Causal Inference on Multivariate and Mixed-Type Data

Alexander Marx and Jilles Vreeken

Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany
{amarx,jilles}@mpi-inf.mpg.de

A Appendix

In the appendix we give additional results and describe the multivariate cause-effect pairs and their corresponding datasets in more detail.

A.1 Encoding of the internal nodes

A.2 Synthetic data

Here we show the additional plot about how both scores deal with dimensionality, i.e. when the number of dimensions $k = l$ increases. In Figure 1 we show the results for $k, l \in \{2, 3, 5, 7, 11\}$ on nominal, numeric and mixed-type data for Crack$_\Delta$ and Crack$_\delta$. Crack$_\delta$ performs better on mixed-type data and is equally good on the single-type data sets. In general, both approaches work well in high dimensions.

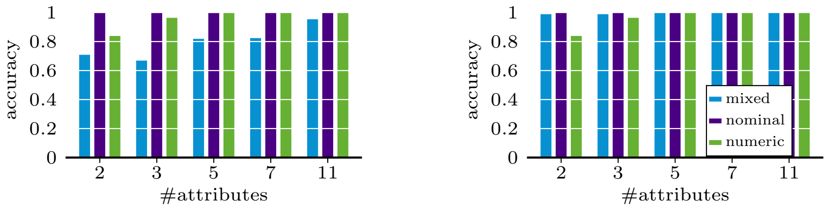

Fig. 1. Accuracy of Δ (left) and *NCI* (right) on symmetric dimensions $k \in \{2, 3, 5, 7, 11\}$ for nominal, numeric and mixed-type data.

Data sets *Haberman* is a data set on medical case studies describing the survival of patients who had undergone surgery for breast cancer between 1958 and 1970 [3]. X consists of the age of the patient at time of operation, the patient's year of operation and the number of positive axillary nodes detected. Y is the survival status, which is binary and divided into longer or at most five years ($X \rightarrow Y$). The *Iris* data set contains data about three types of the Iris plant

(Y) and four features dependent on which the type can be determined [1]. Next, we extract four cause-effect data sets from the *Mammals* data set [4], which consists of both climate data and presence records of 121 mammal species over 2 183 areas of 50×50km in Europe. We assume that elevation, precipitation, average temperature and the annual temperature range (X) cause the presence of a mammal and not contrarily. We created three data sets, *Canis*, *Lepus* and *Martes*, each containing locations of different types of the named species and one data set containing all three of them. Last, we created a data set based on the octet data set [2,6]. Marx and Vreeken [5] created 10 univariate cause effect pairs based on the data set that had all the same effect that we combined to a single multivariate data set.

References

1. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Ann Eugen*, 7(2):179–188, 1936.
2. L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler. Big data of materials science: Critical role of the descriptor. *PRL*, 114, 2015.
3. S. J. Haberman. Generalized residuals for log-linear models. In *Proceedings of the 9th international biometrics conference*, pages 104–122, 1976.
4. H. Heikinheimo, M. Fortelius, J. Eronen, and H. Mannila. Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *J. Biogeogr.*, 34:1053–1064, 2007.
5. A. Marx and J. Vreeken. Telling Cause from Effect by MDL-based Local and Global Regression. In *ICDM*, pages 307–316. IEEE, 2017.
6. J. A. Van Vechten. Quantum dielectric theory of electronegativity in covalent systems. i. electronic dielectric constant. *Physical Review*, 182(3):891, 1969.