Tabular Data Insights and Synthesis with the AutoTable Approach

Josef Spillner
Zurich University of Applied Sciences
Winterthur, Switzerland
josef.spillner@zhaw.ch

***Abstract*—AI convergence platforms such as Google's Unified AI Platform promise to fully interpret and understand any data submitted to them. The business needs of SMEs are however better addressed by tailored tools that smartly parse and interpret data without being locked into a particular vendor platform. With AutoTable, a new tool design for schema, pattern and relation inference as well as training data synthesis has recently become available. This paper explains why AutoTable is smart, unintrusive and yet powerful in working with tabular business data such as CSVs, flat JSON and spreadsheets.**

***Index Terms*—Data intelligence, beyond-schema inference, pattern recognition, MLOps**

I. Motivation

Cloud-based data insights platforms are permitting businesses to perform advanced data science and machine learning tasks easily. System designs range from academic research and training platforms such as XCloud [1] and LiveDataLab [2] to commercially operated platforms such as the Google Cloud's Unified AI Platform. The more advanced platforms load data of any type and automatically add schema information as well as further insights on a statistical level, sometimes even on a level of semantic relations between data records, to enable rapid exploration and postprocessing. While undoubtedly capable of accelerating data intelligence, these platforms also come with negative traits including vendor lock-in, impossibility to extend or customise, and fine-grained but hard to predict cost models. This makes them rather unsuitable for SMEs that look for unintrusive solutions that can be self-hosted or managed by an existing hosting partner, and easily extended in case of limitations.

A particular need exists to gain initial understanding about the nature of data in a given tabular dataset. Tabular data, including timeseries and relational data, has recently attracted the interest of researchers concerned with explainable AI (XAI) who noticed a gap between the focus on tabular engineering, and many XAI techniques not applicable to such formats despite dominance in businesses, especially SMEs [3]. This argument could be extended further: Even basic data preprocessing and delivery of first insights is still tedious with tabular data in conjunction with today's tooling. Data engineering libraries such as Pandas and Spark provide only minimal schema inference capabilities on a columnar level, and NoSQL databases such as HBase require even manual inference [4]. Tools like SDV are highly capable, but also dependency-heavy and focused on model training rather than fast detailed schema inference.

Hence, a flexibly deployable solution is needed to gain first insights into data. AutoTable has been designed to address this need. It is an unintrusive tool that fits into the early stages of data processing pipelines and enables data exploration, but also synthetic training and test data generation, by examining schematic, structural and relational properties of small calibration datasets. In this paper, AutoTable is introduced and evaluated with tabular traffic data from Zurich state streets.

II. Solution Approach

AutoTable is designed as generic roundtripping tool: It learns structural properties from a small calibration dataset, and is then capable of producing training or test data in arbitrary quantities with the same properties. As such, it can be used in MLOps processes to cause synthetic load on the data analysis pipeline and thus to anticipate larger data volumes or velocities in operation. The learning effort can be adjusted to the trade-off between resource consumption (CPU/memory) and desired level of detail of characteristics. AutoTable supports the following smart features:

1) Flexible data source attachment. AutoTable can load tabular data from several sources, making it a versatile tool at the beginning of data analysis pipelines.
2) Origin format memorisation. AutoTable remembers details of the source formatting such as CSV separators, for the sake of producing compatible output on demand.
3) Column types. Schema inference is conducted with support for rich types, including accurate representation of complex datatypes such as timestamps with N subsecond precision digits.
4) Row relationships. Data patterns across rows are determined. The patterns include (strict) monotonously increasing or decreasing numerals, as well as categorical and numerical value distributions. For instance, the characteristics of a column of type $float$ with values ranging from -1.0 to 1.0 are fully captured.
5) Column relationships. Refers to clustered $GROUPBY$ relationships and corresponding value distributions across columns. For instance, all rows with column

$A = ”X”$ have $B > 0$ whereas rows with $A = ”Y”$ have $B < 0$.

III. Use Case and Evaluation

A *Turing test for data* works as follows: Any data produced by AutoTable is indistinguishable from the calibration data even by domain experts. The test shall be conducted with a small (43 lines) calibration data set containing vehicle measurements from Zurich state streets. Column headers have been translated and simplified for presentation purposes. The tabular data informs about date and time of each measurement, the reference point (street lane), as well as measured and estimated characteristics such as weight, size and class of a vehicle, in addition to its speed.

The sample contains measurements that are sometimes less than a second apart, and sometimes more than eight seconds, hence representing a unique pattern across rows in the $Datetime$ column. Moreover, it contains such patterns also across categorical variables, such as the class of vehicle ($Class$) that is spread across the majority passenger car and the minority lorry. Column relationships follow from the vehicle class – for instance, lorries are longer and heavier ($Length$, $GrossW$, $NetW$), whereas they do not possess unique speed characteristics (v). Furthermore, the net weight column data is always around 0.3 to 0.4 below the corresponding gross weight data, representing another detectable cross-column relationship. A simplified excerpt of the data with translated column names is shown below.

```
Datetime;Det;No;Class;Length;Height;v;GrossW;NetW
27.11.2020 17:36:17.1;M2;640;Car;474;64.3;P;2.1;1.7
27.11.2020 17:36:15.1;M2;637;Lor;702;66.7;P;7.9;7.6
```

AutoTable has no prior knowledge of the data and instead learns all properties through a single parsing pass. An excerpt of the synthetically generated data at best level of knowledge is shown below. It follows the calibration sample logically and chronologically.

```
Datetime;Det;No;Class;Length;Height;v;GrossW;NetW
27.11.2020 17:36:57.8;M1;676;Car;387.4;70.3;24.8;24.4
27.11.2020 17:37:00.8;M2;681;Car;448.5;74.0;8.0;7.6
27.11.2020 17:37:03.3;M2;681;Car;386.0;69.1;31.5;31.2
27.11.2020 17:37:06.2;M1;685;Car;473.1;72.9;19.2;18.8
```

Fig. 1 shows how the level of knowledge about data characteristics, including schema, pattern and relationships, influences the quality of synthetic data generation and roundtripping. Ten million rows of data are produced, while leaving the characteristic distribution intact when detailed distribution information is collected and exploited in the generation process.

AutoTable can be requested to produce enhanced data insights files in JSON format that describe key aspects of the inferred schemas, patterns and relations. This is particularly useful for tracking data changes over time. To build the bridge towards XAI over tabular data, AutoTable is also able to generate human-readable descriptions of column contents, as shown in the following excerpt that includes modestly compute-intensive analysis such as cross-clustering of columns.

```
Column 'Class': - of type string
- categorical, with 2 values; all shown:
  * Car: 95.2%
  * Lor:  4.8%
- contents influence columns:
  * Length: clustered
  * GrossW: not clustered
```

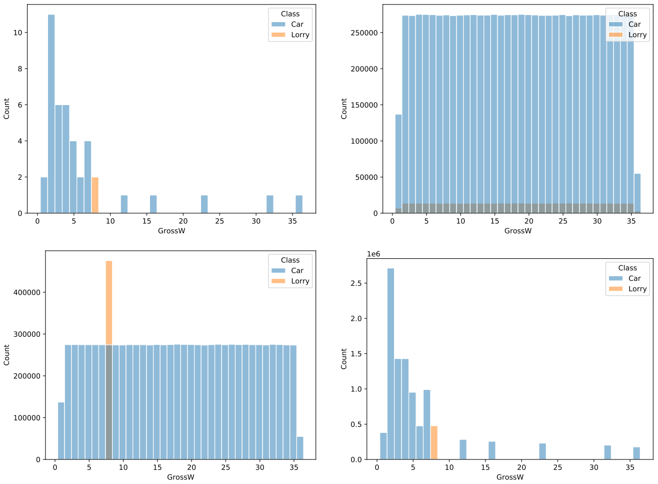

Fig. 1. Comparison of histograms of calibration dataset (upper left), synthetically generated dataset without constraints (upper right), with constraints but without distribution information (lower left), and with distribution information (lower right; synthesis with full level of knowledge)

IV. Conclusion and Future Work

AutoTable is able to generate deep insights about the schemas, row patterns and column relations in tabular data. Furthermore, AutoTable is able to represent this knowledge in machine-readable and human-readable ways, and to exploit the knowledge for synthetic data generation with the same characteristics. A functional prototype of AutoTable is available as open source software and knowledge JSON schema[1].

Future iterations of AutoTable shall be able to interpret and explain schema evolution based on the produced JSON insights files. For instance, a renamed column (e.g. $Name$ to $FamilyName$) will become detectable by comparing its data characterics even without explicit information about the renaming. This corresponds to challenges in enterprise application integration where unannounced schema changes take place occasionally.

References

[1] Lu Xu and Yating Wang. XCloud: Design and Implementation of AI Cloud Platform with RESTful API Service, 2019.

[2] Aaron Green and ChengXiang Zhai. LiveDataLab: A Cloud-Based Platform to Facilitate Hands-on Data Science Education at Scale. In *Proceedings of the Sixth ACM Conference on Learning @ Scale*, L@S ’19, New York, NY, USA, 2019. Association for Computing Machinery.

[3] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access*, 9:135392–135422, 2021.

[4] Angelo Frozza, Eduardo Defreyn, and Ronaldo Mello. A Process for Inference of Columnar NoSQL Database Schemas. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 175–180, Porto Alegre, RS, Brasil, 2020. SBC.

[1] AutoTable link: https://github.com/serviceprototypinglab/autotable