

Supporting the Design of Data Preparation Pipelines

(Discussion Paper)

Camilla Sancricca¹, Cinzia Cappiello¹

¹Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria

Abstract

The availability of a large amount of data facilitates spreading a data-driven culture in which data are used and analyzed to support decision-making. However, data-based decisions are effective only if the considered input data sources are not affected by poor quality and biases. For this reason, the data preparation phase is crucial for guaranteeing an appropriate output quality. There is a strong evidence in the literature that dealing with data preparation is not simple: it is the most resource consuming step in data analysis and most of the times it is performed using a trial and error approach. Considering this, we aim to support users in the design of data preparation pipelines by identifying the most suitable data transformation/cleaning operations to apply and the order in which they have to be executed. In order to achieve such a goal, using different datasets and ML algorithms, we conducted a series of experiments designed to assess the impact of different types of errors on the quality of the output. The idea is to develop a framework that provides users with guidelines that recommend to address the data quality issues with the highest negative impact first. A preliminary validation has confirmed that following the system suggestions yields better results.

Keywords

Data Quality, Data Preparation, Bias, Decision-making

1. Introduction

Nowadays, organizations are increasingly adopting a data-driven culture in which data are used and analyzed to support decisions. However, in order to get valuable results from data analysis, input data should be reliable to avoid the well known Garbage In Garbage Out (GIGO) effect. Unfortunately, real world data are often affected by errors, inconsistencies, incomplete values, or biases (e.g., the data source does not exactly represent the considered population). In order to avoid having erroneous and unusable outputs, data preparation has become a crucial step in the data analysis pipeline for guaranteeing an appropriate quality output. It is worth noting that data preparation is not a simple task: usually it is a time consuming activity and it is usually performed by using a trial and error approach. This paper aims to propose a framework to support users in the identification of the most appropriate data transformation/cleaning activities to perform. Recommendations are designed to address the data quality issues that affect more the reliability of the analysis results first. Data Quality (DQ) is often defined as “fitness for use”, i.e., the ability of a data collection to meet user requirements [1]. Data quality


SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

✉ camilla.sancricca@polimi.it (C. Sancricca); cinzia.cappiello@polimi.it (C. Cappiello)

ORCID 0000-0001-6062-5174 (C. Cappiello)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

is a multi-dimensional concept that refers to several aspects that affect data from various perspectives. The most used data quality dimensions are accuracy, completeness, consistency and timeliness [2]. We base our approach on the fact a specific data quality issue can be solved by using one or more data preparation actions. Therefore, once that a preliminary data quality assessment reveals the dimensions that need to be improved, it is possible to identify the related data preparation actions to perform. Moreover, a series of experiments allowed us to discover that the impact of different quality issues is dependent on the context of the analysis, where the context is modeled as the analytics application and the characteristics of the considered data sources. These findings helped us in designing an adaptive system able to provide users with guidelines about the sequence of data preparation activities to perform in a particular context to maximize the output quality. The effectiveness of such guidelines has been proven testing them with different combinations of data sets and analytics algorithms. The tests confirmed that applying the suggested sequence of data preparation tasks yields better final results. The paper is organized as follows. Section 2 presents the proposed approach and shows the experiments conducted to understand the impact of the data quality errors on the results of data analysis. Section 3 shows a preliminary validation of the effectiveness of the presented system. Section 4 discusses previous literature contributions. Finally, Section 5 draws conclusion and discusses future work.

2. A framework to support the design of data preparation pipelines

This section aims to describe the framework we designed for supporting users in the data preprocessing phase. Sections 2.1 and 2.3 present the architecture and the experiments conducted for feeding the knowledge base used to provide valuable recommendations. Section 2.2 clarifies the data quality aspects and biases we consider in this work.

2.1. The general architecture

A data analytics pipeline is usually composed of two main phases: data preprocessing and data analysis. The former collects and processes the data for guaranteeing a certain level of quality while the latter performs the data analysis tasks. This paper focuses on the data preprocessing

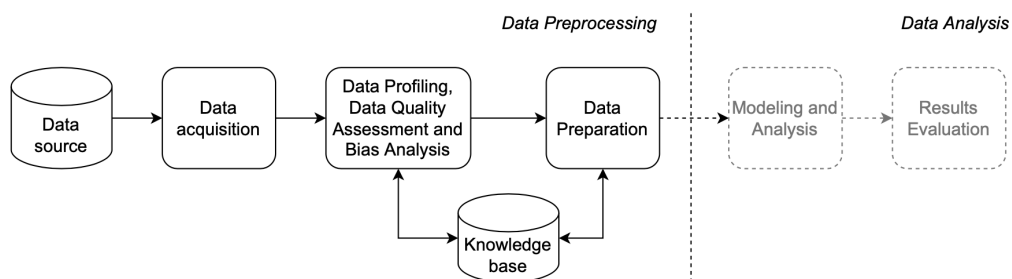


Figure 1: The Data Preparation framework: the high level architecture

phase and proposes a framework that aims to guide the user through the design of the data preparation pipeline, suggesting the most suitable activities to perform. As depicted in Figure 1, the system allows the user to specify the context of analysis in terms of the data sources to consider and the analytics application to use. Once the data are collected, they are inspected and analyzed by using data profiling techniques and data quality and bias assessment algorithms. The considered data quality and bias metrics are described in Section 2.2. Users can access the results of this phase in order to understand the content of the datasets and their initial suitability for the task at hand. The results are sent to the data preparation module that has to identify the most appropriate task to perform and support its execution.

The Data Preparation module is supported by a knowledge base that contains information to infer the data preparation tasks to suggest. It contains, for each data quality dimension, the association between the considered dimension and the data preparation activities able to improve it. Moreover, it registers the impact of the issues related to a quality dimension on the results of an analytics application. Note that such an impact depends on the chosen data analysis algorithm and the data source profile. These relationships together with the data provided by the data profiling and quality/bias assessment module are the input for the design of the data preparation pipeline that defines the data preparation techniques to consider and the order with which they have to be executed. In details, the most suitable data preparation tasks are identified on the basis of a ranking that sorts the data quality dimensions to improve. Such ranking is obtained by merging the quality level of each dimensions with its impact of the quality results.

Note that in the envisioned approach, we assume that the users are free to follow the suggestions or not, letting them building their own data preparation pipeline and executing it. When the data preparation phase is completed, the data analysis phase can start.

2.2. Data Quality and Bias

As stated in the Section 1, Data Quality is a multidimensional concept: a DQ model is composed of *DQ dimensions* that represent the different aspects to consider. In our work, we focused on the accuracy and completeness dimensions. Accuracy is defined as the closeness between a data value v and a data value v' , considered as the correct representation of the real-life phenomenon that the data value v aims to represent [2]. In the literature two types of accuracy are defined: Syntactic accuracy and Semantic accuracy. *Syntactic accuracy* is the closeness of a value v to the elements of the corresponding definition domain D [2]. If v is one of the values in D , then v is accurate otherwise it is not accurate. *Semantic accuracy* is defined as the distance between a data value v and a data value v' . In this case v is a value of the domain D . Semantic inaccurate values are those that, despite belonging to the domain, are not the correct ones. *Completeness* characterizes the extent to which the table represents the corresponding real world [2]. It is related to the presence of null values and a simple way to assess completeness in a table is to calculate the ratio between the number of non-null values and the number of cells of the table. The proposed framework also aims to alert the user of the presence of potential biases that can affect the data. In fact, a high data quality level does not guarantee the representativeness of the database: a bias analysis is needed. Currently, we focus on three metrics for quantifying the bias: coverage, density and diversity [3] [4]. *Coverage* is defined as the degree to which the dataset is

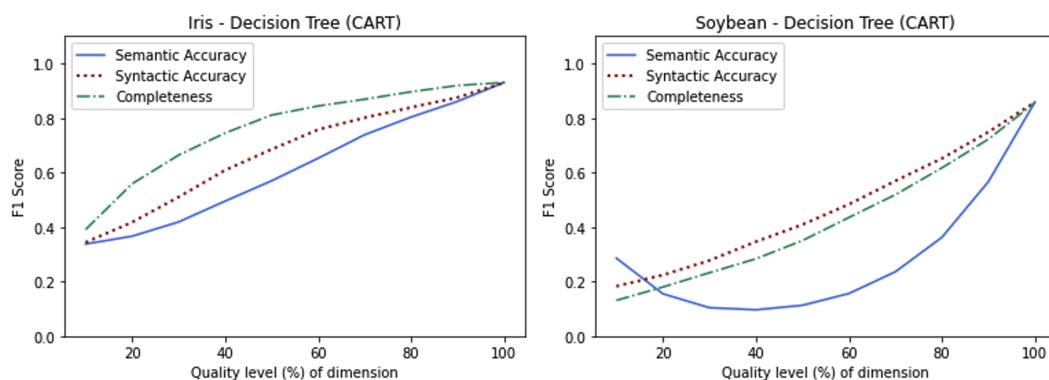


Figure 2: Results related to the impact of data quality issues on Decision Tree results

representative of the real-world. It can be measured as the ratio between the number of different entities populating the dataset, and the total number of real-world entities, for which the value can be approximated or given by the user. *Density* is the degree to which different entities occur into the dataset. Given an attribute, for each distinct value, it is defined as its occurrence in one column. Density can be also computed for the whole attribute, representing the degree to which distinct values in one column are uniformly distributed. *Diversity* is associated with the concept of entropy. In the area of relational databases, entropy relies on how much an attribute is informative. To assess diversity the Shannon entropy, also known as Shannon’s diversity index [5], can be used.

2.3. Evaluation of the impact of poor Data Quality on Machine Learning

This section describes the experiments carried on to understand the impact of the data quality dimensions on the quality of the analytics results. So far we focused on two data quality dimensions, i.e., Accuracy and Completeness and five ML algorithms, i.e., Decision Tree, k-Nearest Neighbors and Naïve Bayes for classification, k-Means as clustering algorithm and Ordinary Least Squares as linear regression method. Note that, as regards the accuracy, we analyzed the impact of errors related to the syntactic accuracy (e.g., typos or values outside the admissible domain) and the semantic accuracy (e.g., an admissible value that is not the correct one). The method we used to evaluate the discussed impact is based on a fault injection approach. Starting from a dataset and a quality dimension we introduced errors with a uniform distribution varying their quantity (from 0 to 90% with a step of 10%). In this way, we obtained nine instances of the same dataset with which we fed a ML algorithm and collected the data related to the performances of the results. We reiterated this procedure for all the considered data quality dimensions. Figure 2 shows an example of the results obtained by applying the method described above on two datasets (i.e., Iris¹ and Soybean² datasets) that are characterized by low dimensionality (i.e., limited number of attributes) but different data types: Iris dataset is numerical and Soybean dataset is categorical. As ML application, a Decision Tree classifier is

¹<https://archive.ics.uci.edu/ml/datasets/iris>

²[https://archive.ics.uci.edu/ml/datasets/Soybean+\(Large\)](https://archive.ics.uci.edu/ml/datasets/Soybean+(Large))

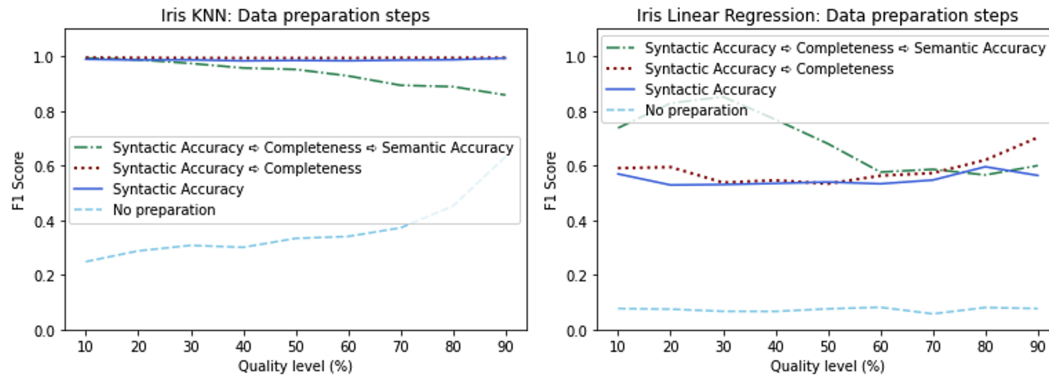


Figure 3: Evaluation of data preparation guidelines on the Iris dataset applying different ML applications

considered and its results are evaluated by using the F1 Score. The results of the experiments show that the quality impact depends on the analyzed datasets that in this case differ in the data types. In this example, on the basis of the discussed impact it is possible to understand that both for Iris and Soybean datasets the first dimension to improve is the Semantic Accuracy while the second one is the Syntactic Accuracy for Iris and the Completeness for Soybean. In summary, these experiments allowed us to identify the data quality dimensions of which issues have a greater impact on the results. Such dimensions are the ones that should be improved at first. Once identified the quality dimension to improve we can retrieve from the knowledge base the data preparation techniques able to improve it and suggest them to the user.

3. Experimental Results

This section provides an example of the experiments conducted for evaluating the effectiveness of the data preparation guidelines provided by the proposed system. Considering a dataset, a ML application and the related data quality dimensions ranking, we conducted the experiments using the following process: (i) we injected errors in the dataset creating nine instances following the same procedure described in the previous section. Each instance contains the same quantity of errors for all the considered quality dimensions. (ii) the ML application is executed on the created instances and the output quality is registered; (iii) following a specific ranking of the data quality dimensions, one dimension at a time is improved and the ML application is executed on the enhanced dataset instances. This method has been applied both following the impact ranking described in Section 2.1 and also using different rankings obtained changing the order of the quality dimensions. The results confirmed the usefulness of our approach. In fact, it has been discovered that, performing data preparation following the suggestions yields better final results than applying the same preparation activities in a different order. Figure 3 shows an example of the results obtained running the above method for the Iris dataset. The considered quality dimensions are the same as before: Semantic and Syntactic Accuracy and Completeness. Instead, the selected ML applications are k-Nearest Neighbors for classification and Ordinary Least Squares for regression. The quality of the results has been evaluated by using the F1

Score. Both algorithms are characterized by the same impact ranking: Syntactic Accuracy, Completeness and Semantic Accuracy. Figure 3 shows the results of the experiments where each curve represents the trend of the performance obtained every time a quality dimension is improved, following a specific order. It is possible to notice that the highest performance is reached by improving the first dimension of the extracted ranking. Data preparation actions performed subsequently bring only minimal improvements. It means that improving the most impacting dimension leads to obtain quickly good results. Therefore, the suggestions might help the user in saving some time in performing data preparation, since the suggestions avoid the adoption of a trial and error approach and the first suggested actions are already effective on the obtained results.

4. Related work

Data-driven decision making relies on the use of advanced analysis tools able to deal with high volume of data. Unfortunately such data are often affected by poor quality that might negatively impact on the quality of the data analysis results and, thus, on the decisions outcome. To address data quality issues, within the processing pipeline a set of data preparation tasks can be performed. The main commercial data preparation tools are surveyed in [6], where their features are collected illustrating the data preparation task in which they are involved, for example, "Locate missing values", "Locate outliers" or "Change date & time format". Then, these features are re-organized in six categories, i.e., data discovery, data validation, data structuring, data enrichment, data filtering and data cleaning. Moreover, recent studies [7] [8] started to analyze the impact of data quality issues on the output quality of ML algorithms. These contributions investigate the effect of missing, inconsistent and conflicting data on the results of different ML tasks and quantify such an impact with an evaluation metric, called sensibility, which measure the sensibility of an algorithm to the data quality. They provide guidelines on: (i) detecting the errors rates (e.g., missing rate, inconsistent rate, conflicting rate) in the given data (ii) selecting the least sensitive ML algorithm according to the error types that have higher rates and (iii) clean each type of dirty data until reaching its corresponding *keeping point*, a metric that identifies at which point the corresponding error rate is acceptable for the selected ML algorithm. However, this approach does not provide any guidelines to clean the poor quality data. Moreover, the evaluation metrics used in this work are Precision, Recall and F-measure for all the ML algorithms. This can become a problem in testing some ML methods because the computation of such metrics requires the original class labels which, in some cases, may not be provided in the final prediction. In our approach, in fact, we consider different metrics to evaluate the output quality of different ML methods, e.g., F1 Score for classification or Silhouette Score for clustering. Several studies focus instead on the design of data preparation pipelines with the goal to get better results. [9] proposes a reinforcement learning method that finds out the optimal sequence of data preparation tasks able to maximize the performance of a ML model. The approach takes as input a data source, a model and a performance metric to optimize and, supported by Q-learning, explores all the possible data preparation tasks, determining at each step the next preprocessing method to execute to maximize the given metric. Our approach differs from this method since we have a knowledge base containing information

about the impact that several types of quality errors have on different ML applications, and the corresponding data preparation tasks that should be performed to address such quality issues. This allows us to suggest an optimal sequence of preparation activities, knowing in advance the problems which can mainly compromise the results. Instead, [9] does not have any knowledge and it needs to compute the optimal sequence running the system from scratch every time. A method aimed to decrease the time needed for the data preparation phase is described in [10]. This work offers a toolkit, which provides a set of key quality metrics related to the context of a ML project. The goal is to assess a set of metrics that are able to detect the specific data issues connected to this field, e.g., class overlap, feature relevance or data homogeneity. Then, they consider the identified issues and build a specific pipeline to detect, explain and address such problems.

Recently, data quality has been also connected to the ethics context, since the fact that data are correct is a typical ethical requirement and, moreover, data should conform high ethical standard to be considered of high quality [4]. [4] has combined the most widely used ethical requirements with data quality, defining a set of ethical quality dimensions, i.e., data transparency, diversity and data protection and fairness. Another relevant work is [11] in which a wide variety of aspects related to bias and fairness in the field of machine learning are discussed. This work investigates different real-world examples that have been affected by biases and the different sources of bias in AI systems. Moreover, it creates a taxonomy grouping many existing definitions for fairness. This contribution highlights the fact that biases need to be considered in a data analysis pipeline as our approach suggests.

5. Conclusions and Future Work

This work addresses the problem of defining reliable guidelines to support users in the data preparation process. Results of the conducted experiments above led to an important conclusion: (i) it is confirmed that different data quality issues have different impact on the quality of the results depending both on the algorithm applied and the dataset features; (ii) following the guidelines leads to quality results in an efficient and effective way. In fact, results show that, following the proposed suggestions yields better results than applying the same preparation activities in a different order. This work opens up a number of new research opportunities. Human-in-the-loop (HITL) techniques should be better exploited in order to further improve the data preparation process and the provided guidelines. Future work will also better investigate issues related to bias: we aim to extend the data preparation techniques by considering bias mitigation techniques (e.g., data enrichment) in addition to data quality improvement techniques.

References

- [1] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, *Journal of Management Information Systems* 12 (1996) 5–33. URL: <https://doi.org/10.1080/07421222.1996.11518099>. doi:10.1080/07421222.1996.11518099. arXiv:<https://doi.org/10.1080/07421222.1996.11518099>.

- [2] C. Batini, M. Scannapieco, *Data and Information Quality - Dimensions, Principles and Techniques, Data-Centric Systems and Applications*, Springer, 2016. URL: <https://doi.org/10.1007/978-3-319-24106-7>. doi:10.1007/978-3-319-24106-7.
- [3] F. Naumann, J. C. Freytag, U. Leser, Completeness of integrated information sources, *Inf. Syst.* 29 (2004) 583–615. URL: <https://doi.org/10.1016/j.is.2003.12.005>. doi:10.1016/j.is.2003.12.005.
- [4] D. Firmani, L. Tanca, R. Torlone, Ethical dimensions for data quality, *ACM J. Data Inf. Qual.* 12 (2020) 2:1–2:5. URL: <https://doi.org/10.1145/3362121>. doi:10.1145/3362121.
- [5] L. Jost, Entropy and diversity, *Oikos* 113 (2006) 363–375.
- [6] M. Hameed, F. Naumann, Data preparation: A survey of commercial tools, *SIGMOD Rec.* 49 (2020) 18–29. URL: <https://doi.org/10.1145/3444831.3444835>. doi:10.1145/3444831.3444835.
- [7] Z. Qi, H. Wang, Dirty-data impacts on regression models: An experimental evaluation, in: C. S. Jensen, E. Lim, D. Yang, W. Lee, V. S. Tseng, V. Kalogeraki, J. Huang, C. Shen (Eds.), *Database Systems for Advanced Applications - 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11-14, 2021, Proceedings, Part I, volume 12681 of Lecture Notes in Computer Science*, Springer, 2021, pp. 88–95. URL: https://doi.org/10.1007/978-3-030-73194-6_6. doi:10.1007/978-3-030-73194-6_6.
- [8] Z. Qi, H. Wang, A. Wang, Impacts of dirty data on classification and clustering models: An experimental evaluation, *J. Comput. Sci. Technol.* 36 (2021) 806–821. URL: <https://doi.org/10.1007/s11390-021-1344-6>. doi:10.1007/s11390-021-1344-6.
- [9] L. Berti-Équille, Active reinforcement learning for data preparation: Learn2clean with human-in-the-loop, in: *10th Conference on Innovative Data Systems Research, CIDR 2020, Amsterdam, The Netherlands, January 12-15, 2020, Online Proceedings*, [www.cidrdb.org](http://cidrdb.org), 2020. URL: http://cidrdb.org/cidr2020/gongshow2020/gongshow/abstracts/cidr2020_abstract59.pdf.
- [10] N. Gupta, H. Patel, S. Afzal, N. Panwar, R. S. Mittal, S. C. Guttula, A. Jain, L. Nagalapatti, S. Mehta, S. Hans, P. Lohia, A. Aggarwal, D. Saha, Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets, *CoRR abs/2108.05935* (2021). URL: <https://arxiv.org/abs/2108.05935>. arXiv:2108.05935.
- [11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, *CoRR abs/1908.09635* (2019). URL: <http://arxiv.org/abs/1908.09635>. arXiv:1908.09635.