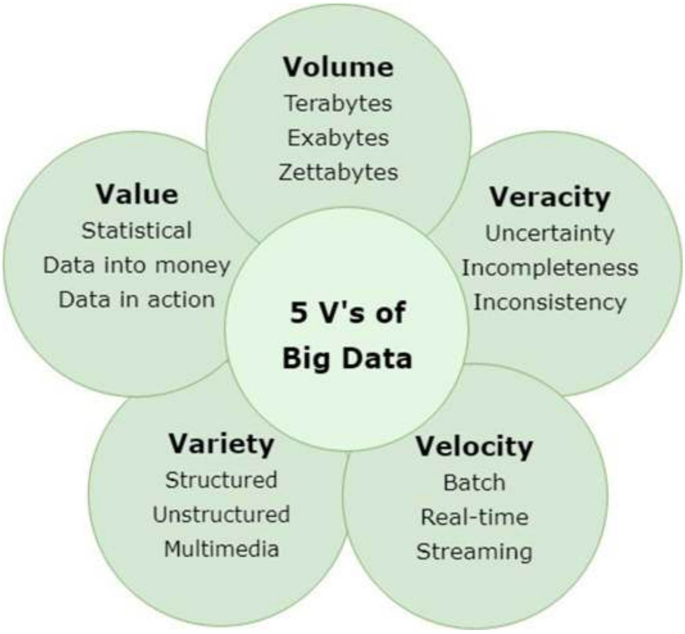

Fig. 1 Common big data characteristics

analytics algorithms face challenges for handling multi-modal, incomplete and noisy data. Because such techniques (e.g., data mining algorithms) are designed to consider well-formatted input data, they may not be able to deal with incomplete and/or different formats of input data [7]. This paper focuses on uncertainty with regard to big data analytics, however uncertainty can impact the dataset itself as well.

Efficiently analysing unstructured and semi-structured data can be challenging, as the data under observation comes from heterogeneous sources with a variety of data types and representations. For example, real-world databases are negatively influenced by inconsistent, incomplete, and noisy data. Therefore, a number of data preprocessing techniques, including data cleaning, data integrating, and data transforming used to remove noise from data [27]. Data cleaning techniques address data quality and uncertainty problems resulting from variety in big data (e.g., noise and inconsistent data). Such techniques for removing noisy objects during the analysis process can significantly enhance the performance of data analysis. For example, data cleaning for error detection and correction is facilitated by identifying and eliminating mislabeled training samples, ideally resulting in an improvement in classification accuracy in ML [28].

Velocity comprises the speed (represented in terms of batch, near-real time, real time, and streaming) of data processing, emphasizing that the speed with which the data is processed must meet the speed with which the data is produced [8]. For example, Internet of Things (IoT) devices continuously produce large amounts of sensor data. If the device monitors medical information, any delays in processing the data and sending the results to clinicians may result in patient injury or death (e.g., a pacemaker that reports emergencies to a doctor or facility) [20]. Similarly, devices in the cyber-physical domain often rely on real-time operating systems enforcing strict timing standards on execution, and as such, may encounter problems when data provided from a big data application fails to be delivered on time.

Veracity represents the quality of the data (e.g., uncertain or imprecise data). For example, IBM estimates that poor data quality costs the US economy $3.1 trillion per year [21]. Because data can be inconsistent, noisy, ambiguous, or incomplete, data veracity is categorized as good, bad, and undefined. Due to the increasingly diverse sources and variety of data, accuracy and trust become more difficult to establish in big data analytics. For example, an employee may use Twitter to share official corporate information but at other times use the same account to express personal opinions, causing problems with any techniques designed to work on the Twitter dataset. As another example, when analyzing millions of health care records to determine or detect disease trends, for instance to mitigate an outbreak that could impact many people, any ambiguities or inconsistencies in the dataset can interfere or decrease the precision of the analytics process [21].

Value represents the context and usefulness of data for decision making, whereas the prior V's focus more on representing challenges in big data. For example, Facebook, Google, and Amazon have leveraged the value of big data via analytics in their respective products. Amazon analyzes large datasets of users and their purchases to provide product recommendations, thereby increasing sales and user participation. Google collects location data from Android users to improve location services in Google Maps. Facebook monitors users' activities to provide targeted advertising and friend recommendations. These three companies have each become massive by examining large sets of raw data and drawing and retrieving useful insight to make better business decisions [29].

Uncertainty

Generally, "uncertainty is a situation which involves unknown or imperfect information" [30]. Uncertainty exists in every phase of big data learning [7] and comes from many different sources, such as data collection (e.g., variance in environmental conditions and issues related to sampling), concept variance (e.g., the aims of analytics do not present similarly) and multimodality (e.g., the complexity and noise introduced with patient health records from multiple sensors include numerical, textual, and image data). For instance, most of the attribute values relating to the timing of big data (e.g., when events occur/have occurred) are missing due to noise and incompleteness. Furthermore, the number of missing links between data points in social networks is approximately 80% to 90% and the number of missing attribute values within patient reports transcribed from doctor diagnoses are more than 90% [31]. Based on IBM research in 2014, industry analysts believe that, by 2015, 80% of the world's data will be uncertain [32].

Various forms of uncertainty exist in big data and big data analytics that may negatively impact the effectiveness and accuracy of the results. For example, if training data is biased in any way, incomplete, or obtained through inaccurate sampling, the learning algorithm using corrupted training data will likely output inaccurate results. Therefore, it is critical to augment big data analytic techniques to handle uncertainty. Recently, meta-analysis studies that integrate uncertainty and learning from data have seen a sharp increase [33–35]. The handling of the uncertainty embedded in the entire process of data analytics has a significant effect on the performance of learning from big data [16]. Other research also indicates that two more features for big data, such as multimodality (very complex types of data) and changed-uncertainty (the modeling and measure of uncertainty for big data) is remarkably different from that of small-size data. There is also a positive correlation in increasing the size of a dataset to the uncertainty of data itself and data processing [34]. For example, fuzzy sets may be applied to model uncertainty in big data to combat vague or incorrect information [36]. Moreover, and because the data may contain hidden relationships, the uncertainty is further increased.

Therefore, it is not an easy task to evaluate uncertainty in big data, especially when the data may have been collected in a manner that creates bias. To combat the many types of uncertainty that exist, many theories and techniques have been developed to model its various forms. We next describe several common techniques.

Bayesian theory assumes a subjective interpretation of the probability based on past event/prior knowledge. In this interpretation the probability is defined as an expression of a rational agent's degrees of belief about uncertain propositions [37]. *Belief function theory* is a framework for aggregating imperfect data through an information fusion process when under uncertainty [38]. *Probability theory* incorporates randomness and generally deals with the statistical characteristics of the input data [34]. *Classification entrop*y measures ambiguity between classes to provide an index of confidence when classifying. Entropy varies on a scale from zero to one, where values closer to zero indicate more complete classification in a single class, while values closer to one indicate membership among several different classes [39]. *Fuzziness* is used to measure uncertainty in classes, notably in human language (e.g., good and bad) [16, 33, 40]. Fuzzy logic then handles the uncertainty associated with human perception by creating an approximate reasoning mechanism [41, 42]. The methodology was intended to imitate human reasoning to better handle uncertainty in the real world [43]. *Shannon's entropy* quantifies the amount of information in a variable to determine the amount of missing information on average in a random source [44, 45]. The concept of entropy in statistics was introduced into the theory of communication and transmission of information by Shannon [46]. Shannon entropy provides a method of information quantification when it is not possible to measure criteria weights using a decision–maker. *Rough set theory* provides a mathematical tool for reasoning on vague, uncertain or incomplete information. With the rough set approach, concepts are described by two approximations (upper and lower) instead of one precise concept [47], making such methods invaluable to dealing with uncertain information systems [48]. Probabilistic theory and Shannon's entropy are often used to model imprecise, incomplete, and inaccurate data. Moreover, fuzzy set and rough theory are used for modeling vague or ambiguous data [49], as shown in Fig. 2.

Evaluating the level of uncertainty is a critical step in big data analytics. Although a variety of techniques exist to analyze big data, the accuracy of the analysis may be negatively affected if uncertainty in the data or the technique itself is ignored. Uncertainty models such as probability theory, fuzziness, rough set theory, etc. can be used to augment big data analytic techniques to provide more accurate and more meaningful results. Based on the previous research, Bayesian model and fuzzy set theory are common for modeling uncertainty and decision-making. Table 1 compares and

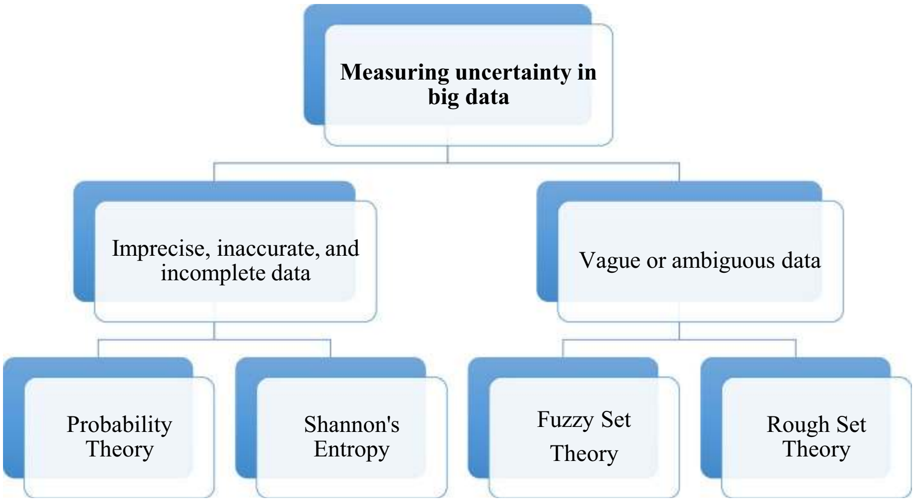

Fig. 2 Measuring uncertainty in big data

Table 1 Comparison of uncertainty strategies

Uncertainty models	Features
Probability theory Bayesian theory Shannon's entropy	Powerful for handling randomness and subjective uncertainty where precision is required Capable of handling complex data [50]
Fuzziness	Handles vague and imprecise information in systems that are difficult to model Precision not guaranteed Easy to implement and interpret [50]
Belief function	Handle situations with some degree of ignorance Combines distinct evidence from several sources to compute the probability of specific hypotheses Considers all evidence available for the hypothesis Ideal for incomplete and high complex data Mathematically complex but improves uncertainty reduction [50]
Rough set theory	Provides an objective form of analysis [47] Deals with vagueness in data Minimal information necessary to determine set membership Only uses the information presented within the given data [51]
Classification entropy	Handles ambiguity between the classes [39]

summarizes the techniques we have identified as relevant, including a comparison between different uncertainty strategies, focusing on probabilistic theory, Shannon's entropy, fuzzy set theory, and rough set theory.

Big data analytics

Big data analytics describe the process of analyzing massive datasets to discover patterns, unknown correlations, market trends, user preferences, and other valuable information that previously could not be analyzed with traditional tools [52]. With the formalization of the big data's five V characteristics, analysis techniques needed to be reevaluated to overcome their limitations on processing in terms of time and space [29]. Opportunities for utilizing big data are growing in the modern world of digital data. The global annual growth rate of big data technologies and services is predicted to increase about 36% between 2014 and 2019, with the global income for big data and business analytics anticipated to increase more than 60% [53].

Several advanced data analysis techniques (i.e., ML, data mining, NLP, and CI) and potential strategies such as parallelization, divide-and-conquer, incremental learning, sampling, granular computing, feature selection [16], and instance selection [34] can convert big problems to small problems and can be used to make better decisions, reduce costs, and enable more efficient processing.

With respect to big data analytics, *parallelization* reduces computation time by splitting large problems into smaller instances of itself and performing the smaller tasks simultaneously (e.g., distributing the smaller tasks across multiple threads, cores, or processors). Parallelization does not decrease the amount of work performed but rather reduces computation time as the small tasks are completed at the same point in time instead of one after another sequentially [16].

The *divide-and-conquer* strategy plays an important role in processing big data. Divide-and-conquer consists of three phases: (1) reduce one large problem into several smaller problems, (2) complete the smaller problems, where the solving of each small problem contributes to the solving of the large problem, and (3) incorporate the solutions of the smaller problems into one large solution such that the large problem is considered solved. For many years the divide-and-conquer strategy has been used in very massive databases to manipulate records in groups rather than all the data at once [54].

Incremental learning is a learning algorithm popularly used with streaming data that is trained only with new data rather than only training with existing data. Incremental learning adjusts the parameters in the learning algorithm over time according to each new input data and each input is used for training only once [16].

Sampling can be used as a data reduction method for big data analytics for deriving patterns in large data sets by choosing, manipulating, and analyzing a subset of the data [16, 55]. Some research indicates that obtaining effective results using sampling depends on the data sampling criteria used [56].

Granular computing groups elements from a large space to simplify the elements into subsets, or granules [57, 58]. Granular computing is an effective approach to define uncertainty of objects in the search space as it reduces large objects to a smaller search space [59].

Feature selection is a conventional approach to handle big data with the purpose of choosing a subset of relative features for an aggregate but more precise data representation [60, 61]. Feature selection is a very useful strategy in data mining for preparing high-scale data [60].

Instance selection is practical in many ML or data mining tasks as a major feature in data pre-processing. By utilizing instance selection, it is possible to reduce training sets and runtime in the classification or training phases [62].

The costs of uncertainty (both monetarily and computationally) and challenges in generating effective models for uncertainties in big data analytics have become key to obtaining robust and performant systems. As such, we examine several open issues of the impacts of uncertainty on big data analytics in the next section.

Uncertainty perspective of big data analytics

This section examines the impact of uncertainty on three AI techniques for big data analytics. Specifically, we focus on ML, NLP, and CI, although many other analytics techniques exist. For each presented technique, we examine the inherent uncertainties and discuss methods and strategies for their mitigation.

Machine learning and big data

When dealing with data analytics, ML is generally used to create models for prediction and knowledge discovery to enable data-driven decision-making. Traditional ML methods are not computationally efficient or scalable enough to handle both the characteristics of big data (e.g., large volumes, high speeds, varying types, low value density, incompleteness) and uncertainty (e.g., biased training data, unexpected data types, etc.). Several commonly used advanced ML techniques proposed for big data analysis include feature learning, deep learning, transfer learning, distributed learning, and active learning. *Feature learning* includes a set of techniques that enables a system to automatically discover the representations needed for feature detection or classification from raw data. The performances of the ML algorithms are strongly influenced by the selection of data representation. *Deep learning* algorithms are designed for analyzing and extracting valuable knowledge from massive amounts of data and data collected from various sources (e.g., separate variations within an image, such as a light, various materials, and shapes) [56], however current deep learning models incur a high computational cost. *Distributed learning* can be used to mitigate the scalability problem of traditional ML by carrying out calculations on data sets distributed among several workstations to scale up the learning process [63]. *Transfer learning* is the ability to apply knowledge learned in one context to new contexts, effectively improving a learner from one domain by transferring information from a related domain [64]. *Active learning* refers to algorithms that employ adaptive data collection [65] (i.e., processes that automatically adjust parameters to collect the most useful data as quickly as possible) in order to accelerate ML activities and overcome labeling problems. The uncertainty challenges of ML techniques can be mainly attributed to learning from data with low veracity (i.e., uncertain and incomplete data) and data with low value (i.e., unrelated to the current problem). We found that, among the ML techniques, active learning, deep learning, and fuzzy logic theory are uniquely suited to support the challenge of reducing uncertainty, as shown in Fig. 3. Uncertainty can impact ML in terms of incomplete or imprecise training samples, unclear classification boundaries, and rough knowledge of the target data. In some cases, the data is represented without labels, which can become a challenge. Manually labeling large data collections can be an expensive and strenuous task, yet learning from unlabeled data is very difficult as classifying data with unclear guidelines yields unclear results. Active learning has solved this issue by selecting a subset of the most important instances for labeling [65, 66]. Deep learning is another learning method that can handle incompleteness and inconsistency issues in the classification procedure [15]. Fuzzy logic theory has been also shown to model uncertainty efficiently. For example, in fuzzy support vector machines (FSVMs), a fuzzy membership is applied to each input point of the support vector machines (SVM). The learning procedure then has the benefits of

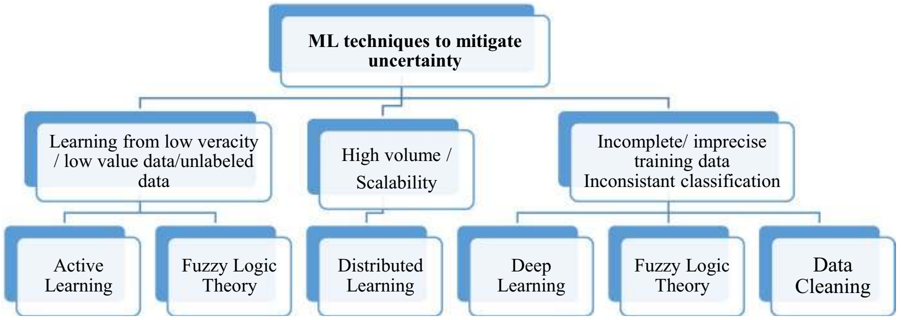
ML techniques to mitigate uncertainty
Learning from low veracity / low value data/unlabeled data
High volume / Scalability
Incomplete/ imprecise training data Inconsistant classification
Active Learning
Fuzzy Logic Theory
Distributed Learning
Deep Learning
Fuzzy Logic Theory
Data Cleaning

Fig. 3 How ML techniques handle uncertainty in big data

flexibility provided by fuzzy logic, enabling an improvement in the SVM by decreasing the result of noises in data points [67]. Hence, while uncertainty is a notable problem for ML algorithms, incorporating effective techniques for measuring and modeling uncertainty can lead towards systems that are more flexible and efficient, respective.

Natural language processing and big data

NLP is a technique grounded in ML that enables devices to analyze, interpret, and even generate text [8]. NLP and big data analytics tackle huge amounts of text data and can derive value from such a dataset in real-time [68]. Some common NLP methods include lexical acquisition (i.e., obtains information about the lexical units of a language), word sense disambiguation (i.e., determining which sense of the word is used in a sentence when a word has multiple meanings), and part-of-speech (POS) tagging (i.e., determining the function of the words through labeling categories such as verb, noun, etc.). Several NLP-based techniques have been applied to text mining including information extraction, topic modeling, text summarization, classification, clustering, question answering, and opinion mining [8]. For example, financial and fraud investigations may involve finding evidence of a crime in massive datasets. NLP techniques (particularly named entity extraction and information retrieval) can help manage and sift through huge amounts of textual information, such as criminal names and bank records, to support fraud investigations. Moreover, NLP techniques can help to create new traceability links and recover traceability links (i.e., missing or broken links at run-time) by finding semantic similarity among available textual artifacts [69]. Furthermore, NLP and big data can be used to analyze news articles and predict rises and falls on the composite stock price index [68].

Uncertainty impacts NLP in big data in a variety of ways. For example, keyword search is a classic approach in text mining that is used to handle large amounts of textual data. Keyword search accepts as input a list of relevant words or phrases and searches the desired set of data (e.g., a document or database) for occurrences of the relevant words (i.e., search terms). Uncertainty can impact keyword search, as a document that contains a keyword is not an assurance of a document's relevance. For example, a keyword search usually matches exact strings and ignores words with spelling errors that may still be relevant. Boolean operators and fuzzy search technologies permit greater flexibility in that they can be used to search for words similar to the desired spelling [70]. Although keyword or key phrase search is useful, limited sets of search terms can miss key information. In comparison, using a wider set of search terms can result in a large set of ‘hits’ that can contain large numbers of irrelevant false positives [71]. Another example of uncertainty impacting NLP involves automatic POS taggers that must handle the ambiguity of certain words (Fig. 4) (e.g., the word “bimonthly” can mean twice a month or every two months depending on the context, the word “quite” having different meaning to American and British audiences, etc.), as well as classification problems due to the ambiguity of periods (‘.’) that can be interpreted as part of a token (e.g., abbreviation), punctuation (e.g., full stop), or both [72, 73]. Although recent research indicates that using IBM Content Analytics (ICA) can mitigate these problems, there remains the open issue in this topic regarding large-scale data [73]. Also, uncertainty and ambiguity impact the POS tagging especially when using biomedical language, which quite different from general English. It has been reported uncertainty and not sufficient tagging accuracy when trained taggers from Treebank corpus and applied to biomedical data [74]. To this end, stream processing systems deal with high data throughput while achieving low response latencies. The integration of NLP techniques with the help of uncertainty modeling such as fuzzy and probabilistic sets with big data analytics may offer the ability to support handling big textual data in real time, however additional work is necessary in this area.

Computational intelligence and big data

CI includes a set of nature-inspired computational techniques that play an important role in big data analysis [75]. CIs have been used to tackle complicated data processes and analytics challenges such as high complexity, uncertainty, and any processes where traditional techniques are not sufficient. Common techniques that are currently available in CI are evolutionary algorithms (EAs), artificial neural networks (ANN), and fuzzy logic [76], with examples spanning search-based problems such as parameter optimization to optimizing a robot controller.

CI techniques are suitable for dealing with the real-world challenges of big data as they are fundamentally capable of handling numerous amounts of uncertainty. For example, generating models for predicting emotions of users is one problem with many potential pitfalls for uncertainty. Such models deal with large databases of information relating to human emotion and its inherent fuzziness [77]. Many challenges

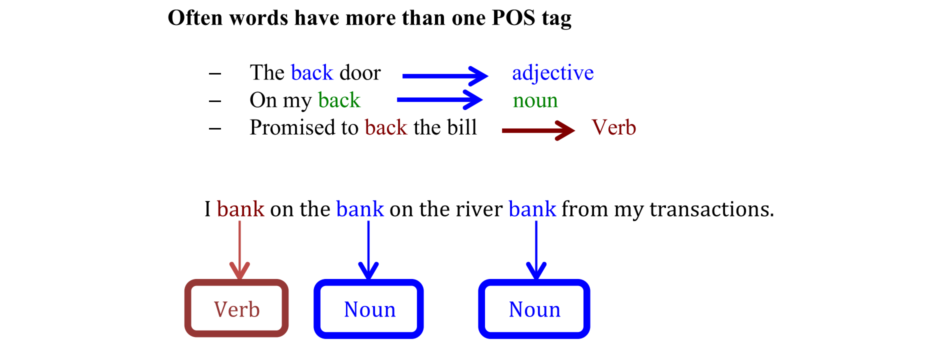
Often words have more than one POS tag
– The back door → adjective
– On my back → noun
– Promised to back the bill → Verb
I bank on the bank on the river bank from my transactions.
Verb
Noun
Noun

Fig. 4 Words with more than one POS tag (ambiguity)

still exist in current CI techniques, especially when dealing with the value and veracity characteristics of big data. Accordingly, there is great interest in developing new CI techniques that can efficiently address massive amounts of data and to have the ability to quickly respond to modifications in the dataset [78]. As reported by [78], big data analysis can be optimized by employing algorithms such as swarm intelligence, AI, and ML. These techniques are used for training machines in performing predictive analysis tasks, collaborative filtering, and building empirical statistical predictive models. It is possible to minimize the complexity and uncertainty on processing massive volumes of data and improve analysis results by using CI-based big data analytics solutions.

To support CI, fuzzy logic provides an approach for approximate reasoning and modeling of qualitative data for uncertainty challenges in big data analytics [76, 79, 80] using linguistic quantifiers (i.e., fuzzy sets). It represents uncertain real-word and user-defined concepts and interpretable fuzzy rules that can be used for inference and decision-making. Big data analytics also bear challenges due to the existence of noise in data where the data consists of high degrees of uncertainty and outlier artifacts. Iqbal et al. [76] have demonstrated that fuzzy logic systems can efficiently handle inherent uncertainties related to the data. In another study, fuzzy logic-based matching algorithms and MapReduce were used to perform big data analytics for clinical decision support. The developed system demonstrated great flexibility and could handle data from various sources [81]. Another useful CI technique for tackling the challenges of big data analytics are EAs that discover the optimal solution(s) to a complex problem by mimicking the evolution process by gradually developing a population of candidate solutions [73]. Since big data includes high volume, variety, and low veracity, EAs are excellent tools for analyzing such datasets [82]. For example, applying parallel genetic algorithms to medical image processing yields an effective result in a system using Hadoop [83]. However, the results of CI-based algorithms may be impacted by motion, noise, and unexpected environments. Moreover, an algorithm that can deal with one of these problems may function poorly when impacted by multiple factors [79].

Summary of mitigation strategies

This paper has reviewed numerous techniques on big data analytics and the impact of uncertainty of each technique. Table 2 summarizes these findings. First, each AI technique is categorized as either ML, NLP, or CI. The second column illustrates how uncertainty impacts each technique, both in terms of uncertainty in the data and the technique itself. Finally, the third column summarizes proposed mitigation strategies for each uncertainty challenge. For example, the first row of Table 2 illustrates one possibility for uncertainty to be introduced in ML via incomplete training data. One approach to overcome this specific form of uncertainty is to use an active learning technique that uses a subset of the data chosen to be the most significant, thereby countering the problem of limited available training data.

Note that we explained each big data characteristic separately. However, combining one or more big data characteristics will incur exponentially more uncertainty, thus requiring even further study.

Table 2 Uncertainty mitigation strategies

Artificial intelligence	Uncertainty	Mitigation
Machine learning	Incomplete training samples Inconsistent classification Learning from low veracity and noisy data	Active learning [65, 66], Deep learning [15, 63], Fuzzy sets [67], Feature selection [9, 60, 61]
	Learning from unlabeled data	Active learning [65, 66]
	Scalability	Distributed learning [12, 63] Deep learning [56]
Natural language processing	Keyword search	Fuzzy, Bayesian [68, 70, 71]
	Ambiguity of words in POS	ICA [73], LIBLINEAR and MNB algorithm [68]
	Classification (simplifying language assumption)	ICA [73], Open issue [68]
Computational intelligence	Low veracity, complex and noisy data	Fuzzy logic, EA [76, 79, 80, 82]
	High volume, variety	Swarm intelligence, EA [78, 81, 82], Fuzzy-logic based matching algorithm, EA [81, 82]

Discussion

This paper has discussed how uncertainty can impact big data, both in terms of analytics and the dataset itself. Our aim was to discuss the state of the art with respect to big data analytics techniques, how uncertainty can negatively impact such techniques, and examine the open issues that remain. For each common technique, we have summarized relevant research to aid others in this community when developing their own techniques. We have discussed the issues surrounding the five V's of big data, however many other V's exist. In terms of existing research, much focus has been provided on volume, variety, velocity, and veracity of data, with less available work in value (e.g., data related to corporate interests and decision making in specific domains).

Future research directions

This paper has uncovered many avenues for future work in this field. First, additional study must be performed on the interactions between each big data characteristic, as they do not exist separately but naturally interact in the real world. Second, the scalability and efficacy of existing analytics techniques being applied to big data must be empirically examined. Third, 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. Fourth, more work is necessary on how to efficiently model uncertainty in ML and NLP, as well as how to represent uncertainty resulting from big data analytics. Fifth, since the CI algorithms are able to find an approximate solution within a reasonable time, they have been used to tackle ML problems and uncertainty challenges in data analytics and process in recent years. However, there is a lack of CI metaheuristics algorithms to apply to big data analytics for mitigating uncertainty.

Abbreviations

IoT: Internet of things; PB: petabytes; ZB: zettabyte; EB: exabyte; IDC: international data corporation; AI: artificial intelligence; ML: machine learning; NLP: natural language processing; CI: computational intelligence; FSVM: fuzzy support vector machines; SVM: support vector machines; POS: part-of-speech; ICA: IBM content analytics; EAs: evolutionary algorithms; ANN: artificial neural networks.

Acknowledgements

The authors would like to thank Rana H. Hariri for her assistance with this paper.

This research has been supported in part by NSF Grant CNS-1657061, the Michigan Space Grant Consortium, the Comcast Innovation Fund, and Oakland University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Oakland University or other research sponsors.

Authors' contributions

RHH proposed the idea of the survey, performed the literature review, analysis for the work, and wrote the manuscript. EMF supervised and provided technical guidance throughout the research and writing phases of the manuscript. KMB assisted with editing the manuscript. All authors read and approved the final manuscript.

Funding

Funding was provided by National Science Foundation (Grant No.CNS-1657061), Arizona Space Grant Consortium, Comcast Innovation, Oakland University.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 9 March 2019 Accepted: 20 May 2019

Published online: 04 June 2019

References

1. Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. Comput Sci Inf Technol (CS & IT). 2014;4:131–40.
2. Marr B. Forbes. How much data do we create every day? 2018. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#4146a89b60ba.
3. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big data: the management revolution. Harvard Bus Rev. 2012;90(10):60–8.
4. Zephoria. Digital Marketing. The top 20 valuable Facebook statistics—updated November 2018. 2018. https://zephoria.com/top-15-valuable-facebook-statistics/.
5. Iafrate F. A journey from big data to smart data. In: Digital enterprise design and management. Cham: Springer; p. 25–33. 2014.
6. Lenk A, Bonorden L, Hellmanns A, Roedder N, Jaehnichen S. Towards a taxonomy of standards in smart data. In: IEEE international conference on big data (Big Data), 2015. Piscataway: IEEE. p. 1749–54. 2015.
7. Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2(1):21.
8. Chen M, Mao S, Liu Y. Big data: a survey. Mobile Netw Appl. 2014;19(2):171–209.
9. Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.
10. Borne K. Top 10 big data challenges a serious look at 10 big data v's. Recuperat de. 2014. https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs. Accessed 11 Apr 2014.
11. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: the next frontier for innovation, competition, and productivity. 2011.
12. Pouyanfar S, Yang Y, Chen SC, Shyu ML, Iyengar SS. Multimedia big data analytics: a survey. ACM Comput Surv (CSUR). 2018;51(1):10.
13. Cimaglobal. Using big data to reduce uncertainty in decision making. 2015. http://www.cimaglobal.com/Pages-that-we-will-need-to-bring-back/velocity-archive/Student-e-magazine/Velocity-December-2015/P2-using-big-data-to-reduce-uncertainty-in-decision-making/.
14. Maugis PA. Big data uncertainties. J Forensic Legal Med. 2018;57:7–11.
15. 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.
16. 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.
17. Villars RL, Olofson CW, Eastwood M. Big data: what it is and why you should care. White Paper IDC. 2011;14:1–14.
18. Laney D. 3D data management: controlling data volume, velocity and variety. META Group Res Note. 2001;6(70):1.
19. Gantz J, Reinsel D. Extracting value from chaos. IDC iview. 2011;1142(2011):1–12.
20. Jain A. The 5 Vs of big data. IBM Watson Health Perspectives. 2017. https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/. Accessed 30 May 2017.
21. IBM big data and analytics hub. Extracting Business Value from the 4 V's of Big Data. 2016. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data.
22. Snow D. Dwaine Snow's thoughts on databases and data management. 2012.
23. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage. 2015;35(2):137–44.
24. Vajjhala NR, Strang KD, Sun Z. Statistical modeling and visualizing open big data using a terrorism case study. In: 3rd international conference on future Internet of things and cloud (FiCloud), 2015. IEEE. p. 489–96. 2015.

25. Marr B. Really big data at Walmart: real-time insights from their 40+ Petabyte data cloud. 2017. https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/#2a0c16916c10.
26. Pokorný J, Škoda P, Zelinka I, Bednárek D, Zavoral F, Kruliš M, Šaloun P. Big data movement: a challenge in data processing. In: Big Data in complex systems. Cham: Springer; p. 29–69. 2015
27. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011.
28. Xiong H, Pandey G, Steinbach M, Kumar V. Enhancing data analysis with noise removal. IEEE Trans Knowl Data Eng. 2006;18(3):304–19.
29. Court D. Getting big impact from big data. McKinsey Q. 2015;1:52–60.
30. Knight FH. Risk, uncertainty and profit, library of economics and liberty. 1921. (Retrieved May 17 2011).
31. DeLine R. Research opportunities for the big data era of software engineering. In: Proceedings of the first international workshop on BIG Data software engineering. Piscataway: IEEE Press; p. 26–9. 2015.
32. IBM Think Leaders. (2014). Veracity of data for marketing: Step-by-step. https://www.ibm.com/blogs/insights-on-business/ibmix/veracity-of-data-for-marketing-step-by-step/.
33. Wang XZ, Ashfaq RAR, Fu AM. Fuzziness based sample categorization for classifier performance improvement. J Intell Fuzzy Syst. 2015;29(3):1185–96.
34. Wang Xizhao, Huang JZ, Wang X, Huang JZ. Editorial: uncertainty in learning from big data. Fuzzy Sets Syst. 2015;258(1):1–4.
35. Xu ZB, Liang JY, Dang CY, Chin KS. Inclusion degree: a perspective on measures for rough set data analysis. Inf Sci. 2002;141(3–4):227–36.
36. López V, del Río S, Benítez JM, Herrera F. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets Syst. 2015;258:5–38.
37. Bernardo JM, Smith AF. Bayesian theory, vol. 405. Hoboken: Wiley; 2009.
38. Cuzzolin F. (Ed.). Belief functions: theory and applications. Berlin: Springer International Publishing; 2014.
39. Brown DG. Classification and boundary vagueness in mapping presettlement forest types. Int J Geogr Inf Sci. 1998;12(2):105–29.
40. Correa CD, Chan YH, Ma KL. A framework for uncertainty-aware visual analytics. In: IEEE symposium on visual analytics science and technology, VAST 2009. Piscataway: IEEE; p. 51–8. 2009.
41. Zadeh LA. Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. J Stat Plann Inference. 2002;105(2002):233–64.
42. Zadeh LA. Toward a generalized theory of uncertainty (GTU)-an outline. Inf Sci. 2005;172(1–2):1–40.
43. Özkan I, Türkşen IB. Uncertainty and fuzzy decisions. In: Chaos theory in politics. Dordrecht: Springer; p. 17–27. 2014.
44. Lesne A. Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. Math Struct Comput Sci. 2014;24(3).
45. Vajapeyam S. Understanding Shannon's entropy metric for information. 2014. arXiv preprint arXiv:1405.2061.
46. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423.
47. Pawlak Z. Rough sets. Int J Comput Inform Sci. 1982;11(5):341–56.
48. Rissino S, Lambert-Torres G. Rough set theory—fundamental concepts, principals, data extraction, and applications. In: Data mining and knowledge discovery in real life applications. New York: InTech; 2009.
49. Tavana M, Liu W, Elmore P, Petry FE, Bourgeois BS. A practical taxonomy of methods and literature for managing uncertain spatial data in geographic information systems. Measurement. 2016;81:123–62.
50. Salahdine F, Kaabouch N, El Ghazi H. Techniques for dealing with uncertainty in cognitive radio networks. In: 2017 IEEE 7th annual computing and communication workshop and conference (CCWC). Piscataway: IEEE. p. 1–6. 2017.
51. Düntsch I, Gediga G. Rough set dependency analysis in evaluation studies: an application in the study of repeated heart attacks. Inf Res Rep. 1995;10:25–30.
52. Golchha N. Big data—the information revolution. IJAR. 2015;1(12):791–4.
53. Khan M, Ayyoob M. Big data analytics evaluation. Int J Eng Res Comput Sci Eng (IJERCSE). 2018;5(2):25–8.
54. 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.
55. Wang XZ, Dong LC, Yan JH. Maximum ambiguity-based sample selection in fuzzy decision tree induction. IEEE Trans Knowl Data Eng. 2012;24(8):1491–505.
56. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2(1):1.
57. Bargiela A, Pedrycz W. Granular computing. In: Handbook on computational intelligence. Fuzzy logic, systems, artificial neural networks, and learning systems, vol 1, p. 43–66. 2016.
58. Kacprzyk J, Filev D, Beliakov G. (Eds.). Granular, Soft and fuzzy approaches for intelligent systems: dedicated to Professor Ronald R. Yager (Vol. 344). Berlin: Springer; 2016.
59. Yager RR. Decision making under measure-based granular uncertainty. Granular Comput. 1–9. 2018.
60. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1–3):389–422.
61. Liu H, Motoda H. (Eds.). Computational methods of feature selection. Boca Raton: CRC Press; 2007.
62. Olvera-López JA, Carrasco-Ochoa JA, Martínez-Trinidad JF, Kittler J. A review of instance selection methods. Artif Intell Rev. 2010;34(2):133–43.
63. 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;2016(1):67.
64. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016;3(1):9.
65. Athmaja S, Hanumanthappa M, Kavitha V. A survey of machine learning algorithms for big data analytics. In: International conference on innovations in information, embedded and communication systems (ICIIECS), 2017. Piscataway: IEEE; p. 1–4. 2017.
66. Fu Y, Li B, Zhu X, Zhang C. Active learning without knowing individual instance labels: a pairwise label homogeneity query approach. IEEE Trans Knowl Data Eng. 2014;26(4):808–22.

67. Lin CF, Wang SD. Fuzzy support vector machines. IEEE Trans Neural Netw. 2002;13(2):464–71.
68. Wang L, Wang G, Alexander CA. Natural language processing systems and Big Data analytics. Int J Comput Syst Eng. 2015;2(2):76–84.
69. Hariri RH, Fredericks EM. Towards traceability link recovery for self-adaptive systems. In: Workshops at the thirty-second AAAI conference on artificial intelligence. 2018.
70. Crabb ES. "Time for some traffic problems": enhancing e-discovery and big data processing tools with linguistic methods for deception detection. J Digit Forensics Secur Law. 2014;9(2):14.
71. Khan E. Addressing bioinformatics big data problems using natural language processing: help advancing scientific discovery and biomedical research. In: Buzatu C, editor. Modern computer applications in science and education. 2014; p. 221–8.
72. Clark A, Fox C, Lappin S. (Eds.). The handbook of computational linguistics and natural language processing. Hoboken: Wiley; 2013.
73. Holzinger A, Stocker C, Ofner B, Prohaska G, Brabenetz A, Hofmann-Wellenhof R. Combining HCI, natural language processing, and knowledge discovery-potential of IBM content analytics as an assistive technology in the biomedical field. In: Human-Computer Interaction and knowledge discovery in complex, unstructured, big data. Berlin, Heidelberg: Springer; p. 13–24. 2013.
74. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, Tsujii J. Developing a robust part-of-speech tagger for biomedical text. In: 10th Panhellenic conference on informatics Volos: Springer; 2005. p. 382–92.
75. Fulcher J. Computational intelligence: an introduction. In: Computational intelligence: a compendium. Berlin, Heidelberg: Springer; p. 3–78. 2008.
76. Iqbal R, Doctor F, More B, Mahmud S, Yousuf U. Big data analytics: computational intelligence techniques and application areas. Technol Forecast Soc Change. 2018. https://doi.org/10.1016/j.techfore.2018.03.024.
77. Wu D. Fuzzy sets and systems in building closed-loop affective computing systems for human-computer interaction: advances and new research directions. In: IEEE international conference on fuzzy systems (FUZZ-IEEE), 2012. IEEE. p. 1–8. 2012.
78. Gupta A. Big data analysis using computational intelligence and Hadoop: a study. In: 2nd international conference on computing for sustainable global development (INDIACom), 2015. Piscataway: IEEE; p. 1397–1401. 2015.
79. Doctor F, Syue CH, Liu YX, Shieh JS, Iqbal R. Type-2 fuzzy sets applied to multivariable self-organizing fuzzy logic controllers for regulating anesthesia. Appl Soft Comput. 2016;38:872–89.
80. Zadeh LA. Fuzzy sets. Inf Control. 1965;8(3):338–53.
81. Duggal R, Khatri SK, Shukla B. Improving patient matching: single patient view for clinical decision support using big data analytics. In: 4th International conference on reliability, infocom technologies and optimization (ICRITO) (trends and future directions), 2015. Piscataway: IEEE; p. 1–6. 2015.
82. Bhattacharya M, Islam R, Abawajy J. Evolutionary optimization: a big data perspective. J Netw Comput Appl. 2016;59:416–26.
83. Augustine DP. Enhancing the efficiency of parallel genetic algorithms for medical image processing with Hadoop. Int J Comput Appl. 2014;108(17):11–6.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.