EasyChair Preprint

№ 4948

Uncertainty handling in big data using Fuzzy logic - Literature Review

Dyari M. Ameen M. Shareef and Sadegh Abollah Aminifar

February 8, 2021

Uncertainty handling in big data using Fuzzy logic – review

Dyari M.Ameen M.Shareef[1], Sadegh Abdulla Aminifar[1]
[1] 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. In this work, 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. We have noted that 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 and the proposed methods very often were better in terms of efficacy, cost-effectiveness and sensitivity. Needless to say that 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

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. 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. 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 [1].

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 [2] [3].

To tackle these uncertainties in the data, fuzzy logic come to our assistance. Fuzzy sets help in the processes to detect the uncertainties in mathematical order [2].

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 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, discussion and the state of the art is presented. In the final section, the conclusion is presented.

2. Technological Innovation for applied AI systems

There are four kinds of learning; Supervised Learning, Unsupervised Learning and Semi-supervised learning. In Supervised Learning, data is classified to tell the machine specifically what patterns it should seek. In supervised learning, the agent observes both the trained data and the target to reach a function that maps from input to output. The input could be the annoying Captcha image accompanying by an output saying “fire hose” or “street”. Supervised machine learning trains itself on a classified data set to predict output in case if there is a need for new data to be classified. Here the human experts play the role of a teacher since s/he should populate the computer with training data containing the trained data (input) with their corresponding categories (their correct answers) being part of it. However, most of the time, the true function that always predicts the right answer cannot be found since the function algorithm based on assumptions made by human beings, in terms of how the computer should learn, and these assumptions can have a bias towards a specific feature. In unsupervised learning, the data is not classified, and it has been populated without any explicit feedback. The computer is mainly trained with unclassified data. In the world of unsupervised machine learning, the agent tries to figure out patterns in the input without any explicit feedback. In here, the computer is the teacher in itself since it might be able to draw some conclusions to teach and show you some patterns in the data. This comes to our assistance when we as human beings don’t know where to look at in the data. The algorithms that are used here try to use techniques on the input data to mine for rules and/or detect patterns which help in finding significant insights in the data. In addition, the most widely used unsupervised learning technique is clustering, which is an activity of finding groups/clusters in the data. For example, when it is being implemented by showing millions of images taken from all over the internet, a computer vision system can detect a large cluster of similar images which a Kurd would call it “Belek” which translates as “black” in English. In Semi-supervised Learning and Reinforcement learning, either you have classified data or unclassified data. Semi-supervised learning is something in the middle since it falls in between the two. Practically, in certain situations, the cost to have classified data is very much high since it requires human beings with the required expertise to do that. As a consequence, if the majority of data are classified but few are unclassified, the best choice is semi-supervised learning. Semi-supervised learning is typically implemented such that there are a few classified

samples that are used to mine more rules from a large set of unclassified samples. Whereas, in reinforcement learning, the agent learns from a collection of reinforcements like rewards and punishments [4]. This paper present background knowledge about feature selection methods, review a number of papers and discuss some future works which are worth to be worked on and investigate.

3. 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 presented.

3.1 Fuzzy Logic

Lotfi Zadeh invented fuzzy logic as he had been reasoning 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 [5, 6, 7].

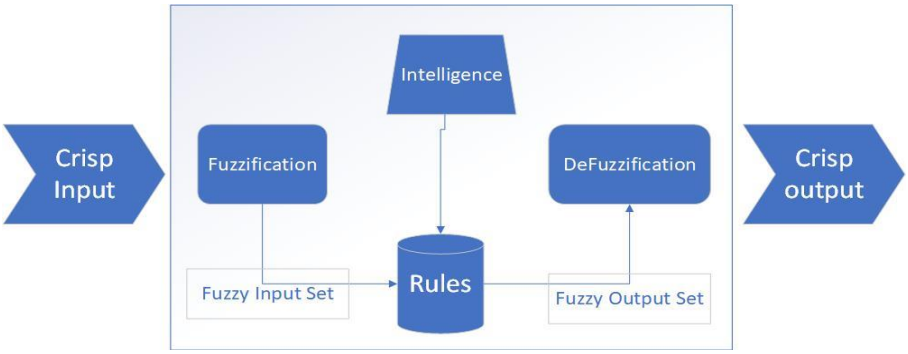

Fig. 1. The architecture of Fuzzy Logic Systems

4. Discussion and state of the art

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

[3] 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 suggests that the de-fuzzified value of the type-2 fuzzy sets provides the same behaviour of a cluster centroid with having fewer instances and fewer computations. In their experiment, the authors divided the data into clusters by k-means algorithm and they then took their cluster centroids and FOU defuzzied centroids. Finally, they used SVR to compare both centroid types. Consequently, the proposed method was more scalable as 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 doesn't only handle uncertainty with fewer costs, but also provides better results at accommodating new data points. However, they didn't use many datasets and all datasets were almost artificial.

Data comes from everywhere including the giant datasets of the social network which leads [11] to propose 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[1]. 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 have negative, positive and neural answers with a degree.

Even though the SQL is a powerful tool, it cannot satisfy the needs for data selection based on linguistic terms. [12] 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 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.

[13] 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

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

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 was better than other classification techniques as they only performed crisp classification. However, The proposed method cannot be applied for data mining in every case.

[2] proposed a method to model the uncertainty in gene expression datasets using Interval type-2 fuzzy uncertainty modelling, symmetrically and asymmetrically. The 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 is 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.

[14] 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 [12, 13].

[15] 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.

[16] 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 were used to calculate the rank factors. Finally, quantifiers were applied in the fuzzy region where uncertainty was being removed. The efficiency of the proposed algorithm was verified to be better by comparing it with standard algorithms like KNN.

[17] 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.

[18] 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.

[19] 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 to which the less the distance, the more similar are the two.

5. Conclusion and future work

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 that 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 neither a 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) nor an 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 [14], 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 [20] has found. With regards to computational complexity, although there have been great achievements, 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 has been used very frequently, so additional studies are highly required to be performed on the relation among the characteristics of big data 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.

.

References

[1] R. Raghava, H. Abid and K. Ravi, "Fuzzy-Set Based Sentiment Analysis of Big Social Data," in *Enterprise Computing Conference*, Ulm, Germany, 2014.

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

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

[4] S. Russell and P. Norvig, "What is AI?," in *Artificial Intelligence: A modern approach: Fourth Edition*, Pearson, 2020, p. 5.

[5] L. Zadeh, "Fuzzy logic," *IEEE,* 1988.

[6] K. G and Y. B, "Fuzzy sets and fuzzy logic," *researchgate.net,* 1995.

[7] M. KekShar and A. Aminifar, "Lookup Table Driven Uncertainty Avoider Based Interval Type-2," *IEEE-SEM,* vol. B, no. 4, 2020.

[8] F. Knight, "uncertainty and proft, library of economics and liberty," 2011.

[9] H. Reihaneh, M. Erik and M. Kate, "Uncertainty in big data analytics: survey,," *Journal of Big data,* 2019.

[10] S. Aminifar and A. Marzuki, "Uncertainty in Interval Type-2 Fuzzy Systems," *Mathematical Problems in Engineering,* vol. 2013, 2013.

[11] R. Raghava, H. Abid and V. Ravi, "Fuzzy-Set Based Sentiment Analysis of Big Social Data," *IEEE,* 2014.

[12] M. Hudec and M. Vujoševic´, “Integration of data selection and classification by fuzzy logic,” *Expert Systems with Applications,* 2012.

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

[14] P. Pendharkar, “Fuzzy classification using the data envelopment analysis,” *Knowledge-Based Systems,* 2012.

[15] Y. Gupta, A. Saini and K. Saxena, “A new fuzzy logic based ranking function for efficient Information Retrieval system,” *Expert Systems with Applications,* 2015.

[16] T. Shweta, S. Bhawna, N. Himanshu, J. Anchit, K. Akshay and G. Sachin, “A new approach for data classification using Fuzzy logic,” *IEEE,* 2016.

[17] R. Iqbal, F. Doctor, B. More, S. Mahmud and U. Yousuf, “Big data analytics: Computational intelligence techniques and application,” *Technological Forecasting and Social Change,* vol. 153, no. 119253, 2020.

[18] E. Kayacan, A. Sarabakha, S. Coupland, R. John and A. Khanesar, “Type-2 fuzzy elliptic membership functions for modeling uncertainty,” *Engineering Applications of Artificial Intelligence,* vol. 70, pp. 170-183, 2018.

[19] R. Duggal, S. Khatri and B. Shukla, “Improving patient matching: Single patient view for Clinical Decision Support using Big Data analytics,” 2015.

[20] A. Aminifar, “Uncertainty Avoider Interval Type II Defuzzification Method,” *Mathematical Problems in Engineering,* 2020.

[21] E. Fokoué, “A Taxonomy of Big Data for Optimal Predictive Machine Learning and Data Mining,” 2015.

[22] M. Berthold, F. Höppner, F. Klawoon and C. Borgelt, in *Guide to Intelligent Data Analysis,* Springer-Verlag, pp. 33-34.

[23] W. Lipo, W. Yaoli and C. Qing, “Feature Selection Methods for Big Data Bioinformatics: A Survey from the,” *Methods,* 2016.

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

[25] K. Normandeau, “Beyond volume, variety and velocity is the issue of big data veracity,” 2013.

[26] C. Hsinchun, H. Roger and C. Veda, “Business Intelligence and Analytics: From Big Data to Big Impact,” *MIS Quarterly,* 2012.

[27] P. C. Zikopoulos, D. Deroos and K. Parasuraman, “Harness the power of big data : the IBM big data platform,” 2013.

[28] R. Stuart and N. Peter, “What Is AI?,” in *Artificial Intelligence: A Modern Approach*, PEARSON SERIES, 2020.

[29] B. Marr, “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read,” [Online]. Available: 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].

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

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

[32] “Artificial Intelligence - Fuzzy Logic Systems,” 2020. [Online]. Available: https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_fuzzy_logic_systems.htm.

[33] D. Nozer and M. Jane, “Membership Functions and Probability Measures of Fuzzy Sets,” *Journal of the American Statistical Association,* vol. 99, no. 467, 2004.

[34] S. Aminifar and A. Marzuki, “Horizontal and Vertical Rule Bases Method in Fuzzy Controllers,” *Hindawi Publishing Corporatio: Mathematical Problems in Engineering,* vol. 2013, 2013.

[35] A. Hamad, S. Aminifar and M. Daneshwar, “An interval type-2 FCM for color image segmentation,” *International Journal of Advanced Computer Research,* vol. 10(46), 2020.