LABORATOIRE D'INFORMATIQUE,
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES

agaetis

A Novel Metric for Measuring Data Quality in Classification Applications

Roxane Jouseau, Sébastien Salva,
and Chafik Samir

Data Quality

"Fitness for the task at hand" → Context dependent, difficult to generalize

Few data quality concepts:

accuracy, completeness, timeliness, and consistency

However, there is **no unified measure** for these concepts

Plenty of work on monitoring data indicators, identifying data errors, and data cleaning but no proposition of a unified metric yet, even for families of tasks such as classification tasks.

Context of Our Work

Structured, numeric data for classification tasks

No metadata,

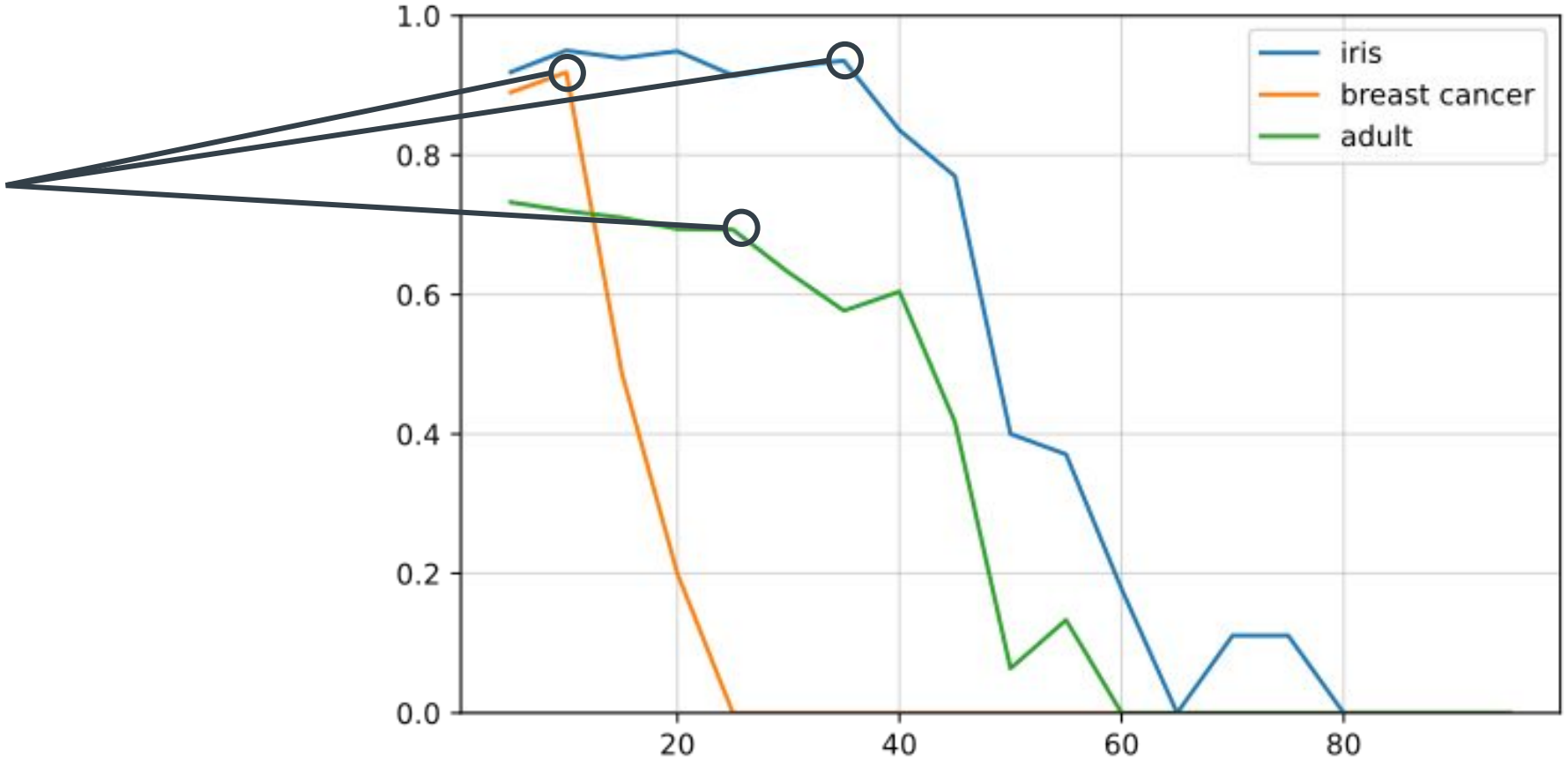
No expert knowledge

Outline

Measuring Data Quality

Example of data quality issues not identified by accuracy score

Accuracy as a function of the percentage of missing values in datasets



$$q_a = \max(q_{a,1}, q_{a,2})$$

- $q_{a,1}$: Accuracy with regards to the accuracy of a random classifier
- $q_{a,2}$: Variations of accuracy when 5% of errors are injected in data

Computed over 12 classification models
(logistic regression, k-nearest neighbors, decision tree, random forest, ada boost, naive bayes, xgboost, support vector classification, gaussian process, multi-layer perceptron, linear model with stochastic gradient descent, gradient boosting)

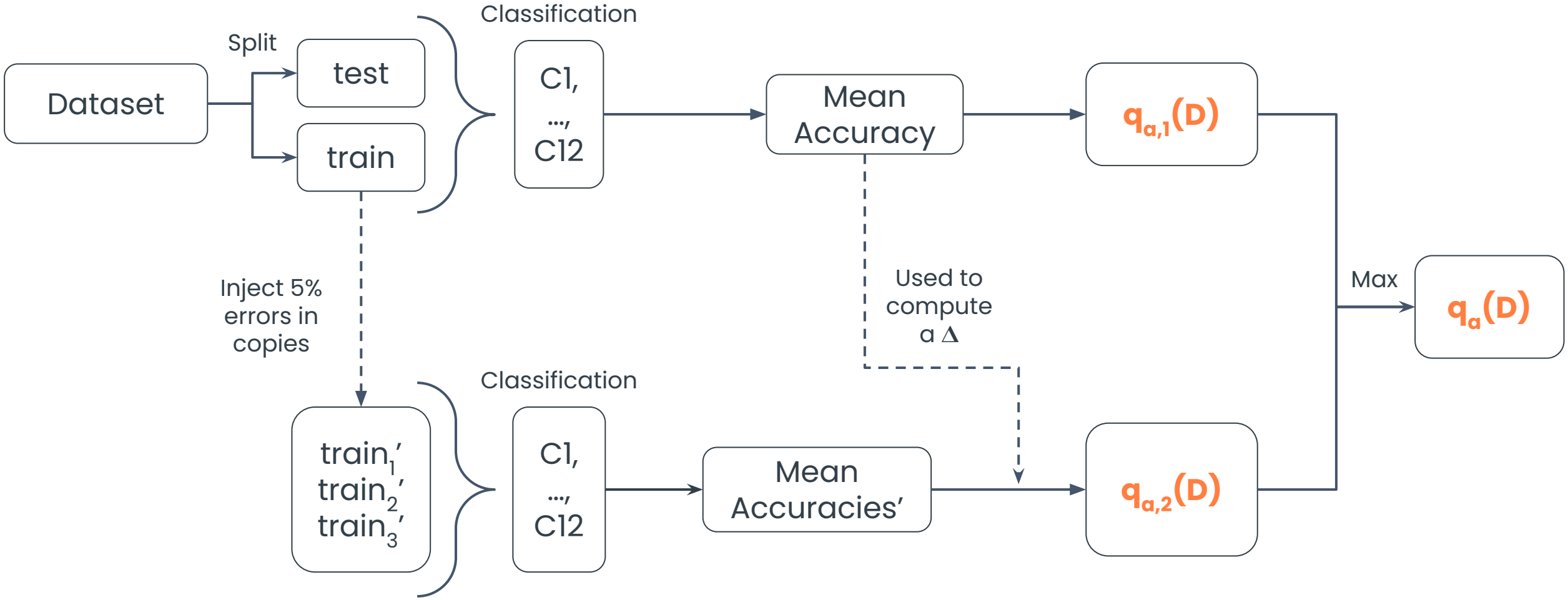
Considering 3 errors: missing values, outliers, and fuzzing injected with a random uniform distribution

$$0 \leq q_a(D) \leq 1$$

$q_a(D) = 0$ best data quality

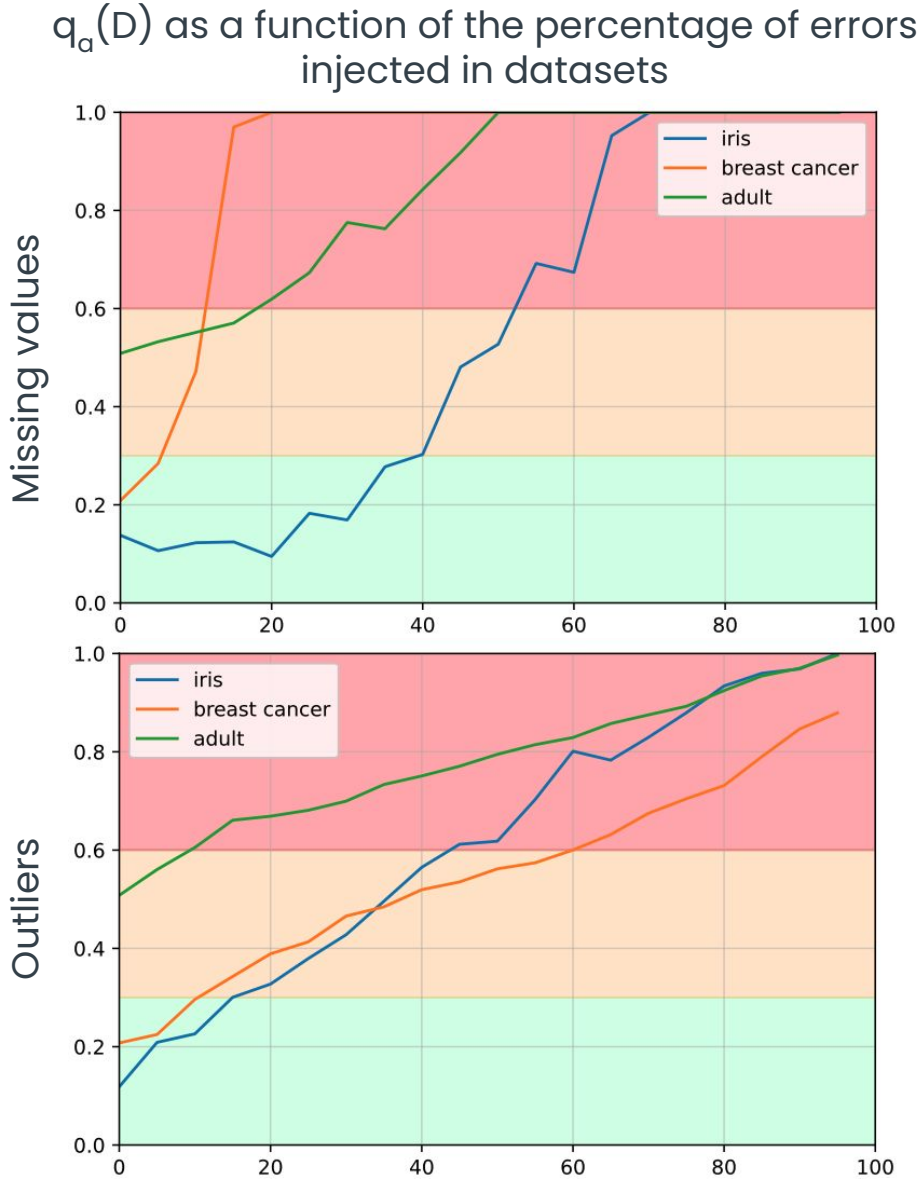
$q_a(D) = 1$ worst data quality

Concepts Behind the Metric



Empirical data quality thresholds based on manual data quality evaluation:

- $0 \leq q_a(D) \leq 0.3$: Good
- $0.3 < q_a(D) \leq 0.6$: Medium
- $0.6 < q_a(D) \leq 1$: Bad



Evaluation of the Metric

Empirical Setup for Evaluation

Start with 5 datasets,

Create 150 datasets by artificially deteriorating datasets through the injection of missing values, outliers, and fuzzing.

We inject these errors separately, in a random uniform way, in 5% increments from 0% to 50%.

Name	Objective	# Samples
Statlog	Predict if a credit risk is good or bad (german credit data)	959
Spambase	Predict if emails are spam	4 601
Abalone	Predict if Abalone shells have 8 or less rings from diverse measures	4 177
Heart Disease	Predict whether or not patients have heart diseases	297
Dry Beans	Predict the type of dry beans from descriptive and contextual data	13 611

Evaluation

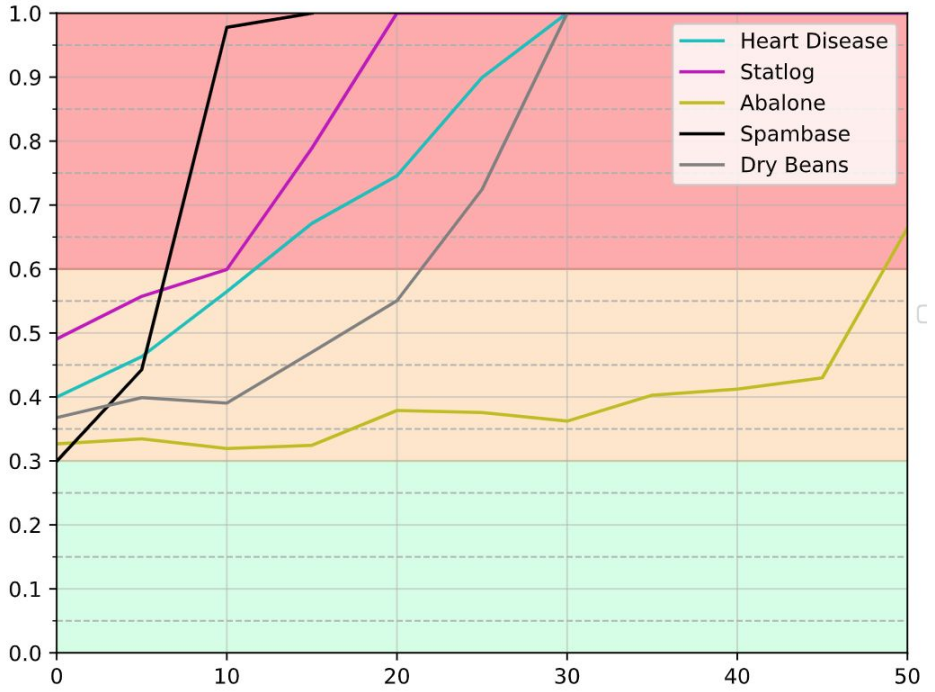
We compare interpretations of q_a with manual evaluations of data quality for the datasets:

- q_a was correct for 83% of datasets
- When q_a was incorrect it was close to interpretation thresholds

q_a correctly quantify data quality levels in most cases, allows comparison between datasets

q_a does not take into account class imbalance, more details are needed close to thresholds

$q_a(D)$ as a function of the percentage of missing values



In additional work we showed that q_a can be computed without a trusted test set through the mean of 30 resamplings

Conclusion and Future Work

Conclusion:

- Proposed a data quality metric q_a
- Proposed an interpretation of q_a
- Evaluation showed that q_a characterize data quality correctly in most cases

Future Work:

- Extension to other performance evaluations (e.g. F1 score)
- Work on measuring data repairability

References

- Azeroual, O., Saake, G., & Abuosba, M. (2018). Data Quality Measures and Data Cleansing for Research Information Systems. *Journal of Digital Information Management*.
- 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*.
- Ehrlinger, L. and Wo ̈ß, W. (2022). A survey of data quality measurement and monitoring tools. In *Frontiers in big data*.
- 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*.
- Haegmans, T., Snoeck, M., & Lemahieu, W. (2016). Towards a precise definition of data accuracy and a justification for its measure. *Proceedings of the International Conference on Information Quality*.
- Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining \& Knowledge Management Process*.
- 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.
- Lettner, C., Stumptner, R., Fragner, W., Rauchenzauner, F., & Ehrliger, L. (2021). DaQL 2.0: Measure Data Quality based on Entity Models. *Procedia Computer Science*.
- Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. In *Communications of the ACM*.
- Rosli, M. M., Tempero, E., & Luxton-Reilly, A. (2018). Evaluating the quality of datasets in software engineering. *Advanced Science Letters*.
- Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A Framework for Analysis of Data Quality Research. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*.

Thank You For Your Attention

agaetis

LIMOS

LABORATOIRE D'INFORMATIQUE,
DE MODÉLISATION ET D'OPTIMISATION DES SYSTÈMES