Towards Unsupervised Data Quality Validation on Dynamic Data

Sergey Redyuk
Technische Universität Berlin
sergey.redyuk@tu-berlin.de

Volker Markl
Technische Universität Berlin
volker.markl@tu-berlin.de

Sebastian Schelter
New York University
sebastian.schelter@nyu.edu

Validating the quality of data is crucial for establishing the trustworthiness of data pipelines. State-of-the-art solutions for data validation and error detection require explicit *domain expertise* (e.g., in the form of rules or patterns) [1] or *manually labeled examples* [7]. In real-world applications, domain knowledge is often incomplete, data changes over time, which limits the applicability of existing solutions. We propose an unsupervised approach for detecting data quality degradation early and automatically. We will present the approach, its key assumptions, and preliminary results on public data to demonstrate how data quality can be monitored without manually curated rules and constraints.

Exemplary use case. Consider a data engineering team at a retail company, which has to regularly ingest product data from heterogeneous sources such as web crawls, databases, or key-value stores, with the goal of indexing the products for a search engine. Errors in the data, such as missing values or typos in the category descriptions, lead to various problems: attributes with missing values might not be indexed, or the products might end up in the wrong category. Ultimately, customers may not be able to find the products via the search engine. Tackling such data quality issues is tedious, as manual solutions require in-depth domain knowledge and result in complex engineering efforts.

Proposed approach. We focus on scenarios where systems regularly ingest potentially erroneous external data. We apply a machine learning-based approach which automatically learns to identify "acceptable" data batches, and raises alerts for data batches that vary significantly from previous observations. We analyze structured data that arrives periodically in batches (e.g., via a nightly ingestion of log files). At time t, we assume that previously ingested data (timestamps 1 to $t-1$) is of "acceptable" quality if it did not result in system crashes or require manual repairing. We use these previously ingested data batches as examples with the goal to identify future erroneous batches. Note that we do not look for erroneous records, but aim to identify errors that corrupt an entire batch, such as the accidental introduction of a large number of missing values in a column.

Figure 1 illustrates our approach: We compute a set of statistics for every column of an observed data batch: completeness, approximate number of distinct values, as well as mean, standard deviation, minimum and maximum values for numeric columns. We record these statistics as time series over multiple batches ❶. We apply time series forecasting methods [4] to estimate the expected data statistics for the next batch (the green area in ❷). When a new batch of data becomes available, we compute its actual statistics ❸ and compare them to the estimate ❹. If the observed statistic differs significantly from the estimated value, we raise an alert about a potential degradation of the data quality.

Preliminary Results. We conducted a preliminary evaluation on datasets of flight information [6] and crawled Facebook posts, for which we have chronological information as well as erroneous and manually cleaned variants. We show a series of "acceptable" data batches from the past to our approach, and have it decide whether the next data batch is "acceptable" or erroneous (we randomly choose either of those for evaluation). We repeat this for multiple timespans and compute binary classification metrics such as accuracy and F1-score. We find that the popular time series forecasting method exponential smoothing combined with a simple decision strategy for outlier detection (inclusion in a 90% confidence interval) works well in many cases, and provides F1-scores of up to 96% for the *Flights* dataset. In contrast, existing baseline solutions such as TFX Data Validation [1] or statistical tests [3] perform with F1-scores of only 64% and 62% respectively.

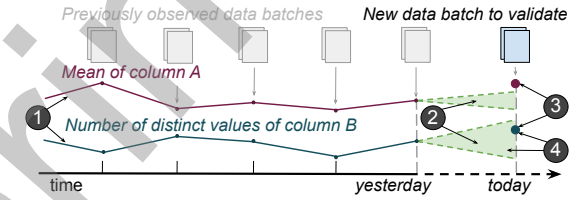

Figure 1: Overview of the proposed approach for data quality monitoring: we maintain time series of column statistics from previously observed data batches of "acceptable" quality. To decide whether a new batch should be accepted, we compare its statistics to a forecast-based estimate of the expected statistics based on the observed time series.

Next directions. We intend to conduct an extensive evaluation on additional datasets against several baselines [2, 5, 8] with respect to the prediction performance, execution time, and scalability. Furthermore, we will investigate the benefits of applying multivariate forecasting methods for our use case.

Acknowledgements. This work was funded by the HEIBRiDS graduate school, with the support of the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data, BBDC 2 (01IS18025A), BZML (01IS18037A), and the Software Campus Program (01IS17052).

REFERENCES

[1] Dennis Baylor et al. 2017. Tfx: A tensorflow-based production-scale machine learning platform. *KDD*, 1387–1395.
[2] Eric Breck, Marty Zinkevich, Neoklis Polyzotis, Steven Whang, and Sudip Roy. 2019. Data Validation for Machine Learning. *SysML*.
[3] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.
[4] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. 2015. *Introduction to time series analysis and forecasting*. John Wiley & Sons.
[5] Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, and Felix Naumann. 2015. Data Profiling with Metanome. *PVLDB* 8, 12 (2015), 1860–1863.
[6] Erhard Rahm and Hong Hai Do. 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* 23, 4 (2000), 3–13.
[7] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. 2017. Holoclean: Holistic data repairs with probabilistic inference. *PVLDB* 10, 11 (2017), 1190–1201.
[8] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating Large-scale Data Quality Verification. *PVLDB* 11, 12 (Aug. 2018), 1781–1794.