Uncertainty handling in big data using Fuzzy logic - Literature Review

Dyari M. Ameen M. Shareef and Sadegh Abollah Aminifar

February 3, 2021

A review of uncertainty handling in big data using Fuzzy logic

Dyari M.Ameen M.Shareef M.Shareef[1], Sadegh Abollah Aminifar[2]

1, corresponding author (dma140h@cs.soran.edu.iq),

2, (sadegh.aminifar@soran.edu.iq),

1,2 Computer Science Department - Faculty of Science - Soran University, Soran, Kurdistan, Iraq

Abstract:

Advances in technology have gained wide attention from both academia and industry as Big Data plays a ubiquities and non-trivial role in the Data Analytical problems. Big Data analysis involves different types of uncertainty, and part of the uncertainty can be handled or at least reduced by fuzzy logic. We have reviewed a number of papers in detail, that have been published in the last decade, to identify the very recent and significant advancements including the breakthroughs in the field. The vast majority of papers, most of the time, came up with methods that are less computational than the current methods that are available in the market. The proposed methods were better in terms of efficacy, cost-effectiveness and sensitivity. Despite the existence of some works in the role of fuzzy logic in handling uncertainty, we have observed that few works have been done regarding how significantly uncertainty can impact the integrity and accuracy of big data.

Keywords: Big data; Fuzzy Logic; Uncertainty handling; Data Analytics

1. Introduction

According to (Marr, n.d.), 2.5 quintillion bytes of data was produced in 2018 and 90% of all data was generated in the last 2 years, and Google alone processes over 40,000 searches each second. People on Facebook publish 300 million posts, 510,000 comments each day. No to mention that as of September 30, 2020, Facebook has 56,653 employees (Noyes, 2018; Marr, n.d.).

As the modern-day technology is in progress and as the list of emerging technologies is largely growing regularly, the available data indescribably and significantly is increasing. With having said that internet applications are generating mammoth amounts of data at unprecedented low costs, such as live streams, or like the data that can be generated by any user on the network via common sources like google forms. Moreover, the majority of sources that we are using nowadays, comprises huge amounts of data, and so in a variety of data formats that come into being at various velocities. Not to mention that the objective, in the policy of the majority of huge companies, is to turn big data sets into big data assets (Raghava, et al., 2014).

Data can be organized and complete, like the data in a database management system, with the utility of having a set of computational means like computing mean, sum, etc. However, over the years the trend has changed and the direction of organized and complete data is continuously being geared towards unorganized and incomplete data. The data used to be organized and stored at a database management system like SQL Server but now it is being stored at various places and at fast speeds, which is mostly unorganized (Amit & Pranab, 2019). Big data comes in different formats, forms and sizes caused by the influence it has on almost all fields including science, technology, medicine, public health, economics, business, linguistics and social science and thus leads many industrial people to seek new ways to increase their revenue. (Fokoué, 2015) proposed that if n is the scale of available data, then, any problem with more than 50 features can be categorized as Big data. As for the most real-world applications, there is the availability of huge datasets that makes the data analysis convenient. However, just because there is a lot of data out there, that necessarily doesn’t mean that you are in a good shape, since one of the overwhelming challenges that hampers huge companies is the problem of turning their growing data into information (Berthold, et al., n.d.).

Data by itself, regardless of how much it is, is not enough because general patterns, structures, and segmentations cannot be detected in it, and it is very often those patterns that are required for a one if desires to get more insights about the data, typically unstructured, to figure out how to increase, for example, his/her revenue. Unfortunately, it's easier said than done to discover such patterns as the current analytical tools are not sufficient to retrieve the information required from such overwhelming amounts of data (Berthold, et al., n.d.).

As there are continues changes and variations in data over time, there is a large possibility that unwanted noise my happen. Not only noise, but there could also be some instances where the data is incomplete, missing or corrupt. Not to mention that in real-world situations, many factors indicate uncertainty, like measurement errors, noisy environments and/or randomness in data gathering. As a result, the mathematical formulations, implementations and modelling of uncertainty can be very effective in many real-world applications (Amit & Pranab, 2019) (Amit, et al., 2020).

To tackle these uncertainties in the data, fuzzy sets come to our assistance. Fuzzy sets help in the decision-making processes to detect the uncertainties in mathematical order. Moreover, the T1 FSs [1]are being used often by researchers to model the uncertainty in the data which gives them the flexibility to model the uncertain parameters. However, T1 FSs' MFs [2]are crisp in nature which means they cannot directly model the uncertainties and so arises interpretability issues. In other words, when uncertainty is generating from more than one source, T1 FSs cannot efficiently represent them since their values are crisp in nature. This was resolved by the appearance of T2 FSs[3]. T2 FSs have MFs which are also fuzzy and thus can help to model the uncertainties that come from multiple sources (Amit & Pranab, 2019).

In this paper, we provide a review of uncertainty handling in big data using Fuzzy logic, that have been applied mostly in the last decade. The review took us roughly two months to finish since it was a requirement mentioned in the regulations defined by the ministry of higher education for master degree programmes. Besides, the rest of the paper is put in order as follows: in the next section, basic definitions, for the mentioned concepts, are presented. In the section after that, the state of art studies is presented. In the final section, the conclusion is presented.

[1] T1 FS stands for Type-1 fuzzy sets where each element of the feasible domain has a membership degree of in between 0 and 1.

[2] MF stands for Membership function which is a function to evaluate the association of a value to a set.

[3] T2 FS stands for Type-2 fuzzy sets in which even elements are fuzzy.

2. Background

In this section, the main concepts that are used in this paper or required to be understood to better benefit from this paper, are discussed.

2.1 Big Data

(Laney, 2001), an Industry analyst, was first to come up with a three-characteristic for the definition of big data to which he called them the three V's (3 V's). He theorized that Volume, Variety and velocity are the characteristics of big data. The other characteristics of big data which were coined in subsequent years are veracity, variability and value. Thus, collectively, there are six characteristics of big data (Hsinchun, et al., 2012; Zikopoulos, et al., 2013).

The six characteristics of big data are briefly described below:

- Volume: it is the size or scale of data. With time, the volume of the data is on the increase.
- Variety: As the name implies, it refers to the various or, say, different types of formats of the data.
- Velocity: it is the sheer rate at which data is coming in, typically at a specific point of time.
- Veracity: it is defined by the IBM, the company which coined the term, as the level of ambiguity.
- Variability: It works incorporation with the characteristic velocity. It is defined by SAS which refers to the variations in the data flow rates.
- Value: It is considered to be almost low-value density.

2.2 Fuzzy Logic

Lotfi Zadeh invented fuzzy logic as he observed that we human beings, unlike computers, have possibilities in between Yes and No, such as EXTREMELY YES, POSSIBLY YES, NOT SURE, ALMOST NO, EXTREMELY NO. Unlike the conventional logic block that computers understand, which takes exact input and outputs a definite response such as zero or one, True or False, etc., which is equivalent to human beings' Yes or No, fuzzy logic produces possibilities in between Yes and No. So, fuzzy logic is a computing-based approach to reasoning that mimics human reasoning. Fuzzy Logic Systems (FLS) reaches a satisfying but definite output in response to incomplete and distorted input. Fuzzy logic is an inevitable tool for a wide figure of various applications that ranges from the control of engineering systems to artificial intelligence. The approach of Fuzzy Logic resembles the way of decision making in human beings that has intermediate levels in between the digital values' YES and No. Fuzzy logic can come to our assistance in terms of commercial and practical purposes like it can have machines under control, it may give inaccurate but acceptable reasoning and it also helps to deal with the uncertainty in engineering. Moreover, it can be implemented in both software and hardware and also in systems where there are various sizes and capabilities like workstation-based control systems (Zadeh, 1988; G & B, 1995).

2.2.1 The architecture of Fuzzy Logic Systems

There are four main parts to Fuzzy Logic Systems, and they are as follows:

Fuzzification Module: Fuzzification is the process of fuzzifying a crisp entity. Here, the crisp input is converted into linguistic variables using membership functions that are stored in the fuzzy knowledge base.

Knowledge base (Intelligence): It stores IF-THEN rules that are written by experts.

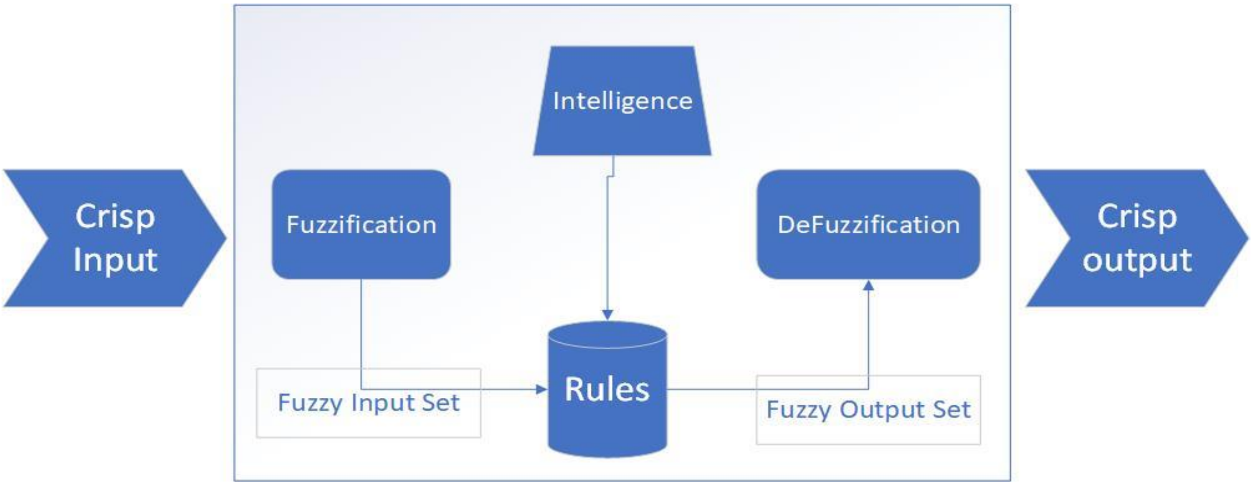

Fig. 1 The schema of Fuzzy Systems

Inference Engine: It stimulates the human reasoning process by applying fuzzy inference on the inputs and IF-THEN rules.

Defuzzification Module: It is the process of de-fuzzifying a fuzzified entity. It transforms the fuzzified set obtained by the inference engine into a crisp value.

2.2.2 Fuzzy Sets

Fuzzy sets are sets in which each and every member has a membership degree of in between 0 and 1, and they can be defined with tilde character (~). In its theory, there exists partial membership since an element of a fuzzy set can also be an element of other sets in the same universe discourse. For example, in the era of biology and in microarray data, one gene can belong to many different clusters/groups. However, in classical sets, that is not possible since one element can only belong to one cluster/groups. So, in here, a gene expression dataset can only be represented with fuzzy sets and that is the very advantage of it (Amit & Pranab, 2019).

2.2.3 Membership Functions

Membership functions are graphical representations of fuzzy sets which allow you to quantify linguistic terms and represent them. A membership function for a fuzzy set S on the universe of discourse [4]X is defined $\mu_A:X \rightarrow [0,1]$. Here, each element of X has a value of in between 0 and 1 which is called membership degree. It quantifies the degree of membership of each and every possible element in X to the fuzzy set A. There are different membership function shapes like triangular, trapezoidal, singleton and Gaussian with the triangular membership function shapes being the most common ones. In addition to that, they all are different from each other by their geometric shapes and structure (Nozer & Jane, 2004).

2.2.4 Uncertainty

“Uncertainty is a situation which involves unknown or imperfect information”, says (Knight, 2011). Uncertainty is everywhere ranging from big data learning to many different sources. Moreover, it has many types that negatively impact the effectiveness and accuracy of the results of big data analytics. For example, if training data is biased towards anything or any dimension, the learning algorithm will still use the biased data and predict the new data points based on that trained data (Reihaneh, et al., 2019).

[4] The universe of discourse is the all-possible values in fuzzy problem. It is the totality of values.

3. State of art studies

In this section, several works regarding uncertainty handling in big data using fuzzy logic, briefly are discussed.

(Amit, et al., 2020) proposed a novel approach that can handle the uncertainty of big data using the Footprint of Uncertainty (FOU) in Interval type-2 fuzzy sets. The proposed method provides the same behaviour of a cluster centroid which is the de-fuzzified value of the type-2 fuzzy sets with having fewer instances and fewer computations. The authors divided the data into clusters by k-means algorithm and then took their cluster centroids and FOU defuzzied centroids. Finally, they used SVR to compare both centroid types. As a consequence, the proposed method was more scalable, as in the experimental section, they observed that at the initial phase, almost all the cluster centroids and de-fuzzified centroids overlapped but when they added millions of instances, they saw the cluster centroids shifted towards the de-fuzzified centroids. The proposed method didn’t only handle uncertainty with fewer costs, but also provided better results at accommodating new data points.

Data comes from everywhere including the giant datasets of the social network. (Raghava, et al., 2014) proposed a novel computational paradigm on the analysis of social network in which there could be a lot of uncertainty. The authors analyzed the sentiment classifications by constructing both the crisp as well as the fuzzy sets of the artefacts[5]. Since user interactions in current social media platforms and applications are huge sources for big data, the Authors used Facebook data by extracting it from the Social Data Analytics Tool (SODATO), to define operations. Finally, they articulated a formal model with the help of fuzzy set theory, for the benefit of taking care of uncertainty, such that the sentiments had negative, positive and neural with a degree.

Even though the SQL is a powerful tool, it cannot satisfy the needs for data selection based on linguistic terms. (Hudec & Vujoševic´, 2012) proposed linguistic terms for database queries and showed the advantages of using linguistics terms, and the distinguishes between classical and fuzzy approaches. The methodology used the fuzzy set theory to reduce uncertainty and thus led them to gear towards the integration of data selection and data classification into one signal entity

[5] In here, refer to posts, comments, likes and shares.

while the Relational databases' accessibility remains untouched. However, the proposed method is sufficient only in the case of two-valued logics not in many-valued logics.

(Mehta, et al., 2009) proposed a method to fuzzy classification as knowledge representation can have uncertainty like noisy data and needs linguistic terms instead of discrete values in a real-world application. The novelty of the proposed method is its discretization on input data followed by classification as it allows the input data representation in linguistic form and lets you do fuzzy classification which is a natural way of classifying the data. The authors believed that the results can be represented in linguistic terms by using fuzzy discretization which is better than other classification techniques as they only perform crisp classification. However, The proposed method cannot be applied for data mining in every case.

(Amit & Pranab, 2019) proposed a method to model the uncertainty in gene expression datasets using Interval type-2 fuzzy uncertainty modelling, symmetrically and asymmetrically. Authors first turned the data into IT2 fuzzified data, next, they de-fuzzified the data and clustered it using C-means algorithm. Finally, they used four validity measure to validate the process's accuracy. They used a 64bit MATLAB and the idea of parallel programming to process a dataset of 14 cancer gene expression which had 16,063 genes and 54 test samples. As a result, and on average, they observed as the FOU spread, which was the uncertainty band, increased, Partition coefficient values went higher with most clusters. Not to mention that the sensitivity and scalability were also tested and the results were positive.

(Pendharkar, 2012) proposed a technique to perform fuzzy classification by learning fuzzy membership functions in which Data Development Analysis using GMAT and GPA was implemented. They highlighted the sources of uncertainty followed by summarizing prior work for solving classification problems. The novelty of this technique is that no expert participation is required for figuring out membership functions unlike the mentioned papers (Hudec & Vujoševic´, 2012; Mehta, et al., 2009).

(Gupta, et al., 2015) proposed a ranking function method in which they thought of fuzzy logic as a means to transform vagueness and uncertainty of their documents into fuzzy membership function. Additionally, they used CACM and CISI benchmark datasets to validate the proposed methods. As a result, the methods increased the values of precision, average recall and F-measure.

However, it was not tested on large datasets during the experiments. So, the data had some characteristics of big data but not all of them.

(Shweta, et al., 2016) proposed an algorithm to handle uncertainty in data using fuzzy logic as they implemented a new approach to fuzzy classification. The algorithm assessed universities to predicts the probability of admission in linguistic terms. The algorithm divided the dataset into training, validation and test set. Then, Fuzzy rules were generated. Next, based on the generated rules, the regions Acceptance, Fuzzy and Rejection were identified. In Fuzzy region, which was the area of uncertainty, fuzzy c-means was applied and outliers were achieved which later are used to calculate the rank factors. Finally, quantifiers are applied in the fuzzy region where uncertainty is removed. The efficiency of the proposed algorithm is verified to be better by comparing it with standard algorithms like KNN.

(Iqbal, et al., 2020) investigated the influence of big data in today’s life and discussed various Big Data analytics’ challenges as they considered many computational intelligence techniques. While they presented a method for data modelling, which relies on a hybrid method which is based on the structure and architecture of the mammalian brains, the authors also demonstrated that how efficiently fuzzy logic systems can handle inherent uncertainties related to the data.

(Kayacan, et al., 2018) presented the type-2 fuzzy logic for uncertainty handling. The capability of the prediction of the elliptic Membership Function was tested using interval type-2 fuzzy logic on the dataset oil price prediction which dated back to the years in between 1985 and 2016. As such, the authors used elliptic Membership Function in the type-2 fuzzy to model the uncertainty.

(Duggal, et al., 2015) studied the problem of matching patient records and proposed a solution by using Big Data Analytic techniques. The authors used the Fuzzy logic-based matching algorithms and MapReduce to perform big data analytics in which it checks the similarity between two pieces of information by calculating the distance between them. The less the distance, the more similar are the two.

4. Conclusion

In this paper, we have read and studied over 100 sources including conferences, journal papers, books and/or articles, of which almost 30 papers in a semi-detailed and 10 in a fully detailed manner. We have observed not-enough-work has been done regarding how significantly uncertainty can impact the confidence of big data and data analytics that are currently available. Moreover, even though there is some little work about choosing the most appropriate Membership Function in literature, there is no certain systematic way to figure out the most appropriate fuzzy membership function for the desired context (e.g., to obtain a better uncertainty modelling capability). Not to mention that there is no objective criterion either to check the performance of them.

According to the best of our knowledge, even though there are attempts to automate the choice of rules and membership functions like the mentioned paper (Pendharkar, 2012), the majority of papers are doing so manually. So, the very few solutions available in the market regarding this can be expanded by Improved learning algorithms to turn the choice of rules, membership functions, and even type reduction and defuzzification algorithms into automatic activities without any human being interference. Additionally, even though there are a huge number of defuzzification algorithms out there, there is still a wide domain for improving these methods like the one (Aminifar, 2020) has found. With regards to computational complexity, even though there have been great achievements, but they still should be expanded since the process of reaching the most appropriate fuzzy rules and membership functions is computationally expensive in itself.

To sum up, big data is an inevitable area in the world where fuzzy logic have been used very frequently, so additional studies are highly required to be performed on the relation among the characteristics and how fuzzy logic can help reducing uncertainty in big data. Additionally, more works should be done to find out which characteristic of big data is being the best handled with the help of fuzzy logic.

.

5. Bibliography

Aminifar, A., 2020. Uncertainty Avoider Interval Type II Defuzzification Method. *Mathematical Problems in Engineering.*

Amit, K., Megha, Y., Sandeep, K. & Pranab, K., 2020. Veracity handling and instance reduction in big data using interval type-2 fuzzy sets. *Engineering Applications of Artificial Intelligence.*

Amit, K. & Pranab, K., 2019. Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets. *The International Journal of Intelligent Real-Time Automation.*

Anon., 2019. *Fuzzy Logic | Introduction.* [Online]
Available at: https://www.geeksforgeeks.org/fuzzy-logic-introduction/
[Accessed 31 10 2019].

Anon., 2020. *Artificial Intelligence - Fuzzy Logic Systems.* [Online]
Available at:
https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_fuzzy_logic_systems.htm

Berthold, M., Höppner, F., Klawoon, F. & Borgelt, C., n.d. In: *Guide to Intelligent Data Analysis.* s.l.:Springer-Verlag, pp. 33-34.

Duggal, R., Khatri, S. & Shukla, B., 2015. *Improving patient matching: Single patient view for Clinical Decision Support using Big Data analytics.* s.l., s.n.

Fokoué, E., 2015. A Taxonomy of Big Data for Optimal Predictive Machine Learning and Data Mining.

G, K. & B, Y., 1995. Fuzzy sets and fuzzy logic. *researchgate.net.*

Gupta, Y., Saini, A. & Saxena, K., 2015. A new fuzzy logic based ranking function for efficient Information Retrieval system. *Expert Systems with Applications.*

Hsinchun, C., Roger, H. & Veda, C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly.*

Hudec, M. & Vujoševic´, M., 2012. Integration of data selection and classification by fuzzy logic. *Expert Systems with Applications.*

Iqbal, R. et al., 2020. Big data analytics: Computational intelligence techniques and application. *Technological Forecasting and Social Change,* 153(119253).

Kayacan, E. et al., 2018. Type-2 fuzzy elliptic membership functions for modeling uncertainty. *Engineering Applications of Artificial Intelligence,* Volume 70, pp. 170-183.

Knight, F., 2011. uncertainty and proft, library of economics and liberty.

Laney, D., 2001. 3d data management: Controlling data volume, velocity and variety. *META Group Res. Note 6.*

Lipo, W., Yaoli, W. & Qing, C., 2016. Feature Selection Methods for Big Data Bioinformatics: A Survey from the. *Methods.*

Marr, B., n.d. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.* [Online]
Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=6b4ceddb60ba
[Accessed 21 May 2018].

Mehta, G., Rana, P. & Zaveri, A., 2009. A Novel Fuzzy Based Classification for Data Mining using Fuzzy Discretization. *Computer Science and Information Engineering.*

Normandeau, K., 2013. Beyond volume, variety and velocity is the issue of big data veracity.

Noyes, D., 2018. *https://zephoria.com/top-15-valuable-facebook-statistics/.* [Online]
Available at: https://zephoria.com/top-15-valuable-facebook-statistics/

Nozer, D. & Jane, M., 2004. Membership Functions and Probability Measures of Fuzzy Sets. *Journal of the American Statistical Association,* 99(467).

Pendharkar, P., 2012. Fuzzy classification using the data envelopment analysis. *Knowledge-Based Systems.*

Raghava, R., Abid, H. & Ravi, K., 2014. *Fuzzy-Set Based Sentiment Analysis of Big Social Data.* Ulm, Germany, s.n.

Raghava, R., Abid, H. & Ravi, V., 2014. Fuzzy-Set Based Sentiment Analysis of Big Social Data. *IEEE.*

Reihaneh, H., Erik, M. & Kate, M., 2019. Uncertainty in big data analytics: survey,. *Journal of Big data.*

Shweta, T. et al., 2016. A new approach for data classification using Fuzzy logic. *IEEE.*

Stuart, R. & Peter, N., 2020. What Is AI?. In: *Artificial Intelligence: A Modern Approach.* s.l.:PEARSON SERIES.

Zadeh, L., 1988. Fuzzy logic. *IEEE.*

Zikopoulos, P. C., Deroos, D. & Parasuraman, K., 2013. Harness the power of big data : the IBM big data platform.