From Data Imputation to Data Cleaning – Automated Cleaning of Tabular Data Improves Downstream Predictive Performance

Sebastian Jäger [1] Felix Bießmann [1, 2]

[1]Berlin University of Applied Sciences and Technology (BHT)

[2]Einstein Center Digital Future

Introduction and Problem Setting

Controlling data quality remains one of the most impactful and difficult-to-automate parts of ML applications [1]. Here, we focus on one of the most common and relevant use cases of ML applications: we assume that an *ML model was trained on clean data*, and at inference time, the data quality deteriorates, impacting the predictive performance.

For missing values, we already showed [3] that using ML-based approaches to capture statistical dependencies between columns are efficient. In this study, we combine these imputation methods with conformal prediction (CP) to automatically detect and clean erroneous cells of heterogeneous tabular data.

Research Question and Hypothesis

Can calibrated ML models reliably and without manual intervention predict whether a single cell is erroneous and clean it if necessary?

We hypothesize that using conformal inference [6] to turn models into set predictors helps to automate data cleaning problems.

ML-based Data Imputation

Consider a dataset represented as a table or matrix $X_{n\times d}$. We train an imputation model $\hat{f}_c$ for each column $c \in \{1, ..., d\}$ that predicts the value in cell i, c given the values in row i except for the value in column c, i.e.:

$$X_{i,c} = \hat{f}_c(X_{\{1,...,d\}\setminus\{c\}})$$

Conformal Predictors

Conformal predictors are uncertainty quantification methods that allow the calculation of statistically rigorous confidence intervals (regression) or sets (classification) from any point estimator for a user-defined error rate [6].

1. sample D_{train} and D_{calib} i.i.d for the tabular dataset $\mathcal{D} := \mathcal{X} \times \mathcal{Y}$
2. fit a (arbitrary) predictor $\hat{f}$ to the training data D_{train}
3. compute nonconformity sores R_{calib} using nonconformity score function S:
$$\hat{y}_{calib} = \hat{f}(X_{calib})$$
$$R_{calib} = S(\hat{y}_{calib}, y_{calib})$$
4. compute the k-th empirical quantile of R_{calib}, where $\alpha \in [0, 1]$ is the user-chosen error rate:
$$k = \frac{\lceil (n+1)(1-\alpha) \rceil}{n}$$
$$\hat{q} = quantile(R_{calib}, k),$$

For new and unseen test data X_{test} construct the prediction set $\mathcal{C}$, which depends on the chosen nonconformity score function S. The conformal framework guarantees that $\mathcal{C}(X_{test})$ contains y_{test} (the true label) with at least probability $1-\alpha$:

$$\mathbb{P}(y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha$$

If the model $\hat{f}$ fits the data D_{train} well, the prediction sets $\mathcal{C}$ will be small. However, if $\hat{f}$ performs poorly, the prediction sets will be larger to satisfy this Statement, which is known as *(marginal) coverage*. However, in this work, we use conditional conformal prediction, for more information see the paper Appendix B.

Conformal Data Cleaning (CDC)

CDC uses conformal prediction to calibrate the ML models of the above described imputation approach and turn them into set predictors.

1. **Error Detection:** For new and unseen test data $D^{test}_{n\times d}$ and error rate, e.g., $\alpha = 0.01$, $\widehat{cleaner}$ predicts confidence sets $\mathcal{C}_{i,c}$, where $\forall i \in \{1, ..., n\}$ and $\forall c \in \{1, ..., d\}$. If $D^{test}_{i,c} \notin \mathcal{C}_{i,c}$, we assume $D^{test}_{i,c}$ as incorrect and compute a boolean matrix $B^{test}_{n\times d} \subset \{0, 1\}$, which represents incorrect values of D^{test} as 1.
2. **Error Cleaning:** Knowing which cells are erroneous, i.e., B^{test}, allows to remove those and treat the situation as a missing value problem. Therefore, we once more leverage $\widehat{cleaner}$'s underlying ML models to impute them.

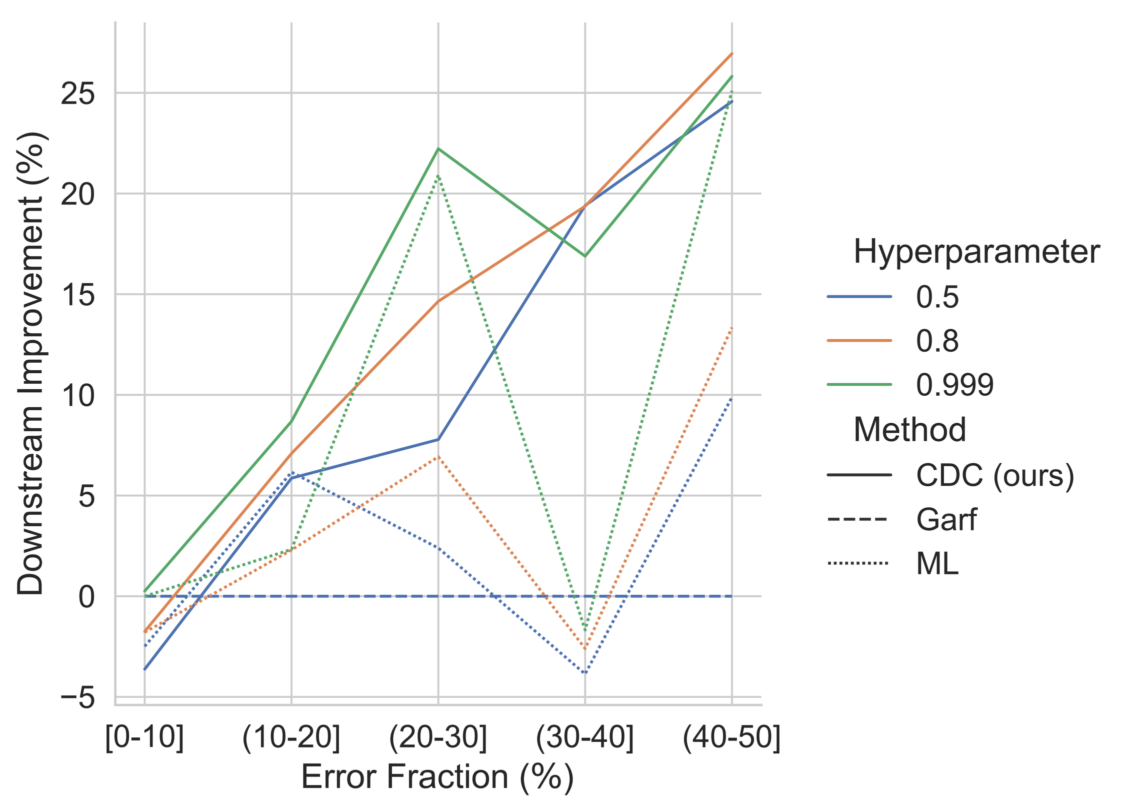
Experiments

- Cleaning methods are trained on high-quality training data without errors
- **Datasets:**
  - 16 heterogeneous tabular datasets [2] from OpenML
  - $80/20$ split into training and test dataset
  - We use Jenga [5] to corrupt the test datasets
    - four error types: swapping values between columns, random scaling, Gaussian noise, and shifts of categorical value distributions
    - five error fractions: 1%, 5%, 10%, 30%, and 50%
  - 320 corrupted datasets with 0% to about 41% with $11\% \pm 14$ errors on average
- **Baselines:**
  - Not calibrated ML models that are otherwise applied in the same way
  - Garf [4] uses a SeqGAN to learn functional dependencies between columns and generate data repair rules applied for data cleaning
- **Evaluation:**
  - True positive rate and false positive rate of error detection
  - Downstream performance improvement relative to the corrupted performance

Results: Error Detection

- **True Positive Rate (↑)**
  - lower hyperparameter values lead to higher (better) TPR
  - CDC is more robust against error fraction
- **False Positive Rate (↓)**
  - lower hyperparameter values lead to lower (worse) FPR
  - hyperparameter's values have more influence on the FPR (ML more than CDC)
  - increasing hyperparameter reduces difference between ML and CDC
  - CDC has in $\sim 80\%$ of the experiments fewer false alarms

Results: Downstream Improvement



- CDC is more robust against error fraction
- in $\sim 61\%$ of the experiments, CDC leads to better downstream improvements
- in $\sim 66\%$ of the experiments, higher confidence level leads to better downstream performance

Results: Relative Confidence Set Size

- sort by relative confidence set sizes (easy, moderate, difficult)
- difficult experiments mostly degrade downstream performance
- easy experiments mostly improve downstream performance
- opens possibilities for data monitoring

Conclusion

- CDC can detect and clean erroneous values of heterogeneous tabular data without user interventions
- CDC outperforms the baselines in about $\sim 61\%$ of our experiments
- CDC using high confidence level improves downstream performance in $\sim 60\%$ of the cases
- Results highlight potential of automated imputation combined with modern calibration methods to tackle data quality problems

Future Work

- Iterative cleaning similarly to multiple imputation could further increase CDC's performance
- Apply CDC as data quality monitoring and data cleaning approach

Limitations

- *Tabular datasets* as defined by Grinsztajn et al. [2]
  - five to 15 columns (mixed types)
  - $4,800$ to $89,000$ rows (no missing values)
  - regression, binary classification, and multi-class classification
- *High-quality training data* without any errors

Acknowledgments

We thank the anonymous reviewers for their helpful and constructive feedback and Philipp Jung for valuable discussions. This research was supported by the German Federal Ministry of the Environment grant number 67KI2022A and the German Federal Ministry of Education and Research grant number 16SV8856.

Further Information

References

[1] Felix Biessmann, Jacek Golebiowski, Tammo Rukat, Dustin Lange, and Philipp Schmidt. Automated data validation in machine learning systems. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2021.

[2] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS 2022 Datasets and Benchmarks Track*, New Orleans, United States, November 2022.

[3] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. *Frontiers in Big Data*, 4, 2021.

[4] Jinfeng Peng, Derong Shen, Nan Tang, Tieying Liu, Yue Kou, Tiezheng Nie, Hang Cui, and Ge Yu. Self-Supervised and Interpretable Data Cleaning with Sequence Generative Adversarial Networks. *Proceedings of the VLDB Endowment*, 16(3):433–446, November 2022.

[5] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. JENGA - A Framework to Study the Impact of Data Errors on the Predictions of Machine Learning Models. In *Proceedings of the 24th International Conference on Extending Database Technology, {EDBT} 2021, Nicosia, Cyprus, March 23 - 26, 2021*, pages 529–534. OpenProceedings.org, 2021.

[6] Vladimir Vovk, A. Gammerman, and Glenn Shafer. *Algorithmic learning in a random world.* Springer, New York, 2005.