A Comparative Study of Methods for Deciding to Open Data

Luthfi, Ahmad; Janssen, Marijn

DOI
10.1007/978-3-030-24854-3_14

Publication date
2019

Document Version
Accepted author manuscript

Published in
Business Modeling and Software Design - 9th International Symposium, BMSD 2019, Proceedings

A Comparative Study of Methods for Deciding to Open Data

Ahmad Luthfi[1,2 *[0000-0001-5416-1529]] and Marijn Janssen[1[0000-0001-6211-8790]],

[1] Delft University of Technology
Faculty of Technology, Policy and Management
Jaffalaan 5, 2628 BX Delft, the Netherlands
{a.luthfi, M.F.W.H.A.Janssen}@tudelft.nl
[2] Universitas Islam Indonesia, Yogyakarta, Indonesia
ahmad.luthfi@uii.ac.id

Abstract. Governments may have their own business processes to decide to open data, which might be supported by decision-making tools. At the same time, analyzing potential benefits, costs, risks, and other effects-adverse of disclosing data are challenging. In the literature, there are various methods to analyze the potential advantages and disadvantages of opening data. Nevertheless, none of them provides discussion into the comparative studies in terms of strengths and weaknesses. In this study, we compare three methods for disclosing data, namely Bayesian-belief networks, Fuzzy multi-criteria decision-making, and Decision tree analysis. The comparative study is a mechanism for further studying the development of a knowledge domain by performing a feature-by-feature at the same level of functionalities. The result of this research shows that the methods have different strengths and weaknesses. The Bayesian-belief Networks has higher accuracy in comparison, and able to construct the causal relationships of the selected variable under uncertainties. Yet, this method is more resource intensive. This study can contribute to the decision-makers and respected researchers to a better comprehend and provide recommendation related to the three methods comparison.

Keywords: Methods, Decision-making, Open Data, Bayesian-belief Networks, Fuzzy Multi-criteria Decision Making, Decision Tree Analysis.

1 Introduction

The disclosing of public sector information through open government data initiatives can provide numerous advantages to the public domain at a large scale [1, 2]. Opening the various types of dataset might drive high demand from stakeholders like business enablers, researchers, and non-governmental organizations for specific purposes [3, 4]. At this moment, the governments may have their own business process to avoid human or technological system mistakes from open data decisions [5]. In reality, the way to analyze the risks, costs, and other effects-adverse of disclosing data to the potential stakeholders are cumbersome [6].

There have been works of literature introduce the methods to analyze the potential advantages and disadvantages of opening data and its consequences [6-8]. Methods like Bayesian-belief networks, Fuzzy multi-criteria decision-making, Decision tree analysis, and privacy risks scoring model were used to analyze the potential risks and benefits of opening data [6, 7, 9]. However, none of them provides insight into the comparative studies in terms of strengths and weaknesses. The comparative method is a mechanism for further studying of a knowledge domain by performing comparison a feature-by-feature of selected parameters at the same level of functionalities [10, 11].

In this study, the comparison method is divided into three main groups. First, input parameters that consist of three variables, namely experimental data, data type, and posterior probability. Second, output parameters are decomposed into four variables, namely efficiency, easiness, effectiveness, and complexity. Third, output parameters structure into three variables, namely understandability, subjectivity, and accuracy. We use systematic literature as the main sources to compare each parameter.

The goal of a comparative study conducted in this paper is to explain a better comprehension of the causal process in terms of an event, feature, and relationships by presenting together their complexities in the explanatory parameters [10, 11]. This research can contribute to the decision-makers and respected researchers to a better understanding and provide recommendation related to the three methods comparison. This paper decomposes of five main sections. In Section 1, the current issues and problems definitions are described. Section 2 reviews of related literature are provided. Section 3 the comparison methods and its parameter are defined. Section 4 the comparative studies between three methods are presented. Finally, the paper will be summarized in Section 5.

2 Literature review

In this paper, the comparative studies use three approaches namely Bayesian-belief networks, Fuzzy multi-criteria decision-making, and Decision tree analysis. There are several reasons for using the Bayesian-belief Networks method in open data studies. First, the Bayesian-belief networks are able to capture causal knowledge between selected variables [12]. Second, this theory provides an efficient integration between empirical data and expert's judgment [13]. Third, the Bayesian-belief networks can improve a better understanding of the causal relationships and its consequences [14]. Moreover, the use of Fuzzy set theory in the open data domain is aiming to manage problem complexities of the decision alternatives [15]. The main function of the Fuzzy logic is to capture the expertise of open experts and to express it with computational approach [16, 17]. The properness of the alternative compares to the criteria and the priority weights of each criterion can be analyzed and computed using linguistic matrix values reflected by the fuzzy [17, 18]. The scores for each criterion are summed up to rank the importance of the alternatives decision in open data [28, 29].

The use of decision tree analysis, furthermore, is to construct a feasible decision from the complex problems in the open data domain. A decision-tree is a decision

support theory that uses a schematic tree-shaped diagram of decisions and their reasonable consequences of the conditional control arguments [19, 20]. In addition, decision tree analysis can serve a number of purposes when complicated problems in the decision-making process of releasing data are found. There are some advantages in using decision tree analysis to the decision-making problems [21]. First, Decision tree analysis is able to create comprehensible rules and easy to interpret. Second, Decision tree analysis is able to take into account both continuous and categorical decision variables. Third, Decision tree analysis can provide a clear indication of which variable is becoming the most priority in predicting the outcome of the alternative decisions. Fourth, Decision tree analysis can perform a classification without requiring a computational background in depth.

From the systematic literature of the three selected approaches in analyzing the risks and benefits of opening data, we summarize the specific functionalities of each parameter. First, Bayesian-belief Networks present a directed cyclic graph based on the probabilities of event occurrence [22, 23]. Besides, Bayesian belief networks can perform quantitative judgments by considering the probability distribution to the degree of belief an event both top-down and bottom-up reasoning [6, 24]. Second, Fuzzy Multi-criteria decision-making constructs a hierarchical structure to adjust many types of problem definitions easily, but not focus on incentive data and its consequences [17]. Third, Decision tree analysis predicts the rate of return of various investment strategies to handle the multi-factors response [19]. The important interaction between decision nodes can determine the worst, best, and expected values for the different cases and their problem complexities [21].

3 Research Approach

3.1 Comparison Parameters

In this study, the comparison parameters will be divided into three main parts. First, the input parameter consists of three variables, namely experimental data, data types, and posterior probability. Second, the process parameter decomposes into four variables, namely efficiency, easiness, effectiveness, and complexity. Third, the output parameter consists of three variables, namely understandability, subjectivity, and accuracy. The three sub-parameters used can be explained in detail, as follows:

1. **Input**
   First, experimental data refers to data produced in measurable activities by doing an experimental or quasi-experimental design [25]. The experimental data may be quantitative or qualitative platform using different investigation methods. Second, dataset type refers to a specific type of dataset presented in tabular form and each column of the table represents a specific meaning of values [26]. Third, posterior probabilities define as an uncertainty proposition of the conditional probability that is allocated after the relevant evidence is considered [6].

2. **Process**
   It is started from an efficiency parameter refers to the ability to avoid wasting efforts, energy, and time in doing the evaluation process. In a mathematical sense, it is a measurable instrument of the selected variable to ensure the effort to produce and establish a specific outcome with a less or minimum amount of costs and unnecessary endeavor [25]. Second, the easiness of the selected method in analyzing the selected method means the ease of manner and rules of the evaluation process [22]. Third, the effectiveness refers to the capability of generating the desired result, which means it has an expected outcome and a clear impression [10]. Fourth, the complexity of the process refers to the behavior of a system in interacting components into multiple ways and reasonable [25].
3. **Output**
   The first sub-parameter considers to the understandability of the process. Understandability means that the process of the evaluation is easy to recognize and being understood. Second, subjectivity refers to a subject's personal insights and judgments influenced by individual feelings, desires, expertise in discovering, and level of beliefs in terms of phenomena [25]. Third, the accuracy of the results in evaluating means the accuracy and precision of measurements [22]. A measurement system in specific could be accurate but not precise and vice versa.

4 Result

The following Table 1 gives a summary of comparative study using three methods in open data domain.

Table 1. Comparative studies of methods in opening data

Parameter	Bayesian-belief Networks	Fuzzy Multi-criteria Decision Making	Decision Tree Analysis
Input			
Experimental data	Data is summarized based on the likelihood function from the observe dataset [6, 10]	Data is summarized based on the pairwise comparison matrix [7]	Data is summarized based on the assign payoffs process of possible investments [19]
Data type	Numerical and categorical [27]	Numerical and categorical [17]	Numerical and categorical [20]
Probability	Posterior probability distribution [10]	Posterior probability distribution [17]	Posterior and conditional probability distribution [28]
Process			
Efficiency	Time consuming (maximum) [27]	Time consuming (moderate) [18]	Time consuming (minimum) [21]
Easiness	Highly difficult to understand and interpret the model. Ad-	Moderately difficult to understand and interpret the model. Ad-	Relatively easy to understand and interpret the model. Basic in the

	vanced in the mathematical background is required [29]	vanced in the mathematical background is required [30]	mathematical background is required [21]
Effectiveness	Constructing a causal relationship between variables and provide decision recommendation [25]	Constructing a hierarchy of decisions including its alternative and ranking them into best options [16]	Constructing a structured decisions estimation and its consequences [28]
Complexity	Require the size of the belief-network to simulate and construct complex conditional probabilities [6]	Require rule base analysis to construct a pairwise comparison matrix [16]	Changing variables during the analysis process might be possible to redraw the existing tree. Irrational expectations can lead to flaws and errors in the decision tree [20]
Output			
Understandability	Require high level to comprehend the process and expected results [22]	Require high level to comprehend the process and expected results [31]	Require a moderate level to comprehend the process and expected results [19]
Subjectivity	The elicitation data and information from the experts might possible bias of the quantification process [25]	The elicitation data and information from the experts might somewhat bias of the quantification process [17]	The elicitation data and information from the experts might possible bias of the quantification process [19]
Accuracy	The expected of the value is more accurate when there is less uncertainty in the input parameter. The output is distributed over a range of uncertainties [6, 27]	The estimation result is more consistent compared to reference data approach [31]	The expected result is accurate and able to predict the future outcome [19]

Table 1 describes some different characteristics of the three methods in terms of similar parameters. To classify the different and similarity including its consequences, more explanation can be given as follows:

Bayesian-belief Networks requires maximum allocation time in processing the evaluation instead of the other two methods. This approach is noticeably difficult to under-stand and interpret the proposed model. The subjectivity of this method is potentially found during the process because of the limited resources to quantify the risks and benefits factors. The decision-makers of dataset officer require the capability in mathematics background. However, the advantage of using this method is the expected of the result is more accurate in a range of uncertainties.

Fuzzy Multi-criteria Decision Making consumes time in moderate level in evaluating the dataset. This method is relatively difficult to comprehend and interpret the model. The pairwise comparison tasks may also need an advanced level in mathematics because there are some applied calculus formulations to be used. The expected results show moderate bias in the quantification process. The benefits in using this method are the dataset consistently estimates the selected parameter

Decision Tree Analysis is summarized based on assign payoffs the number of values of the possible investment. This method has a constraint when decision-makers are changing variables during the analysis process, it might be possible to redraw the existing tree. However, the advantage of using Decision Tree Analysis is relatively easy to understand and interpret the model.

In summary, all models have their pros and cons. Which one is favoured is dependent on the needs.

5 Conclusion

Currently, various works of literature have introduced the methods to analyze the potential advantages and disadvantages including its consequences in the open data domain. However, none of them provides insight into their strengths and weaknesses. The comparative study in this paper results in some important findings. First, Bayesian-belief Networks is advance in accuracy because of the very tight steps and rules, but in some cases, this method tends to inefficiency and too complex to be used. Second, Fuzzy Multi-criteria Decision-making is successfully constructing decision-making alternatives and expects to provide the rank of decision options. This method consumes many times to process the entire evaluation because of the many mathematical works at the same time. Third, Decision Tree Analysis is relatively easy to understand and interpret the model. Yet, changing variables during the analysis process might be possible to require redrawing the existing tree. This paper can contribute to the researchers and decision-makers to a better understanding of the method comparison in analyzing the risks and benefits of opening data. In future work, we recommend adding some other approaches like clustering analysis and artificial neural networks to obtain different insights. In addition, to develop a method based on the best parts of each method are not having its disadvantages. Of such an effort is feasible or utopia has to be researched.

References

1. Zuiderwijk, A. and M. Janssen, *Open Data Policies, Their Implementation and Impact: A Framework for Comparison.* Government Information Quarterly, 2013. **31**(1).
2. Zuiderwijk, A. and M. Janssen, *Towards decision support for disclosing data: Closed or open data?* Information Polity, 2015. **20**(2-3): p. 103-107.
3. Veenstra, A.F.V. and T.A.v.d. Broek, *Opening Moves – Drivers, Enablers and Barriers of Open Data in a Semi-public Organization*, in *International*

Conference on Electronic Government. 2013, Springer: Porto, Portugal. p. 50-62.

4. Ubaldi, B., *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives.* OECD Working Papers on Public Governance, 2013. **22**: p. 60.
5. Luthfi, A., M. Janssen, and J. Crompvoets. *Framework for Analyzing How Governments Open Their Data: Institution, Technology, and Process Aspects Influencing Decision-Making*. in *EGOV-CeDEM-ePart 2018*. 2018. Donau-Universität Krems, Austria: Edition Donau-Universität Krems.
6. Luthfi, A., M. Janssen, and J. Crompvoets. *A Causal Explanatory Model of Bayesian-belief Networks for Analysing the Risks of Opening Data*. in *8th International Symposium, BMSD 2018*. 2018. Vienna, Austria: Springer International Publishing AG.
7. Luthfi, A., et al. *A Fuzzy Multi-criteria Decision Making Approach for Analyzing the Risks and Benefits of Opening Data*. in *17th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2018*. 2018. Gulf University for Science and Technology (GUST), Kuwait: Springer LNCS 11195.
8. Luthfi, A. and M. Janssen, *A Conceptual Model of Decision-making Support for Opening Data*, in *7th International Conference, E-Democracy 2017*. 2017, Springer CCIS 792: Athens, Greece. p. 95-105.
9. Ali-Eldin, A.M.T., A. Zuiderwijk, and M. Janssen. *Opening More Data: A New Privacy Scoring Model of Open Data*. in *Seventh International Symposium on Business Modelling and Software Design (BMSD 2017)*. 2017. Barcelona, Spain: SCITEPRESS - Science and Technology Publication, Lda.
10. Nojava, F., S. Qian, and C. Stow, *Comparative analysis of discretization methods in Bayesian networks.* environmental Modelling & Software, 2017. **87**: p. 64-71.
11. Jan, B., et al., *Deep Learning in Big Data Analytics: A comparative Study.* Computers and Electronic Engineering, 2019. **75**: p. 275-287.
12. Preece, A. *Building the right system right - Evaluating V&V methods in knowledge engineering*. in *The Eleventh Workshop on Knowledge Acquisition, Modelling and Management*. 1998. Voyager Inn, Banff, Alberta, Canada.
13. Uusitalo, L., *Advantages and Challanges of Bayesian Networks in Environmental Modelling.* Ecology Modelling, 2007. **203**(3-4): p. 312-318.
14. Fenton, N. and M. Neil, *Risks Assessment and Decision Analysis with Bayesian Networks*. 2012: Noca Raton: CRC Press.
15. Teicher, M., *Interviewing Subject Matter Experts*, in *International Cost Estimating and Analysis Association (ICEAA)*. 2015.
16. Rezaei, P., et al., *Application of Fuzzy Multi-Criteria Decision Making Analysis for Evaluating and Selecting the Best Location for Construction of Underground Dam.* Acta Polytechnica Hungarica, 2013. **10**(7): p. 187-205.
17. Ceballos, B., M.T. Lamata, and D. Pelta, *Fuzzy Multicriteria Decision-MakingMethods: A Comparative Analysis.* International Journal of Intelligent Systems, 2017. **32**: p. 722-738.

18. Kahraman, C., S.C. Onar, and B. Oztaysi, *Fuzzy Multicriteria Decision-Making: A Literature Review.* International Journal of Computational Intelligence System, 2015. **8**(4): p. 637-666.
19. Delgado-Gómez, D., J. C.Laria, and D. Ruiz-Hernández, *Computerized adaptive test and decision trees: A unifying approach.* Expert Systems with Applications, 2019. **117**: p. 358-366.
20. Yeoa, B. and D. Grant, *Predicting service industry performance using decision tree analysis* International Journal of Information Management, 2018. **38**(1): p. 288-300.
21. Yuanyuan, P., B.A. Derek, and L. Bob, *Rockburst prediction in kimberlite using decision tree with incomplete data.* Journal of Sustainable Mining, 2018. **17**: p. 158-165.
22. Chakraborty, S., et al., *A Bayesian Network-based customer satisfaction model: a tool for management decisions in railway transport.* Journal of Decision Analytics, 2016. **3**(4): p. 2-24.
23. Xiong, J., et al., *Personalized visual satisfaction profiles from comparative preferences using Bayesian inference.* Energy Procedia, 2017. **122**: p. 547-522.
24. Castillo, E., et al., *Complexity Reduction and Sensitivity Analysis in Road Probabilistic Safety Assessment Bayesian Network Models: Complexity reduction and sensitivity analysis.* Computer-Aided Civil and Infrastructure Engineering, 2017. **32**(6).
25. Beuzen, T., *A comparison of methods for discretizing continuous variables in Bayesian Networks.* Environmental Modelling & Software, 2018. **108**: p. 61-66.
26. Safarov, I., S. Grimmelikhuijsen, and A. Meijer, *Utilization of open government data: A systematic literature review of types, conditions, effects and users.* Information Polity, 2017. **22**(1): p. 1-24.
27. Herland, K., H. Hämmäinen, and P. Kekolahti, *Information Security Risks Assessment of Smartphones Using Bayesian Networks.* Journal of Cyber Security, 2016. **4**: p. 65-85.
28. Adina Tofan, C., *Decision Tree Method Applied in Cost-based Decisions in an Enterprise.* Procedia Economics and Finance, 2015. **32**: p. 1088-1092.
29. Horný, M., *Bayesian Networks*, in *Technical Report No. 5*. 2014, Department of Health Policy & Management: Boston University School of Public Health.
30. Mohsen, D., et al., *A Combined Fuzzy MCDM Approach for Identifiying the Suitable Lands for Urban Development: An Example from Bandar ABBS, Iran.* Journal of Urban and Environmental Engineering, 2014. **8**(1): p. 11-27.
31. Werro, N., *Fuzzy Classification of Online Customers.* Fuzzy Management Methods, 2015.