
Data Cleaning using Probabilistic Models of Integrity Constraints

Simão Eduardo
School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, UK
s.eduardo@ed.ac.uk

Charles Sutton
The Alan Turing Institute &
School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, UK
csutton@inf.ed.ac.uk

Abstract

In data cleaning, data quality rules provide a valuable tool for enforcing the correct application of semantics on a dataset. Traditional rule discovery techniques assume a reasonably clean dataset, and fail when faced with a dirty one. Enforcement of these rules for error detection is much less effective when mined on dirty data.

In the databases literature, a popular and expressive type of logic-based data quality rule (or Integrity Constraint) is the *constant Conditional Functional Dependency* (cCFD) [Fan et al., 2011], which can be easily understood by a data analyst.

We introduce a probabilistic model that combines error detection and rule induction (cCFDs), we show that this methodology performs better than just traditional logic-based error detection. Moreover, after inference is performed, we provide a set of rules which is statistically sound and with low redundancy. To the best of our knowledge this is the first work to combine statistical anomaly detection with logic-based approaches to data cleaning.

1 Introduction

In industry, most of the data usually comes in tabular (e.g. CSV files, Excel sheets), or in relational form (e.g. relational database systems).

Error detection is an important step in data cleaning, which can be carried out using data quality rules if enforced on the data. These rules should be kept up-to-date to ensure the data is clean, and thus predictions using it are sound.

There is a need to infer, apply and monitor data quality rules, particularly for tabular datasets. Often these rules are either inferred automatically or by a data analyst. Indeed, these rules are often part of the schema (i.e. blueprint) of relational databases. Nowadays, it is unrealistic to expect a human to perform rule discovery without any tools, given the intricacy of data.

We tackled this problem from a probabilistic point-of-view, trying to provide an algorithm for error detection, and robust rule induction for cCFDs - by removing redundant and spurious rules from a candidate set, given by ZART [Szathmary et al., 2007]. Traditional approaches for CFD and cCFD induction are defined in [Fan et al., 2011].

Our probabilistic model was implemented using *Structural Expectation Maximization* (SEM) in [Friedman, 1998]. We attempt robust rule induction, and show that traditional discovery techniques do not perform as well.

We obtained good results for error detection with our model: both with the set of rules induced, and directly from the model itself. The final set of induced rules was reduced, and thus less redundant. We obtain better results than traditional techniques under significant noise.

2 Related work

In the data cleaning pipeline, one of the first steps towards cleaning the dataset is to detect errors. Often, error detection can be reduced to the problem of anomaly detection, particularly in tabular datasets. In tabular datasets, quantitative or logic-based methods can be used to detect anomalies. The quantitative approach is statistically inspired, meanwhile the logic-based can use integrity constraints or data quality rules, as well as user-defined data transformations.

Formally, error detection using integrity constraints (ICs) usually involves detecting the tuples (rows) that violate a set of constraints seen as describing the dataset. Recently, two good surveys have been published [Ilyas and Chu, 2015] and [Fan, 2015] on data cleaning, mostly focusing on logic-based approaches.

On the other hand, in quantitative error detection, there has been considerable work in outlier detection for quantitative data, as seen in survey [Hellerstein, 2008]. Most of this work is based on robust estimators, and methods for univariate and multivariate outlier detection, but it also contains observations on relational data. A tutorial on outlier detection can be found in [Kriegel et al., 2009]. Methods have also been developed for distributional change detection in [Dasu et al., 2009].

3 Results

For our experiments we used the Adult dataset (UCI Machine Learning Repository), with both categorical and continuous features, injected with random errors (outliers and typos).

We compared our model (Prob-Log) to two other methods in error detection and rule induction (cCFDs) performance. The first traditional method, which assumes a moderately clean dataset is CFDMiner [Fan et al., 2011], mines for rules with confidence 1, in the *Association Rule* sense. The second is ZART [Szathmary et al., 2007] which is a non-redundant *Association Rule* mining algorithm that was modified to obtain cCFDs, its definition allows for us to mine rules with confidence less than 1, thus more robust to noise. Intuitively, error detection with CFDMiner and ZART can be achieved by searching for the tuples in the dataset that violate their induced set of cCFD rules.

We present results for both number of rules induced (Table 1), and F-Measure for the error detection process (Figure 1). The noise is injected at random from 0 % to 20 % of the cells in the dataset. In Table 1 *low_conf* are rules mined with ZART which exhibit poor confidence in the dataset, in the *Association Rule* sense, whilst *high_conf* are high confidence rules (near 1).

Results in Figure 1 show that our model (Prob-Log, in purple) obtained good error detection performance, against rule-based detection using the set of cCFDs mined by ZART (Candidate Set fed into Prob-Log, in blue), CFDMiner (Ground-Truth, in red), and the rule set \mathcal{S} induced by our model (in green). Note that CFDMiner had access to the clean dataset to induce its cCFDs, thus we name it Ground-Truth.

Finally, results in Table 1 suggest that Prob-Log offers substantial reduction in number of cCFDs (less redundancy) without much loss in error detection performance. Particularly when the rules are more spurious (less confidence), tagged *low_conf* (Table 1).

Corruption Level	Candidate Type	ZART (Candidate Set)	Prob-Log (Set \mathcal{S})	CFDMiner
0.1 %	high_conf	58	43	1352
1 %	high_conf	46	38	538
1 %	low_conf	265	115	538
3 %	high_conf	58	48	19
5 %	high_conf	69	59	0
5 %	low_conf	248	133	0
7 %	high_conf	71	58	0
10 %	high_conf	70	54	0
10 %	low_conf	265	156	0
15 %	high_conf	66	48	0
15 %	low_conf	270	169	0
20 %	high_conf	128	86	0

Table 1: Number of Rules generated per method, for each injected noise level in Adult dataset - from 0.1% to 20 % erroneous cells, corrupted at random. Ground-Truth cCFD rules using CFDMiner registers 611 rules on the clean dataset.

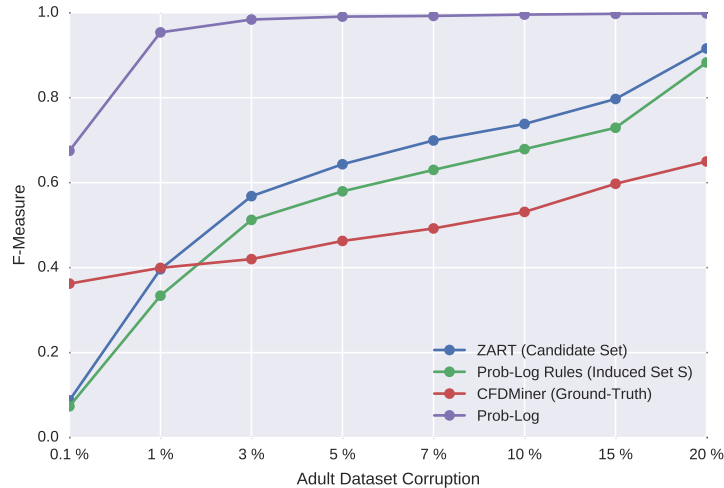


Figure 1: F-Measure of error detection per method, for each injected noise level in Adult dataset - from 0.1% to 20 % erroneous cells, corrupted at random.

Acknowledgments

The authors would like to thank Wenfei Fan, Chris Williams, Floris Geerts and Joeri Rammelaere for their help in producing this work. This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

- Wenfei Fan, Floris Geerts, Jianzhong Li, and Ming Xiong. Discovering conditional functional dependencies. *IEEE Trans. on Knowl. and Data Eng.*, 23(5):683–698, May 2011. ISSN 1041-4347. doi: 10.1109/TKDE.2010.154. URL <http://dx.doi.org/10.1109/TKDE.2010.154>.
- L. Szathmary, A. Napoli, and S. O. Kuznetsov. ZART: A Multifunctional Itemset Mining Algorithm. In *Proc. of the 5th Intl. Conf. on Concept Lattices and Their Applications (CLA '07)*, pages 26–37, Montpellier, France, Oct 2007. URL <http://hal.inria.fr/inria-00189423/en/>.
- Nir Friedman. The bayesian structural em algorithm. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98*, pages 129–138, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-555-X. URL <http://dl.acm.org/citation.cfm?id=2074094.2074110>.
- Ihab F. Ilyas and Xu Chu. Trends in cleaning relational data: Consistency and deduplication. *Foundations and Trends® in Databases*, 5(4):281–393, 2015. ISSN 1931-7883. doi: 10.1561/19000000045. URL <http://dx.doi.org/10.1561/19000000045>.
- Wenfei Fan. Data quality: From theory to practice. *SIGMOD Record*, 44(3):7–18, 2015. doi: 10.1145/2854006.2854008. URL <http://doi.acm.org/10.1145/2854006.2854008>.
- Joseph M Hellerstein. Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE)*, 2008.
- Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Outlier detection techniques. *Tutorial in PAKDD 2009*, 2009.
- T. Dasu, S. Krishnan, D. Lin, S. Venkatasubramanian, and K. Yi. Change (Detection) You Can Believe in: Finding Distributional Shifts in Data Streams. *Lecture Notes in Computer Science*, 5772:21, 2009. doi: 10.1007/978-3-642-03915-7_3.