Uncertainty in Data Mining

Ayad Tareq Imam[1][a] and Rafid Qahtan Allawi[2]
[1]*Faculty of IT, Isra University, Amman 11622, Jordan*
[2]*The Higher Institute of Health, Anbar, Iraq*

Keywords: Data Mining, Uncertainty, Fuzzy Set, Artificial Neural Network, Neuro-Fuzzy, Relative-Fuzzy.

Abstract: This paper aims to articulate the current approaches to handle the uncertainty problem in data mining (DM). The difficulties in DM are given to show the impact of the uncertainty problem in DM's various applications. As it has been stated that is no common DM method to handle the uncertainty problem in DM and based upon the literature of cognitive science, this paper highlights a new classification approach to overcome the uncertainty problem in DM that is the Relative-Fuzzy (RF) approach and its ML/RFL-Based Net software tool.

1 INTRODUCTION

Mining an enormous amount of data to show implicit models of that data for their subsequent use in many applications is the goal of the data mining (DM) topic. The yielded data models are used for prediction, description, identification, and optimization (Zaki & Wagner Meira, 2020). The aforementioned understanding interprets the definition of DM as "the searching to find or extract (in terms of mining) hidden predictive patterns or rules" (Ramez & B., 2000)

Because DM's tools are mainly used to predict upcoming patterns and behaviors, by using the computerized and potential analysis of the past data, they provide the ability to make proactive and knowledge-based decisions, and thus, proved itself as an effective aid tool for many areas that require accurate, fast and proactive decision making to maximize gains and reduce losses to a large extent. DM systems have evolved a lot in which the current emerged applications use the online technique to enhance the value of existing information resources and their integration with new software products (Ramez & B., 2000) (Aggarwal, 2015).

Actually, the currently used methods in DM resulted from intensive research and production efforts that started from the simple data storage techniques. This fact does not mean that DM is engaged only in database applications; DM is used with some other application areas also. On the other hand, the technologies that support DM for business applications are DM algorithms, improved data gathering and management, and rising of computation strength (Aggarwal, 2015) (Witten, Frank, & Hall, 2016).

Machine Learning (ML) approaches and techniques are the main DM's techniques that are used to reveal the tacit different types of models which are classification hierarchies, clustering, association rules, sequential patterns, patterns within time series, summarization, and change & deviation detection (Aggarwal, 2015) (Witten, Frank, & Hall, 2016).

The main DM's techniques, some of them have already been introduced for more than a decade and have lately been implemented as clear, reliable, and stable tools are sort of algorithms as given by the common DM software tools like Weka and RapidMinor (Hofmann & Klinkenberg, 2013) (Holmes, Donkin, & Witten, 1994).

It may be important to mention here that there is no single DM method that can be identified as the best one for all applications. Defining a specific computational tool (s) or different combinations of traditional methods depends entirely on the nature of the application and often requires human intervention to settle the suitability of such an approach (Christo, Halkidi, & Michalis, 2002) (Sushmita & K., 2002).

Data, by natural, comes with uncertainty phenomena, which is actually a serious problem that

[a] https://orcid.org/0000-0002-9942-4772

badly affects the effectiveness of any type of data processing algorithms, including DM algorithms. Recalling that uncertainty is the full or partial loss of information (incompleteness), inaccuracy in data (and a process) that shows the lack of the exact meaning. uncertainty comes in different types, which involve a wider sense of uncertainty like Fuzziness, Nonspecificity, and Strife (Imam, Ayad Tareq, 2010).

The goal of this paper is to articulate the current approaches to handle the uncertainty problem in DM. based on the reported result, there is no one approach to handle all types of uncertainty problem in DM, and referring to the literature of cognitive science, this paper highlights the newly defined approach that id Relative-Fuzzy (RF) approach and its ML/RFL-Based Net software tool to overcome the uncertainty problem in DM.

The rest of the paper is as follows: Section 2 illustrates the challenges in DM to show the impact of the uncertainty problem in DM's various applications. Section 3 reports the approaches to handle the problem of uncertainty in the DM. A summary is given in the last section.

2 DATA MINING CHALLENGES

There is a list of challenges for DM that were outlined by researchers. As this list is by no means exhaustive, it gives the problem categories of DM that need to be handled. The most common challenges are (R, B, & Sofia, 2018) (Kumar, Tyagi, & Tyagi, 2014) (Paidi, 2015):

- **Large Databases:** the increase in the size of databases in its fields, records, and tables, day by day is an emerging phenomenon, which imposes finding more efficient, sampling, approximating methods, and possibly parallel processing with a vast capability.
- **High Dimensionality:** the increase in the number of attributes that are used for forming a model and the complicated relationships, which are the hierarchically structured attributes would negatively affect the capability of the model to perform its function. The current solution to this problem is the use of a deep learning approach.
- **Omitted and Noisy Data:** the loss of important attributes and the noisy training data and causes poor performance of the yielded DM models on test data. Solving this problem could be via using techniques like cross-validation, regularization, and other advanced statistical methods
- **Varying Data and Knowledge:** the rapid change in the data makes the formerly produced patterns worthless. The suggested solutions for this problem involve the using of dynamic methods to cue the updating of the produced models.
- **Comprehensibility of Patterns:** which aims to make discoveries more coherent by humans. The most common solutions for this problem are the visualization of data and knowledge by using graphic representations, natural language generation. Additionally, rule structuring and refinement can be used to address a related problem.

Generally, the Fuzzy set, Artificial Neural Network (ANN), and Genetic Algorithm (GA) are used to deal with different challenges in DM. Fuzzy set and ANNs are used to deal with uncertainty, and GAs are concerned with various optimization and search processes. These methods can be combined collaboratively, to solve some DM's problems. Even though the adaptive property of DM systems keeps increasing, there is a considerable role of human intervention, which is normally required for knowledge representation, manipulation, and processing tasks (Christo, Halkidi, & Michalis, 2002; Sushmita & K., 2002).

Is there a common cause of the difficulties listed above? The atomic process of DM's algorithms that is classification is used to develop a DM model by using either a statistical or a structural approach (Ramez & B., 2000; Luger, 2008). **In the statistical approach**, a model is trained to categorize new, unlabeled information based upon their quantitative features to labeled known data by utility functions that employ probability estimates. **In the structural approach**, the class is characterized based upon the arrangement of the existence of their morphological features. Although each approach has pros and cons, the statistical approach is known as more efficient and flexible than the structural approach (Olszewski, 2001; Witten, Frank, & Hall, 2016)

It is obvious that in both classification approaches, the success or accuracy of the resulted model is conditioned by the quality of the training data, which its importance is increased by the increase in the size of the training data, the number of features (dimensionality) of the domain of discourse, and stability of the training data. We mean by the quality of data how the data is clear or certain or in other words has no uncertainty as much as possible.

3 UNCERTAINTY AND DM

DM is mainly concerned with recognizing data patterns (classes) and describing them in a concise style. Careful filtering of data, qualitative opinions, and adjusting of commonsensical rules are to be represented to set up meaningful and constructive relationships among the data variables. The problem of uncertainty is common in the classification process. This problem is appeared in classifying a mysterious element into more than one class, which will lead to the problem of uncertainty in the identity (classification) of this element. To clarify this, take the example of interpreting a word in a natural language to a specific meaning (i.e. class). This word may have more than one meaning (or class), which confuses the program's work and makes the results unreliable (this example shows that fuzzy sets are naturally functioning with the linguistic domain of knowledge and producing more understands). This form of uncertainty is termed as multiplicity, which causes the problem of vagueness (Luger, 2008) (R B. h., 1988) (Orlenko, et al., 2018) in the results of the DM model as well as other classification-based applications. So, to improve the effectiveness of classification in DM, this problem should be handled as agreed upon by (Christo, Halkidi, & Michalis, 2002).

Fortunately, there are several approaches to deal with the problem of uncertainty in the DM. These approaches have been defined and used and showed good performance with of course some cons. The following paragraphs report these approaches.

3.1 Fuzzy Sets Approach

To a reasonable scope, fuzzy logic can support the natural form of human type reasoning. In addition to its use to quantify knowledge, fuzzy sets are used for modeling and handling of uncertainty the use of fuzzy set around DM is increased shown by the different DM systems that use fuzzy set theory in their implementation. Considering that the analysis of data in DM normally needs handling different data classes and numeric attributes at the same time, the role of fuzzy sets in the DM's techniques are seen in the (Aggarwal, 2015) (Luger, 2008) (Witten, Frank, & Hall, 2016) (Zaki & Wagner Meira, 2020):

- **Clustering:** fuzzy sets facilitate the accurate searching of data attributes of linguistic knowledge and the detecting of dependencies among qualitative and semi-qualitative data. This optimizes the searching for desired patterns in the database as there are many features to be considered, which can produce complex errors.
- **Association Rules:** In the statistical approach of classification, the crisp association uses binary values {0,1} to describe the confidence value of the association between feature(s) and a class. As for the ambiguous classification, the description of belonging is described in partly sense, i.e. [0..1].
- **Functional Dependencies (FDs):** consider two continuous variables X and Y, functional dependency denotes that factor X partially determines the level of factor Y. FDs able the intensive expression of real-world properties via utilizing regression to analyze the relationship between two continuous variables. FDs-based inference uses fuzzy logic which supports the fundamental connections of classic, imprecise, and fuzzy relational models of variables in database relationships.
- **Summarization:** The interactive top-down summarization of data utilizes fuzzy IS-A hierarchies (that implicitly use fuzzy set) as domain knowledge. Recalling that the discovery of summary seizes the core representation of a huge amount of information in a database, this fuzzy-based representational structure comes as a generalized combination of attributes and thus gives the discovery process the ability to deal with the database summaries more precisely.

3.2 Artificial Neural Networks Approach

ANN is the most appropriate model where data are so heavy due to its nature as a parallel process model. ANN contribute to DM tasks in (Witten, Frank, & Hall, 2016) (Zaki & Wagner Meira, 2020):

- **Rule Extraction:** The automation of the connectionist rule extraction is achieved by using a trained ANN. Subsequently, the combination and simplifying of these rules could be applied to obtain a further intelligible rule set.
- **Clustering and Self Organization:** there are many examples for the use of self-organization ANN for clustering data such as Self-Organizing Map (SOM) to huge data, the involving of SOM with partitioning the data in stepwise methodology, the hierarchical clustering of SOMs to treat with the dimensionality of the data, and the combining SOM and Sammon's nonlinear mapping to minimize the dimension of data representation for visualization reasons.

3.3 Neuro-Fuzzy Computing Approach

Each ANN and fuzzy set theory are termed soft computing techniques and they are tools for establishing intelligent systems. Recalling that ANN can learn from its environment, are self-organized, and can adapt interactively, fuzzy systems are not (Ata & Kocyigit, 2010). A neuro-fuzzy system is a hybrid style of processing to get advantages of ANN and fuzzy approaches into one complete intelligent decision-making (fuzzy-rule-based) system that works on a data-rich environment system. Another advantage of this approach is overcoming some limitations of the ANN and fuzzy approach. Incomplete data, quantitative data, linguistic data, or a mixture of them are the types of data that would be presented to the neuro-fuzzy system. The neuro-fuzzy classifier (NFC) is an example of this type of system (Zaki & Wagner Meira, 2020; Aggarwal, 2015; Rutkowska, 2012).

3.4 Relative-Fuzzy Approach

While the neuro-fuzzy approach has been agreed as the most advanced approach, it is still been reported that there is no general best approach that can handle the uncertainty problem in DM. This was the motivation for seeking a new approach that can handle uncertainty in the data from a different point of view.

The studies in cognitive physiology show that a mental structure describes the relationship between meaning and memory (recognition) is defined (Groome, 2014). This structure divides elements into clusters and then characterizes each element by a set of its cluster attributes. Recalling that the clustering of data helps to reduce the complexity of its classification (Zaki & Wagner Meira, 2020) (Aggarwal, 2015) (Rutkowska, 2012). This natural phenomenon was used to invent a new approach to classify data that is the Relative -Fuzzy (RF) approach.

RF is a newly defined approach for controlling the complex ambiguity type of uncertainty. It aims to help to produce more accurate and effective predictive DM classification models. RF approach is based on the multiple worlds' principle. RF quantifies the belonging strength of an element of a class by using a pair of fuzzy values (Fd, Fe), where Fd represents the belonging strength of an item's syntax to a certain domain of knowledge, and Fe represents the belonging strength of the syntax to possible syntax's alternatives of an element. An RF-based neural network software tool that is called ML/RFL-Based Net, was developed to generate predictive data mining models, which encompasses a super stage for clustering elements, and a substage for individual recognition of elements within each cluster (Imam, Ayad Tareq, 2010; Al-Zobaydi, M., & John, 2005).

4 DISCUSSION

In this paper, the problem of uncertainty in the data and the approaches to handle it has been reported. While there are DM's tools that have been carried out using a fuzzy set, ANN, and neuro-fuzzy approaches to handle the uncertainty problem in DM, however, the most advanced approach is the neuro-fuzzy approach, which is a hybrid approach that incorporates the ANN and fuzzy set approach. Still, it has been reported that there is no general best approach that can handle the uncertainty problem in DM. This is due to the complex uncertainty (Imam, Ayad Tareq, 2010) nature that exists in the data used by DM approaches while creating predictive or descriptive models.

This expected reason can be solved by a separate step that is preparing data, which aims to minimize the problems found in the data before using it by DM approaches to create DM models.

RF approach is an interesting new approach, which seems to be the solution for well-preparing data. A hierarchical type of ANN implements RF and can be used to define a new and more general DM approach. This is because RF assigns each data element to one or more categories with an attached degree of certainty.

5 CONCLUSIONS AND FUTURE WORKS

DM aims at creating predictive or descriptive data models via using methods of research on huge amounts of data such that the resulting models represent the common attributes among each class of the examined data.

DM is currently a dominant topic in the information industry, and it is keeps evolving technology. DM proved its effectiveness in a lot of applications not limited to discovery science, business management, business intelligence, bioinformatics, cheminformatics, and loyalty card.

The process of creating predictive or descriptive data models comes as classification, clustering, and association rules, which each one is achieved via sort

of algorithms including ANN, GA, and NNC. While some of DM's theoretical aspects still under development, DM itself adds a new type of data processing to classical ones (such as sorting and searching data, data compression-decompression, data encryption-decryption), which is classification-reclassification of data.

The careful studying of difficulties in DM that have been defined by researchers yielded categories of these difficulties: data missing, changing data, the high dimensionality of attributes, and the accuracy of models. All remain to be well answered.

Our analysis of the "somewhat" limited ability of current common utilized approaches in DM to well handling the complex uncertainty of the data itself, led to a conclusion that a data preparation should be made before submitting it to a DM approach to extract a DM model. Since RF can deal with the problem of complex uncertainty, **we recommend considering** the RF approach (and its ML/RFL-Based Net) software tool to be included in the domain of the standard DM software tools like Weka and RapidMinor.

ACKNOWLEDGEMENTS

To the memory of Prof. Hussien Zedan, the former dean of STRL, De Montfort University, Leicester, U.K.; may Allah rest his soul in peace.

REFERENCES

Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing.

Al-Zobaydi, A. T., M., M., & John, R. I. (2005). *Data Mining for Generating Predictive Models of Automatic Speech Recognition*. International Middle Eastern Multiconference on Simulation and Modelling (MESM'05) (pp. 147-150.). Porto, Portugal: EUROSIS.

Ata, R., & Kocyigit, Y. (2010). *An adaptive neuro-fuzzy inference system approach for prediction of tip speed ratio in wind turbines*. Expert Systems with Applications, 37(7), 5454–5460.

Christo, A., H. M., & Michalis, V. (2002). *UMiner: A Data Mining System Handling Uncertainty and Quality. Advances in Database Technology*. 8th International Conference on Extending Database Technology (EDBT), (pp. 762-765). Prague, Czech Republic,.

Groome, D. (2014). *An Introduction to Cognitive Psychology*. Hove, England: Psychology Press.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications (1st ed.)*. Chapman and Hall/CRC.

Holmes, G., Donkin, A., & Witten, I. H. (1994). *Weka: A machine learning workbench*. Second Australia and New Zealand Conference on Intelligent Information Systems. Brisbane, Australia: Piscataway NJ: Institute of Electrical and Electronics Engineers.

Imam, Ayad Tareq. (2010). *Relative Fuzzy: A Novel Approach for Handling Complex Ambiguity for Software Engineering of Data Mining Models*. Leicester, UK: De Montfort University.

Kumar, A., Tyagi, A. K., & Tyagi, S. K. (2014). *Data Mining: Various Issues and Challenges for Future*. International Journal of Emerging Technology and Advanced Engineering, 4(1), 1-8.

Luger, G. F. (2008). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (6th ed.)*. Pearson.

Olszewski, R. T. (2001). *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*. Carnegie Mellon University, School of Computer Science. Pittsburgh, USA: Carnegie Mellon University.

Orlenko, A., Moore, J. H., Orzechowski, P., Olson, R. S., Cairns, J., Caraballo, P. J., Breitenstein, M. K. (2018). *Considerations for automated machine learning in clinical metabolic profiling: Altered homocysteine plasma concentration associated with metformin exposure*. Pacific Symposium on Biocomputing 2018. Big Island of Hawaii, USA: World Scientific.

Paidi, A. (2015). *Major Research Challenges in Data Mining*. International Journal of Trend in Research and Development, 2(4), 5-9.

R, B. h. (1988). *Treatment of uncertainty in artificial intelligence*. Machine Intelligence and Autonomy Aerospace Systems, 115, 233-247.

R, R., B, S., & Sofia, A. (2018). *Data Mining Issues and Challenges: A Review*. International Journal of Advanced Research in Computer and Communication Engineering, 7(11), 118-121.

Ramez, E., & B., N. (2000). *Fundamentals of Database Systems (3rd ed.)*. Addison Wesley.

Rutkowska, D. (2012). *Neuro-Fuzzy Architectures and Hybrid Learning*. Physica.

Sushmita, M., & K., P. S. (2002). *Data Mining in Soft Computing Framework: A Survey*. IEEE Transactions on Neural Networks, 13(1), 3-14.

Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques (4th ed.)*. Morgan Kaufmann.

Zaki, M. J., & Wagner Meira, J. (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms (2nd ed.)*. London, UK: Cambridge University Press.