Fuzzy Expert Systems for Automated Data Quality Assessment and Improvement Processes

Corinna Cichy[1,2] [0000−0002−1745−0126] and Stefan Rass[1] [0000−0003−2821−2489] ⋆

[1] Alpen-Adria University Klagenfurt, 9020 Klagenfurt, Austria
[2] Volkswagen Bank GmbH, 38112 Braunschweig, Germany

Abstract. Due to its importance for decision-making processes, data quality plays a crucial role in modern data management. However, assessing data quality still involves a number of manual steps. Moreover, these tasks are characterized by subjective decisions performed by domain experts. The goal of this research is to reduce the time spent on these activities by investigating the possibility of imitating an expert's quality judgement via an approximation of the expert reasoning by fuzzy logic. We lay out the steps to (automatically) set up such a system and introduce possible applications. The approach allows us to benefit from combining established data management methods with machine learning as well as knowledge engineering techniques in order to handle the complexity and uncertainty of the presented process in a transparent way.

Keywords: Fuzzy expert systems · Explainable AI · Data quality

1 Introduction

Human capabilities are often insufficient to handle today's data management tasks which can lead to increased costs for the business [4,6]. Consequently, the importance of improving data quality processes is widely recognized [5]. Nevertheless, the prioritization of data quality issues and the assessment of the overall value of a data set is often performed through manual tasks. To some extent, this can be attributed to the decisions made by experts when it comes to evaluating data quality from a business perspective. Yet, existing data quality frameworks give little advice on how to handle such decisions based on the available metadata and the context-dependent aspects of data quality are mostly handled via survey questionnaires [3]. Especially within the domain of regulatory compliance, this causes a high risk. This motivates our research with the goal of imitating an expert's decision-making about data quality. In particular, we propose a combination of fuzzy logic reasoning and regression techniques to "machine-learn" the expert's judgment and ultimately automate this manual labor. The overall aim of the proposed approach is to save time regarding critical data quality processes while providing consistency throughout personnel changes. The choice of applying fuzzy logic is premised on the assumption that domain experts are more likely to express their knowledge-based decisions in terms of natural language, rather than mathematical concepts. Moreover, next to topics such as fairness and accountability, the aspect of transparency within machine learning becomes more and more important as one of the key AI principles [1,8]. The provided degree of explainability enables the domain experts to play a specified role during the development process and, rather than being confronted with plain results, understand the recommended actions proposed by the system.

2 Fuzzy Reasoning about Data Quality

We propose a framework based on fuzzy logic that aims at simulating an expert's behavior throughout his tasks in data quality assessment. In particular, we describe the development of a knowledge-based system that combines the domain knowledge of an expert with existing measurement metrics. This approach produces the recommended actions for a decision process to the expert in comprehensible way, mainly due to the natural interpretation of the fuzzy logic involved. In this section, the necessary steps for developing and implementing such a support system are explained. Moreover, we demonstrate the proposed procedure for two applications.

2.1 Overview of the Method

For the method to be applicable, the following preleminary conditions have to be met: The process should

(i) include *objective measurements* such as data quality metrics or other key performance indicators (KPIs) that provide an informative indication for the subsequent decision process,
(ii) involve an *expert evaluation* that rests upon the measurements from (i) combined with his implicit expert knowledge of the process, and
(iii) be a *recurring process*, i.e. the output represents a regular activity of the expert and historic data on previous decisions are available.

Once these preliminary conditions are met, the approach consists of three main phases and proceeds as follows:

1.) **Capturing the Expert's Knowledge Using Fuzzy Logic**: The expert expresses his decision-making in terms of natural language, which is then transferred and expressed within a number of candidate fuzzy systems.
2.) **Training the Fuzzy Approximation Model**: The relevant fuzzy rules are selected, e.g. by performing a linear regression on historical data, with the regression base functions being fuzzy if-then rules.
3.) **Applying the Fuzzy Approximation Model**: This represents the practical usage of the model with regard to the experts' recurring decision process on a regular basis.

The actions of the data scientist and the input from the data quality expert are distinguished in Figure 1 to emphasize the practical realization of the method. In particular, the domain expert provides his knowledge (textual input) which is captured within the model and combined with the objective measurements (numerical input). These steps are demonstrated in Sections 2.2 to 2.4.

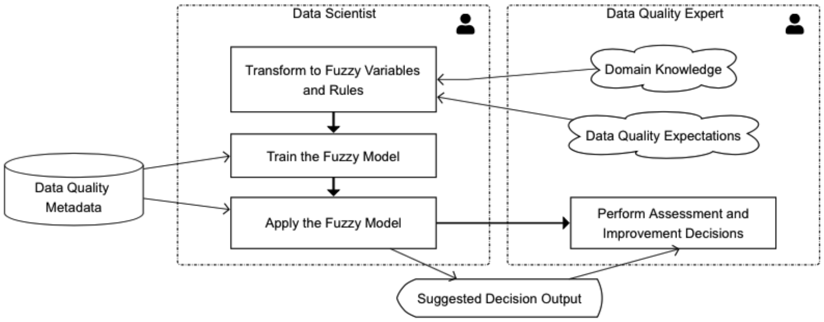
Data Scientist
Data Quality Expert
Transform to Fuzzy Variables and Rules
Domain Knowledge
Data Quality Expectations
Train the Fuzzy Model
Data Quality Metadata
Apply the Fuzzy Model
Perform Assessment and Improvement Decisions
Suggested Decision Output

Fig. 1. Illustration of the Method

2.2 Capturing the Expert's Knowledge Using Fuzzy Logic

The first step of our proposed method is to develop the relevant fuzzy components, i.e. fuzzy variables along with a set of (potentially significant) fuzzy rules. Advantageous about this choice is that fuzzy variables allow for different degrees of membership to various categories. This is especially helpful if the experts are uncertain about assigning clear thresholds (e.g., a record has "good quality" if it is at least 60% complete). As a starting point, the expert names relevant categories and associations (in natural language terms) which builds the basis for the fuzzy variables and their membership functions. Specifically, the expert states his interpretation of existing objective measurements and his subsequent decision process.

Following the principles of constructing a fuzzy expert system, we proceed by defining fuzzy rules for inference towards the quality assessment. Here, expert knowledge is again incorporated since the expert expresses aspects about his decision-making. Based on natural language explanations, rules of the form: IF *condition* THEN *conclusion* are extracted.

The result of this step is a set of candidate (basis) fuzzy if-then rules $R_1, \dots, R_n$, which can be converted into a set of continuous functions $g_1, \dots, g_n : \mathbb{R} \to \mathbb{R}$ representing the basis fuzzy systems. The individual importance of the rules is determined by regression techniques in the next step.

2.3 Training the Fuzzy Approximation Model

To refine the fuzzy rule set towards a best reproduction of the given expert's rating (supervised learning), we refer to the techniques laid out in [2] and [7]. In particular, here we use linear regression to fit a linear combination of fuzzy rules. This has the advantage of assigning weights to the rules that reflect the particular rule's importance. Moreover, the the overall model is open to a variety of further statistical analyses and model diagnostics. Fitting this "expert imitation model" (1) requires training data, which we take from historical (manual) data quality judgments and objective measurements, represented as x_i.

$$f(x_i, \alpha_0, \alpha_1, \ldots, \alpha_k) = \alpha_0 + \alpha_1 g_1(x_i) + \ldots + \alpha_k g_k(x_i) + \varepsilon_i, \qquad (1)$$

for $i = 1, \ldots, n$; and with ε_i being a random error term for x_i. Evaluating the deterministic part of Equation 1 provides the desired output for a given data set. Moreover, the parameters can be interpreted as weights and more specifically, the importance of the fuzzy rules.

2.4 Application to Data Quality Assessment and Improvement

Due to the context-dependency of data quality, it is often inevitable to consult domain experts for its assessment with regard to data quality requirements (e.g. from business side or driven by regulatory guidelines). They incorporate the severity of potential consequences for the business when assessing the overall value of the data. Since the experts have to consider a large amount of metadata for each data set in form of objective measurements, this can be a highly time consuming process.

To use the proposed method within the context of data quality assessment,the people involved in the data quality assessment follow the steps illustrated in Figure 1. The metadata stems from data quality measurements which are commonly obtained via data quality tools, providing automated calculations of key metrics. A fuzzy rule can have the form "If metric 1 is low then quality is poor" in this context. Applying the model (1) is then a trivial matter of evaluating the formula on the data set to be quality-assessed, whose quality metadata $i = 1, 2, \ldots$ directly go into (1) as the variables x_i. For this practical setting, the model produces a data quality score representing the quality indicator of the data record. The explanation of how this assessment was obtained is visible from the linear model, which is simply an aggregation as a weighted average of the rules about quality. The judgment is also explainable, since each g_j (for $j = 1, 2, \ldots, k$) in (1) corresponds to natural-language if-then rules and affects the overall assessment in the direction and to the extent told by its coefficient α_j by sign and magnitude.

For data quality improvement, we consider the prioritization of identified data quality issues. Here, rather than evaluating the formula, we use the coefficients therein as indicators of how much impact an improvement would make to the overall quality. If the coefficient of a fuzzy rule is large, then the priorities on what criteria to improve can be set accordingly.

3 Conclusion

The challenging tasks within the data quality management of a business are characterized by subjective and time consuming processes. However, a suitable adaption of machine learning and knowledge management techniques can provide assistance for such tasks. We propose an approximation by fuzzy rules to support the assessment and improvement of data quality. The expert system can provide assistance throughout the decision-making process of data quality experts and data consumers. We further contribute a set of guidelines to set up such imitation systems in a business. This provides the basis for identifying further processes that are suitable for an implementation. Next steps of our research include the validation of the method in comparison with other machine learning techniques for the assessment and improvement phases in the context of risk data including a test of scalability regarding the number of fuzzy variables, rules and outcomes.

References

1. Adadi, A., Berrada, M.: Peeking inside the black box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Cichy, C., Rass, S.: A fuzzy-approximation approach to explainable information quality assessment. In: Proc. of the 34th International Business Information Management Association Conference (IBIMA'19). pp. 3919–3931. Madrid, Spain (2019)
3. Cichy, C., Rass, S.: An overview of data quality frameworks. IEEE Access **7**, 24634–24648 (2019)
4. Hippold, S.: Watch these data and analytics challenges and trends (2018), https://www.gartner.com/smarterwithgartner/watch-these-data-and-analytics-challenges-and-trends/. Accessed 2 Aug 2020
5. Nagle, T., Redman, T., Sammon, D.: Assessing data quality: A managerial call to action. Business Horizons **63**(3), 325–337 (2020)
6. Redman, T.C.: Getting in Front on Data: Who Does What. Technics Publication, Baskin Ridge, NJ, USA (2016)
7. Riza, L.S., Bergmeir, C., Herrera, F., Benitez, J.M.: frbs (2015), https://cran.r-project.org/web/packages/frbs/frbs.pdf. Accessed 2 Aug 2020
8. Zeng, Y., Lu, E., Huangfu, C.: Linking artificial intelligence principles. arXiv (2018), https://arxiv.org/pdf/1812.04814v1.pdf. Accessed 2 Aug 2020