

A Novel Metric for Measuring Data Quality in Classification Applications

Jouseau Roxane, Salva Sébastien and Samir Chafik

*Université Clermont-Auvergne, CNRS, Mines de Saint-Etienne, Clermont-Auvergne-INP, LIMOS,
Clermont-Ferrand, France*

Keywords: Classification, Data Quality, Machine Learning, Measure, Metric.

Abstract: Data quality is a key element for building and optimizing good learning models. Despite many attempts to characterize data quality, there is still a need for rigorous formalization and an efficient measure of the quality from available observations. Indeed, without a clear understanding of the training and testing processes, it is hard to evaluate the intrinsic performance of a model. Besides, tools allowing to measure data quality specific to machine learning are still lacking. In this paper, we introduce and explain a novel metric to measure data quality. This metric is based on the correlated evolution between the classification performance and the deterioration of data. The proposed method has the major advantage of being model-independent. Furthermore, we provide an interpretation of each criterion and examples of assessment levels. We confirm the utility of the proposed metric with intensive numerical experiments and detail some illustrative cases with controlled and interpretable qualities.

1 INTRODUCTION

During the last decades, data availability has played a crucial role in the development and sophistication of artificial intelligence in general and machine learning models in particular. Yet, few works have been dedicated to deeply investigating the assessment on the apparent accuracy and consistency of data (Cichy and Rass, 2019; Batini et al., 2009; Batini et al., 2016; Pipino et al., 2002). Moreover, the term data quality has been restricted to studying the impact of some standard criteria on the task at hand and user's expectations. Standard criteria have been evaluated very often separately or with a non-rigorous combination of scores. For example, several methods use the term data quality for specific purposes, such as accuracy, completeness, or timeliness, for specific applications and contexts (Gudivada et al., 2017). Unfortunately, this leads to the first limitation related to data quality: The lack of an appropriate definition and, subsequently, an accurate measure.

Classical data evaluation is often related to a given context: External metadata, rules, trusted references, etc. Establishing or extracting these external elements is usually a long process, expert-dependent and error-prone (Ehrlinger and Wöß, 2022). Instead of developing new methods, some previous papers cite commercial products used to measure data quality. A common conclusion states that only a few tools are available.

We believe that this is a consequence of the limitations detailed above. Nevertheless, we have tried to test some of them for the same context, but they were either out of the scope of our study, or we were unable to set up a usable configuration.

In this work, we solve the previous limitations by introducing an original data quality metric. We focus on the use of data in artificial intelligence applications and, more specifically, for learning models in numerical classification. The proposed metric has the significant advantage of providing a consistent means to evaluate the quality of different types of data with different numbers of classes, various domains, and from low to high dimensionality, etc. All the steps were constructed carefully to make the metric interpretable, easy to use, and model-independent. To illustrate how the metric can successfully capture a deterioration, we have shown that the impact makes the classification performance decrease non-linearly with different rates. This is against a non-justified prior when, in some previous works, they have assumed a linear behavior.

1.1 Related Work

Only a few tools are available to directly evaluate the quality of a dataset with a metric, as most tools only offer indicators to monitor and help with data profiling. In the recent survey (Ehrlinger and

Wöß, 2022) the authors evaluated 11 tools: Apache Griffin (Foundation, 2023), Ataccama ONE (Ataccama, 2023), DataCleaner (DataCleaner, 2023), Datamartist (Datamartist, 2023), Experian Pandora (Experian, 2023), InformaticaDQ (Informatica, 2023), InfoZoom & IZDQ (InfoZoom, 2023), MobyDQ (Roland, 2023), OpenRefine (OpenRefine, 2023) & MetricDoc (Bors et al., 2018), SAS Data Quality (SAS, 2023), and Talend Open Studio (Talend, 2023). These tools were classified into 5 categories: accuracy, completeness, consistency, timeliness, and others. Interestingly, most of these tools allow the evaluation of one or two criteria only. They mainly focus on easing the definition of data indicators and assisting data profiling. Ultimately, only 4 tools were distinguished: Apache Griffin, InformaticaDQ, MobyDQ, and MetricDoc.

While many of these tools focus on the attribute level, they are not generalized for higher levels of aggregation. They also considered data rules, which are not considered here because they require expert knowledge and are often unavailable. Additionally, Apache Griffin and MobyDQ require a reference dataset. This makes them less practical as ground truth reference datasets are not always available. InformaticaDQ focuses on textual data, such as elements of postal addresses, email addresses, etc., and cannot be applied to numerical data. MetricDoc offers two different time interval metrics for time-series data, a redundancy metric on the table level and metrics for validity and plausibility, both defined at the attribute level only. Finally, we tried to investigate the Data Quality for AI API proposed by IBM (IBM, 2023). However, despite taking steps to access the free trial version on the website, we could not secure working access to the API, which is a problem that the authors of (Ehrlinger and Wöß, 2022) also seem to have faced, as mentioned in their paper.

1.2 Contributions

In order to build a metric that is independent of learning models, we formulate the problem such that no reference or expertise is needed. Instead, our metric is based upon two main terms: The former evaluates classification performance across a wide range of models, and the latter assesses variations of performance when a low amount of errors is injected into datasets. We show that high variations are consequences of quality issues. All terms are then empirically combined to form a unique evaluation, denoted q_a . Furthermore, we show how to interpret our metric scores and express the notions of good, medium, and bad qualities. Then, we evaluate the metric q_a on 110

datasets of known quality and show that q_a is able to measure the quality correctly. We also discuss the information that can be captured by q_a , leading to an easy connection with a given result.

To the best of our knowledge, this is the first method that proposes a clear and rigorous formulation for measuring data quality. We summarize the main contributions of this work as follows: We propose a novel rigorous metric to evaluate data quality for numerical classifications. Our method can be adapted and extended to regression. The proposed metric is model-independent. This proposed metric does not require external elements or expert supervision. We use a constructive approach, step by step, to formulate the metric that could be generalized to other contexts.

The rest of the paper is organized as follows: Section 2 proposes a definition for the data quality metric and some thresholds to ease interpretation. Then, the metric is evaluated on 155 datasets of varied levels of quality in different contexts. Possible threats to validity are also discussed in Section 3. Finally, we conclude the paper in Section 4.

2 MEASURING DATA QUALITY

In (Jouseau et al., 2022), we investigated the relevance of repairing datasets according to the impact of the amount of errors included in datasets on the classification performance. This study allowed the observation of some distinct characteristics related to data quality. In particular, we observed the two following properties:

1. Model accuracy decreases along with data quality when errors are injected into data. This decrease varies across classification models and numbers of classes;
2. The decrease in accuracy is nonlinear. It is low when data is of good quality or when data is extremely deteriorated. However, the decrease in accuracy is significantly higher between these two states.

We illustrate these observations on three datasets (Iris, Breast cancer, Adult). We chose 12 standard classification models available in (Pedregosa et al., 2011): Logistic regression, K-Nearest Neighbors, Decision tree, Random forest, Ada boost, Naive Bayes, XGboost, Support vector classification, Gaussian process, Multi-layer perceptron, Stochastic gradient descent, and Gradient boosting. We also illustrate these observations for the error types missing values, outliers, and fuzzing, a.k.a. partial duplicates. We chose these error types because we observed that they have different impacts on model performance: Outliers and

missing values have the most impact on accuracies and f1 scores; fuzzing tends to have less impact and offers the benefit of simulating data generation. In Figure 1, we present mean accuracies computed over 30 iterations of injecting controlled percentages of errors, randomly generated with a uniform distribution, in training data. We inject up to 95% of errors with a 5% increment. The 12 classification models are then trained on these deteriorated datasets with a random split using 80% for training and 20% for testing. Figures 1a, 1b, and 1c present the mean accuracies when missing values, outliers, and fuzzing are respectively injected in training data.

In Figures 1a and 1b, as per our first observation, accuracies are high when none or a low percentage of errors are present in the data. Unsurprisingly, these accuracies decrease as data deterioration increases. But, as we stated in our second observation, the decrease in accuracy is non-linear. This is especially visible in Figure 1a. We can also see in Figures 1a and 1b that the mean accuracy can drop significantly with only a 5% increment of errors. For instance, in Figure 1a, for the dataset Iris, the mean accuracy stays over 0.8 up to the injection of 35% of missing values into training data. However, at 40% of missing values, the mean accuracy drops to 0.6, approximately. The data fuzzing observed in Figure 1c is a particular case with no loss of information, which is why accuracies stay quite steady.

The analysis and formalization of these observations allow us to propose the first data quality measurement using the notion of data deterioration. This metric, denoted q_a , is composed of two parts, denoted $q_{a,1}$ (Eq. 4) and $q_{a,2}$ (Eq. 7), which respectively encode these characteristics:

- $q_{a,1}$: the accuracies across a set of classification models (Observation 1). $q_{a,1}$ also accounts for the number of classes in datasets in order to allow comparisons of data quality levels between datasets with different numbers of classes;
- $q_{a,2}$: variations of accuracies when a low percentage of errors are injected in training sets (Observation 2). It aims to capture abnormally high accuracy variations over small dataset perturbations.

Next, we formalize these two observations in a rigorous way.

2.1 Definition of $Q_{a,1}$

We use the following notations in the remainder of the paper: The set of models is denoted M , the set of error types is denoted E , and D is the dataset under evaluation. $A(m, D)$ stands for the accuracy of the model

$m \in M$ on D . We define $A_M(D)$ as :

$$A_M(D) := \frac{1}{\text{card}(M)} \sum_{m \in M} A(m, D) \quad (1)$$

Observing the mean accuracy $A_M(D)$ alone is not sufficient to express data quality. Moreover, given the number of classes c in D , we want an accuracy that is better than a random choice:

$$1 \geq A_M(D) > \frac{1}{c}, \quad c > 1 \quad (2)$$

When the accuracy of a model is lower than $\frac{1}{c}$, we consider the quality to be the lowest. The function δ_1 capture this statement:

$$\delta_1(A_M(D)) := \begin{cases} 1 & \text{if } A_M(D) > \frac{1}{c} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Finally, we define $q_{a,1}$ as:

$$q_{a,1}(D) := 1 - \frac{cA_M(D) - 1}{c - 1} \delta_1(A_M(D)) \quad (4)$$

We have $0 \leq q_{a,1}(D) \leq 1$, with $q_{a,1}(D) = 0$ as the best quality.

2.2 Definition of $Q_{a,2}$

The main idea encoded by $q_{a,2}$ is that the mean accuracies of models trained on good or bad-quality datasets are not sensitive to a small data deterioration. This is not the case for datasets of medium quality. We verify this hypothesis by computing variations of accuracy with the set of classification models M when a percentage p of an error type $e \in E$ is injected in D . We denote $D_{e,p}$ the resulting dataset.

For $\Delta_{A_{M,e}}(D)$ to express the variation of accuracy for a specific error e , we introduce:

$$\Delta_{A_{M,e}}(D) := \frac{1}{\text{card}(M)} \sum_{m \in M} |A(m, D) - A(m, D_{e,p})| \quad (5)$$

To avoid any bias in our metric, we assume that errors are injected randomly, with a uniform distribution in training data. In our experiments, we noticed in Figure 1a and 1b that a small percentage of errors $p = 5\%$ is sufficient to capture accuracy variations. According to our experiments, a higher value of p is possible but may lead to less precise measurements of the performance loss. The goal of this parameter is to simulate small perturbations in data that can happen in real-life situations. Although small variations of accuracies are expected, we are only interested in abnormal variations. To exclude these minor variations, we define δ_2 as:

$$\delta_2(\Delta_{A_{M,e}}(D)) := \begin{cases} 1 & \text{if } \Delta_{A_{M,e}}(D) > p \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

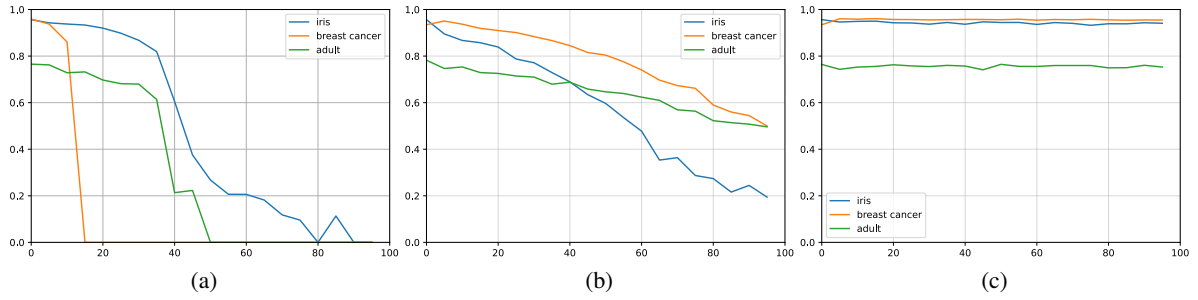


Figure 1: Evolution of the mean accuracy as a function of missing values (a), outliers (b), and fuzzing (c).

$q_{a,2}$ is then defined as:

$$q_{a,2}(D) := \min\left(\frac{10}{\text{card}(E)} \sum_{e \in E} \Delta A_{M,e}(D) \delta_2(\Delta A_{M,e}(D)), 1\right) \quad (7)$$

We chose to add a factor of 10 in $q_{a,2}(D)$ because we consider that an accuracy variation of 10% or more when we inject a small amount of error in data, indicates bad data quality. However, this factor does not keep the result bounded by 1, so we use the minimum function to define $q_{a,2}$. It is worth noting that this parameter may be easily changed to meet user preferences.

2.3 Definition of the Quality Metric Q_a

We define q_a as:

$$q_a(D) := \max(q_{a,1}(D), q_{a,2}(D)) \quad (8)$$

We use the maximum to ensure that the metric captures the most variation of data quality, given by $q_{a,1}$ and $q_{a,2}$. We are now ready to present, in Algorithm 1, all the steps required for evaluating the quality of D with $q_a(D)$. If no trusted test is available, we compute $q_a(D)$ over 30 resamplings of test and train. We studied the validity of this method and presented it in an extended version of this paper (Jouseau et al., 2023b). $q_{a,1}$ and $q_{a,2}$ are then computed as the means of the $q_{a,1}$ and $q_{a,2}$ obtained from the resampled datasets.

2.4 Interpretation

In this section, we discuss the interpretation of q_a . More specifically, we propose to extract thresholds for q_a that express the notions of good, medium, or bad quality. To do this, we computed q_a on 114 datasets derived from Iris, Breast Cancer, and Adult when errors are injected in 5% increments from 0% to 95%. We only considered the error types missing values and outliers, as fuzzing does not tend to affect classifier performance. Since we control the level of errors, we estimate the different levels of quality for these data-

Data: Dataset D

Result: $q_a(D)$ and its interpretation.

if D is made up of a trusted test dataset **then**

$LD = (D)$;

else

 Generate the list $LD = (D^1, \dots, D^{30})$ of resampled versions of D ;

end

foreach $D_i \in LD$ **do**

 Compute $q_{a,1}(D^i)$ as defined in Eq. 4;

foreach error type $e \in E$ **do**

 Create a new dataset $D_{e,p}^i$ by injecting D^i with $p\%$ of error e , randomly generated with a uniform distribution;

end

 Compute $q_{a,2}(D^i)$ with Eq.(7);

end

$$q_{a,1}(D) = \frac{1}{30} \sum_{i=1}^{30} q_{a,1}(D^i);$$

$$q_{a,2}(D) = \frac{1}{30} \sum_{i=1}^{30} q_{a,2}(D^i);$$

 Compute $q_a(D)$ with Eq.(8);

 Interpret $q_a(D)$;

Algorithm 1: $q_a(D)$ computation.

sets. Iris and Breast Cancer are considered good quality when used with classification models in the literature, while the Adult is of medium quality. We expect data quality to be good for the datasets obtained from Iris and Breast Cancer with up to 10% of errors, medium between 10% and 30%, and bad for over 30% of errors. For the datasets obtained from Adult, we expect medium quality up to 10 % of errors and then bad quality.

We present, in the figures 2a, 2b, the evolution of q_a for three dataset examples when controlled percentages of missing values, outliers, and fuzzing are respectively injected with 5% increments. Colors on the y axis depict the proposed thresholds for which q_a

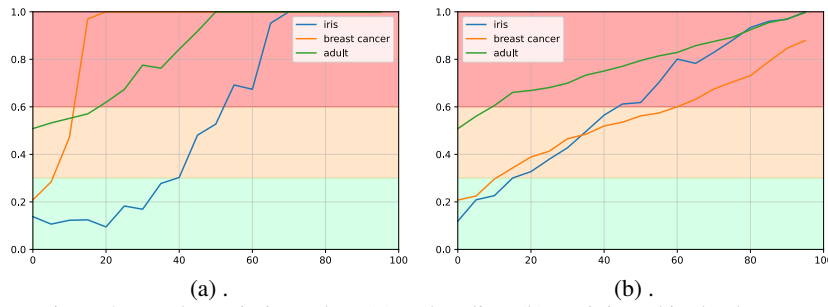


Figure 2: q_a when missing values (a) and outliers (b) are injected in the datasets.

indicates good, medium, or bad data quality.

These thresholds were empirically chosen based on the following observations. In Figure 2, we see that without any deterioration, q_a measurements for these datasets are respectively 0.11, 0.2, and 0.5. Thus, we expect 0.11 and 0.2 to indicate good quality and 0.5 to indicate medium quality. Furthermore, in Figure 2b for around 10% of errors, when data quality is expected to be altered from good to medium, we observe that values of q_a are close to 0.3 for the datasets Breast Cancer and Iris. This prompts us to set the upper threshold for good quality at 0.3, which is consistent with our first observations. Additionally, in Figure 1a, we observe that the mean accuracy for the dataset Adult starts to decrease significantly between 15% and 20% of missing values injected. This corresponds to q_a measurements between 0.55 and 0.7 in Figure 1a. From these elements, we propose to set the upper threshold for medium quality to 0.6. We find this to be a reasonable limit since, if we take the case of a dataset with two classes, being over this threshold either means that its mean accuracy is below 0.7 or that it experiences a mean variation of accuracy over 0.18 when injected with 5% of errors. These requirements either indicate a mean accuracy low enough or variations of accuracy high enough to characterize a dataset as bad quality.

We, therefore, propose the following thresholds to interpret q_a : if $q_a(D) \leq 0.3$, this means that D is of good quality; If $0.3 < q_a(D) \leq 0.6$ D can be considered to be of medium quality, and examining the values of $q_{a,1}(D)$ and $q_{a,2}(D)$ is necessary to decide whether D can be used. Finally, $0.6 < q_a(D)$ means that D is of bad quality.

However, in Figure 2a, we can see that for 20% of missing values, we measure $q_a(\text{Iris}) < 0.3$. This result is unexpected since, for this percentage of errors, we do not expect good data quality. Nonetheless, it is consistent with Figure 1a where we can observe that at 20% of missing values, accuracy is still high with Iris and is not about to experience a significant drop.

3 EMPIRICAL EVALUATION

The experiments presented in this section aim to evaluate the relevance of q_a . For the remainder of the paper, we express this relevance through 2 questions:

- Q1: Can q_a characterize a dataset of good or medium quality?
- Q2: Can q_a characterize a dataset of bad quality?

3.1 Empirical Setup

First, we use the classification model set M listed in Section 2 along with the three error types: missing values, outliers, and fuzzing. We also set the percentage of injected errors $p = 5\%$ as discussed in Section 2.

We evaluated q_a on 155 numeric datasets. We used five distinctive datasets: Spambase, Heart Disease, Abalone, Dry Beans, and Statlog (Markelle Kelly, 1999), as well as 150 datasets we modified by injecting controlled amounts of errors. We selected these datasets for their varied dimensions, number of classes, number of attributes, and domains of application. The objective of the Spambase dataset is to predict whether an email is a spam. The Heart Disease dataset is used to predict the presence of heart diseases in patients. The Abalone dataset is used for the prediction of the number of rings present in abalone shells (which indicates their ages) from physical measurements. The class imbalance in this dataset is too high to achieve reasonable accuracies. Therefore, we chose to work on a more straightforward classification task by aggregating the classes into two groups: up to 8 shell rings and over eight shell rings. The Dry Beans dataset is used for the prediction of the varieties of dry beans from features by the market situation. Finally, the dataset Statlog is used to classify people described by attributes as good or bad credit risks.

We manually evaluated the quality of these five datasets by measuring 7 dimensions and characteris-

tics. The results, along with our quality estimations, are provided in Table 1. The dataset Spambase is estimated to be of good quality as its number of samples (4 601) is relatively high even for its less populated class (1 813) compared to its number of attributes and classes (57 and 2). Besides, its mean accuracy across classification models is high (0.9). Datasets Abalone, Statlog, Dry Beans, and Heart Disease are estimated to be of medium quality mainly because their mean accuracies are lower (respectively 0.84, 0.76, 0.68, and 0.79). Additionally, the datasets Abalone and Dry Beans present high levels of class imbalances, which are usually considered quality issues. The mean accuracy for the dataset Dry Beans can seem very low compared to the other evaluation datasets, but in the context of a seven-class classification problem, it is, in fact, rather high.

105 additional datasets were derived from the five previous ones by injecting missing values, outliers, and fuzzing, separately and randomly with a uniform distribution. Two strategies were followed: injection of 5 and 10% of errors to build datasets of good or medium qualities, and injection of 30 up to 50% of errors with increments by 5% to build datasets of bad quality. Datasets, results, and a prototype version of our tool allowing to compute q_a are available in (Jouseau et al., 2023a).

3.2 Q1: Can Q_a Characterize a Dataset of Good or Medium Quality?

To answer this question, we computed q_a for the five datasets presented in Table 1 as well as for the 30 datasets created by injecting 5% and 10% of missing values, outliers, or fuzzing. In Table 2, we present q_a for the five datasets without deterioration, q_a is presented with the details of $q_{a,1}$, $q_{a,2}$, and the corresponding data quality levels defined in Section 2.4. q_a was computed over the 12 classification models and 30 resamplings of training and test data. We compare these quality levels with the data quality estimations given in Table 1.

We observe with Tables 2 and 1 that q_a evaluates the data quality of the five first datasets correctly. The only dataset evaluated as of good quality by q_a is Spambase. We can also see that its values for $q_{a,1}$ and $q_{a,2}$ are low. This means that the classification models perform with high levels of accuracy ($q_{a,1}$) and that these performances would not drop significantly upon small perturbations of the dataset ($q_{a,2}$). Indeed, its mean accuracy is high (0.9), and even if the dataset presents a class imbalance, its least populated class is still relatively populated (1 813 samples). The other datasets in Table 2 are all classified

as medium quality by q_a . Again, this is consistent with our estimated data qualities given in Table 1. The quality issues are different, though. For instance, for the dataset Abalone, $q_{a,1}$ is relatively high, but $q_{a,2}$ is low. This means that the accuracy across models is not very high, but its accuracy does not vary much when 5% of errors are injected in the training set. Moreover, the mean accuracy across classification models for the dataset Abalone is 0.84 (Table 1). Since the dataset only has two classes, this is not considered a very high accuracy. However, depending on the application, it can be regarded as reasonable.

Figure 3a, 3b, and 3c now respectively illustrate the values of q_a when injecting missing values, outliers, and fuzzing in the previous datasets with a 5% increment from 0 to 50%. For this question, we only focus on the 30 datasets obtained when 5% and 10% of errors are injected. We expect them to be of medium data quality. The colored zones in Figure 3 correspond to the thresholds we set in Section 2.4, for which quality is considered good, medium, or bad.

We observe in these figures that q_a indicates a good or medium quality level for 28 of these 30 datasets. In Figure 3a, q_a indicates a bad quality for the datasets Spambase and Statlog at 10% of missing values injected. For the dataset Statlog, with 10% of missing values, we observed that the mean accuracy score is equal to 0.7. This dataset also presents a class imbalance that can be worsened by the injection of missing values. Its quality evaluation is, therefore, consistent with our observations. For Spambase, the mean accuracies of the datasets with 0, 5, and 10% of missing values are 0.9, 0.85, and 0.56. These results tend to show that the dataset with 10% of errors is of bad quality for classification tasks. In Figure 3b, we observe that q_a indicates medium data quality for all datasets between 0% and 10% of missing values, which is consistent with our expectations.

It is worth noting that, in Figure 3c, q_a stays relatively constant, which is what we expect.

Out of the 35 datasets studied in this section, q_a is indicative of either good or medium data quality for 33 of them, which is the result we expected. For the 2 remaining cases where q_a scores are indicative of bad quality, we studied carefully the datasets and observed that q_a is correct. Thus, we conclude that q_a is able to characterize datasets of good or medium quality.

3.3 Q2: Can Q_a Characterize a Dataset of Bad Quality?

To investigate this question, we computed q_a for the 50 datasets obtained after injecting controlled per-

Table 1: Overview of the evaluation datasets.

Dataset	Number of classes	Samples total	Number of attributes	Features	Class imbalance	Missing data	Mean accuracy	Estimated data quality
Spambase	2	4 601	57	integers, reals	Yes (1 813 samples in the least populated class)	None	0.90	good
Abalone	28 (2 post-processing)	4 177	8	categorical, integers, reals	Yes (1 407 samples in the least populated class post-processing)	None	0.84	medium
Dry Beans	7	13 611	16	categorical, integers, reals	Yes (522 samples in the least populated class)	None	0.68	medium
Statlog	2	1 000 (959 post-processing)	23	integers	Yes (275 samples in the least populated class)	on 41 samples	0.76	medium
Heart Disease	5 (2 post-processing)	303 (297 post-processing)	13	categorical, integers, reals	Yes (No after post-processing)	on 6 samples	0.79	medium

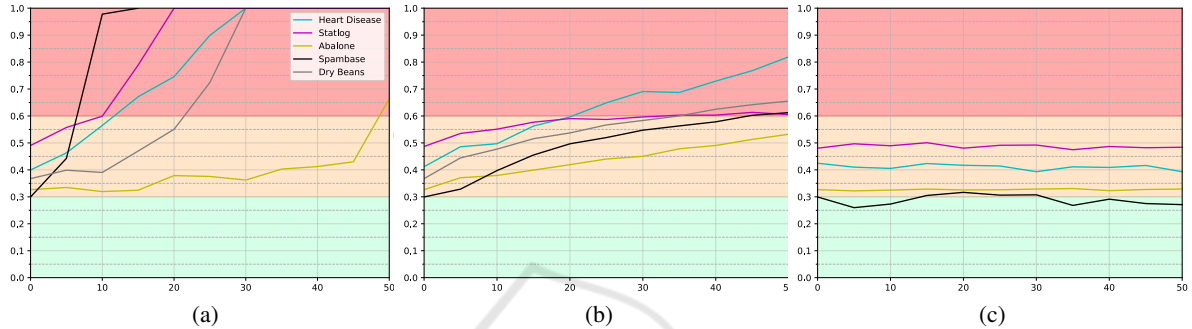

Figure 3: q_a when missing values (a), outliers (b), and fuzzing (c) are injected in all datasets.

Table 2: q_a calculation and the related quality levels.

Dataset	q_a	$q_{a,1}$	$q_{a,2}$	Quality level
Spambase	0.29	0.18	0.29	good
Abalone	0.32	0.32	0.02	medium
Dry Beans	0.36	0.36	0.17	medium
Statlog	0.48	0.48	0.13	medium
Heart Disease	0.42	0.42	0.11	medium

centages of missing values and outliers in the initial datasets presented in Table 1. We injected from 30 to 50% of errors with a 5% increment. At this level of errors for missing values and outliers, we assumed that the 50 resulting datasets were of bad quality. We do not consider fuzzing here as we expected that the injection of this error type should keep the quality close to the quality of the original datasets. The computed q_a scores are again presented in Figure 3.

Out of the 50 datasets studied in this section, q_a indicated bad data quality for 37 of them. For example, we can see in Figure 3a that when 30% of missing values are present in datasets, q_a correctly indicates bad data quality, except for the dataset Abalone up to 45%. We observe similar results for outliers. We hence investigated why q_a indicates medium data quality for the 9 datasets obtained from Abalone after injecting 30 to 45% of missing values or outliers. The mean accuracies measured from these datasets stay over 75% and then drop (abruptly in Figure 3a). This is actually consistent with the Abalone properties given in Table 1, where we observe that Abalone has a high number of samples (4 177) for a low num-

ber of classes and attributes (2 and 8). Hence, our own interpretation of quality was wrong for these 9 datasets. We can indeed estimate that these datasets are of medium quality with regard to these characteristics. The mean accuracy of the dataset Spambase stays over 0.7 up until the injection of 45% of outliers. Its quality evaluation without the injection of errors was good (as opposed to medium for the other evaluation datasets). This could explain its higher resistance to outliers. Moreover, for the dataset Dry Beans at 30% of outliers, the mean accuracy is evaluated at 0.5, which seems low but is still much better than a random guess with seven classes ($\frac{1}{7}$).

In summary, q_a indicates bad data quality for 37 datasets. For the 13 cases where q_a indicates medium data quality when bad data quality was expected, we observed that the data characteristics and mean accuracies were consistent with a medium data quality level. These observations allow us to conclude that q_a is able to characterize datasets of bad quality.

3.4 Threats to Validity

In this section, we address 8 possible threats to the validity of this study. We identified 3 internal threats and 4 external threats.

The internal threats we identified are the implementation of the classification models, the hyper-parametrization of the classification models, and the number of datasets used to define the thresholds to

interpret q_a . To address the first threat, we implemented classification models using scikit-learn (Pedregosa et al., 2011), a widely used library. To limit the second threat, we used a grid search to set the hyper-parameters for classification models on all datasets without any deterioration and then used these settings for the rest of the experiments. To limit the third threat, we selected three widely used datasets, along with the 114 datasets obtained by injecting controlled percentages of errors.

The 4 external threats are the choice of the datasets for the evaluation, the choice of classification models, the generation of errors, and the combination of errors. We tried to limit the first threat by choosing datasets that are widely used and have different dimensions. We also selected datasets that cover various applications and ranges of dimensions. We selected a wide range of classification approaches to address the second threat. To limit the third one, we decided to generate errors randomly, with a uniform distribution. Finally, we decided to study errors separately to limit the fourth threat. However, we plan to extend this work to error combinations in future work.

4 CONCLUSION

In this paper, we have introduced a novel metric to measure data quality. The main advantage of the proposed metric is being independent of learning models and expert knowledge. Furthermore, it does not require external reference data. As a consequence, it offers the possibility to compare different datasets. We have extensively tested and evaluated the proposed metric and have shown that it is able to characterize data quality correctly.

REFERENCES

- Ataccama (2023). Ataccamaone. <https://www.ataccama.com/platform>.
- Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. In *ACM computing surveys*.
- Batini, C., Scannapieco, M., et al. (2016). *Data and information quality*. Springer.
- Bors, C., Gschwandtner, T., Kriglstein, S., Miksch, S., and Pohl, M. (2018). Visual interactive creation, customization, and analysis of data quality metrics. In *Journal of Data and Information Quality (JDIQ) ACM*.
- Cichy, C. and Rass, S. (2019). An overview of data quality frameworks. In *IEEE Access*.
- DataCleaner (2023). Datacleaner. <https://datacleaner.github.io/>.
- Datamartist (2023). Datamartist. <http://www.datamartist.com/>.
- Ehrlinger, L. and WöB, W. (2022). A survey of data quality measurement and monitoring tools. In *Frontiers in big data*.
- Experian (2023). User manual version 5.9. <https://www.edq.com/globalassets/documentation/pandora/pandora\manual\590.pdf>.
- Foundation, A. (2023). Apache griffin user guide. <https://github.com/apache/griffin/blob/master/griffin-doc/ui/user-guide.md>.
- Gudivada, V., Apon, A., and Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. In *International Journal on Advances in Software*.
- IBM (2023). Ibm data quality for ai api. <https://developer.ibm.com/apis/catalog/dataquality4ai--data-quality-for-ai/Introduction>.
- Informatica (2023). What is data quality? <https://www.informatica.com/resources/articles/what-is-data-quality.html>.
- InfoZoom (2023). Infozoom & izdq. <https://www.infozoom.com/en/products/infozoom-data-quality/>.
- Jouseau, R., Salva, S., and Samir, C. (2022). On studying the effect of data quality on classification performances. In *23rd International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*. Springer.
- Jouseau, R., Salva, S., and Samir, C. (2023a). Additional resources for the reproducibility of the experiment. <https://gitlab.com/roxane.jouseau/measuring-data-quality-for-classification-tasks>.
- Jouseau, R., Salva, S., and Samir, C. (2023b). A novel metric for measuring data quality in classification applications (extended version). <https://arxiv.org/abs/2312.08066>.
- Markelle Kelly, Rachel Longjohn, K. N. (1999). The uci machine learning repository. <https://archive.ics.uci.edu>.
- OpenRefine (2023). Openrefine. <https://github.com/OpenRefine/OpenRefine>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. In *Journal of Machine Learning Research*.
- Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. In *Communications of the ACM*.
- Rolland, A. (2023). Mobydq. <https://ubisoft.github.io/mobydq>.
- SAS (2023). Dataflux data management studio 2.7: User guide. <http://support.sas.com/documentation/onlinedoc/dfdmstudio/2.7/dmpdmsug/dfUnity.html>.
- Talend (2023). Talend open studio for data quality – user guide 7.0.1m2. http://download-mirror1.talend.com/top/user-guide-download/V552/TalendOpenStudio_DQ_UG_5.5.2.EN.pdf.