BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

A Robust Approach for Discovering Functional Dependencies using Machine Learning Approaches

von

Philipp Jung

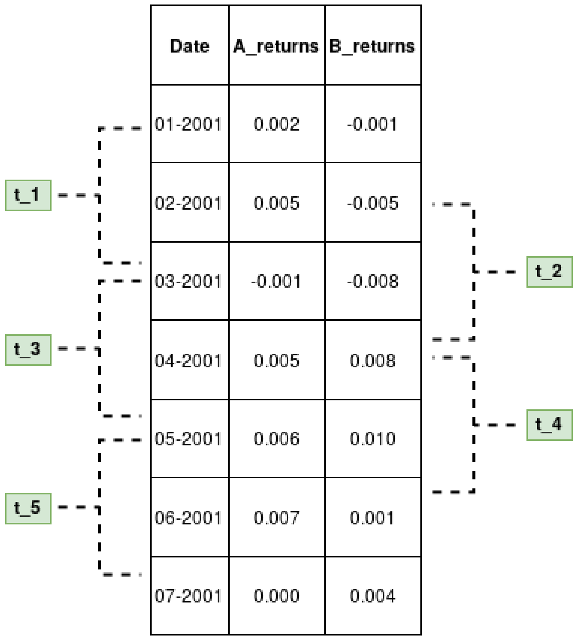
Philipp Jung
Matrikelnummer: 872855
16.03.2019

Gutachter:
Prof. Felix Biessmann
Dr. Zweit Gutachterin

ABSTRACT. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Contents

1 Introduction



Date	A_returns	B_returns
01-2001	0.002	-0.001
02-2001	0.005	-0.005
03-2001	-0.001	-0.008
04-2001	0.005	0.008
05-2001	0.006	0.010
06-2001	0.007	0.001
07-2001	0.000	0.004

Figure 1: Illustration of the rolling-window approach for a time-series containing seven time-steps filled with mock-data. Five subsets of length 3 divide the time-series.

This approach is schematically described in figure 1.

2 Theory

Functional dependencies(FDs) are a way of expressing "a priori knowledge of restrictions or constraints on permissible sets of data"[Mai83, p.42] in relational database theory. In order to give a definition of FDs, some concepts stemming from relational database theory need to be introduced beforehand.

2.1 Relational Database Theory

A *relation scheme*[1] $\boldsymbol{R}$ is a finite set of *attribute names* $\{A_1, A_2, \ldots, A_n\}$, where to each attribute name A_i corresponds a set D_i, called *domain* of A_i, $1 \leq i \leq n$.

Let $\boldsymbol{D} = D_1 \cup D_2 \cup \cdots \cup D_n$, then a *relation* r on relation scheme $\boldsymbol{R}$ is a finite set of mappings $\{t_1, t_2, \ldots, t_p\}$ from $\boldsymbol{R}$ to $\boldsymbol{D}$:

$$t_i : \boldsymbol{R} \to \boldsymbol{D}, \tag{1}$$

where we call those mappings *tuples* under the constraint that [Mai83, p.2]

$$t(A_i) \subseteq D_i. \tag{2}$$

2.2 Definition of a Functional Dependency

Consider a relation r on scheme $\boldsymbol{R}$ with subset $X \subseteq \boldsymbol{R}$ and a single attribute $A_i \in \boldsymbol{R}$. A FD $X \to A$ is said to be *valid* in r, if and only if

$$t_i[X] = t_j[X] \Rightarrow t_i[A] = t_j[A] \tag{3}$$

holds for all all pairs of distinct tuples $t_i, t_j \in r$.[Abe+19, p. 21] We say that X *functionally determines* A[Mai83, p. 43] and name X the *left side*, whilst calling A the *right side*.

2.3 Approximate Functional Dependencies

In the field of data profiling an extensive body of theory and algorithms for FD detection has been created in the past decades.[Pap+15] These mainly consider FDs as defined in formula 3. Howevever, the strict detection of FDs yields results that are solely applicable in a strictly controlled environment. Real-world datasets faced by data-scientists or database engineers are often *noisy*. Entries might be corrupted by missing data, wrongly entered entries or incomplete datasets. Inconsistencies are to be expected. Thus, functionally dependent column-combinations might not be detected as such. This may result in misleading insights when searching for FDs.

To illustrate this, table 1 shows an example of noisy data. The potential FD **Town → ZIP** is not captured by the definition given in equation 3. Due to a type-error, the potential FD is invalidated. To still capture meta-information, a different dependency-measure than given in equation 3 is needed.

Approximate FDs (AFDs), sometimes called *Relaxed FDs*, improve the applicability of FDs, "in that they relax one or more constraints of the canonical FDs"[CDP16, p. 1].

[1] also called *relational schema* in literature[Abe+19, p.21]

While there are AFDs introducing general error measures, others are defined "aiming to solve specific problems"[CDP16, p. 1].

	Data			
ID	First name	Last name	Town	ZIP
1	Alice	Smith	Munich	19139
2	Peter	Meyer	Muinch	19139
3	Hannah	Parker	Munich	19139
4	John	Pick	Berlin	12055

Table 1: Even though the ZIP-Code functionally determines the town (and vice-versa) in the given example, a FD is not capable of displaying this fact. A type-error in the dataset with ID 2 invalidates the functional dependency.

The error measure for

3 Execution

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

3.1 Begriffsdiskussion

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

4 Discussion

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

4.1 Begriffsdiskussion

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

References

[Abe+19] Ziawasch Abedjan et al. *Data Profiling.* 2019. ISBN: 9781681734477. DOI: `https://doi.org/10.2200/S00878ED1V01Y201810DTM052`.

[CDP16] Loredana Caruccio, Vincenzo Deufemia, and Giuseppe Polese. "Relaxed Functional Dependencies - A Survey of Aproaches". In: *IEEE Transactions on Knowledge and Data Engineering* 28.1 (2016), pp. 147–165. ISSN: 2150-8097.

[Mai83] David Maier. *The Theory of Relational Databases.* Computer Science Pr, 1983. ISBN: 0914894420. URL: `http://web.cecs.pdx.edu/~maier/TheoryBook/TRD.html`.

[Pap+15] Thorsten Papenbrock et al. "Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms". In: *Proc. VLDB Endow.* 8.10 (June 2015), pp. 1082–1093. ISSN: 2150-8097. DOI: `10.14778/2794367.2794377`. URL: `https://doi.org/10.14778/2794367.2794377`.