Eliciting Structure in Data

Anders Holst[1], Ahmad Al-Shishtawy[1], Juhee Bae[2], Mohamed-Rafik Bouguelia[3], Onur Dikmen[3], Göran Falkman[2], Olof Görnerup[1], Alexander Karlsson[2], Sławomir Nowaczyk[3], Sepideh Pashami[3], and Amira Soliman[1]

[1] RISE Research Institutes of Sweden
[2] University of Skövde
[3] Halmstad University

Abstract. In the BIDAF-project we have developed machine learning tools for discovering and visualizing different kinds of structure in data. This can be especially useful in the important initial stage of any machine learning or data analysis task, of exploring and understanding the data. Focus has been on unsupervised statistical machine learning methods, and the types of structure considered are: clusters, anomalies, causal relations, and higher order similarity relations.

The BIDAF-project has been a 5-year collaboration project between RISE, Halmstad University, and University of Skövde, funded by KKS. The scientific objective was to develop a framework for data analysis and machine learning suitable for massive, streaming and distributed data sets, and to provide this functionality in an intuitive and easy to use way so that it can be used also by non-specialists of machine learning.

In line with this we decided to focus on statistical machine learning methods for unsupervised exploration and visualization of structure in data. The motivation is that this is something needed in the first stage of any data analysis or machine learning project, to get an understanding of the characteristics of the data, which is then used to guide the subsequent modelling and machine learning steps. Nevertheless, intuitive and easy to use tools to explore the different types of structure in data are not very well developed, in spite of many years of research efforts in corresponding areas.

There are several types of possible structure. Fig. 1 highlights three types of structure. In BIDAF we have explored two kinds of vertical structure, i.e. relations between features: Causal relations and Higher order similarity relations. We have explored two kinds of horizontal structure, i.e. relations between samples in the data: Clusters of samples, and Anomalous samples. Time series data is very common, and sometimes separate time points can be considered as separate samples, sometimes as separate features. However, often it is more appropriate to treat it as a separate dimension of the data, with its own characteristics and complications. Structure along this dimension has in the BIDAF project been considered by aggregating features such as mean, variance and slope, calculated at one or more selected time scales.

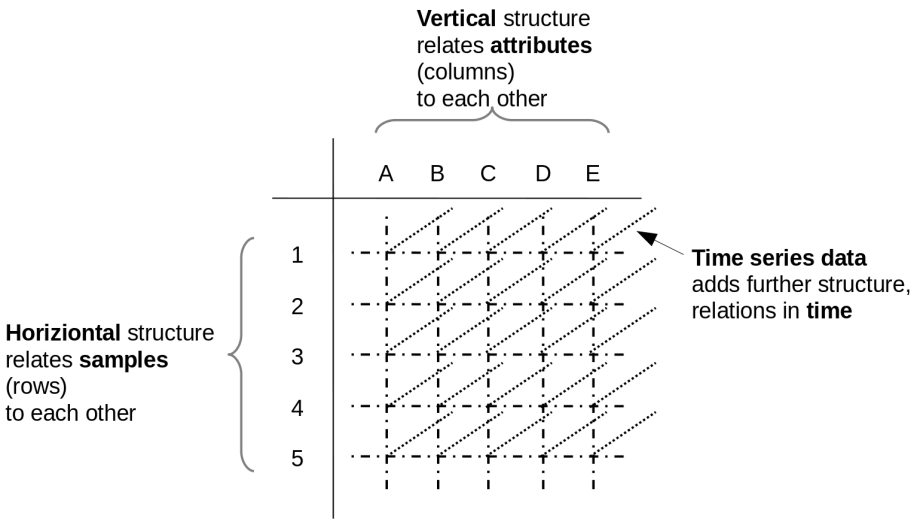

Fig. 1. Different types of structure in data.

The vision is similar in some aspects to the automatic statistician[1], in that in the initial stage of a data analysis project, an unseen dataset could be fed into the system, and it will report back with various characteristics found in that dataset. However, there are important differences: The automatic statistician is mainly a static tool, taking a batch of data and returning the final result. Our tool allows for interactive exploration of the structure in data, and also for streaming data making it possible to analyse data sets that are unlimited or too large to read in at once. Further, the automatic statistician focuses on supervised learning and time series characteristics, whereas we focus on the four types of unsupervised structure described above.

We have provided the developed tools as an open source package[2] written in python. The tool can read in a (structured) data file, and visualizes the found structures in a common view, as shown in Fig. 2. A scatter plot with the detected clusters and anomalies are shown to the lower right. Anomalous samples are colored in gray to black, whereas the other colors represent the different clusters. The corresponding time series are shown to the upper right. To the lower left is a graph showing the relations between attributes, indirect correlations are gray, direct correlations are blue, infered causal arrows are purple, and higher order similarities are green.

We will in the present some of the techniques used in the tool, focusing on clustering and causal inference, and also show a demo of the tool.

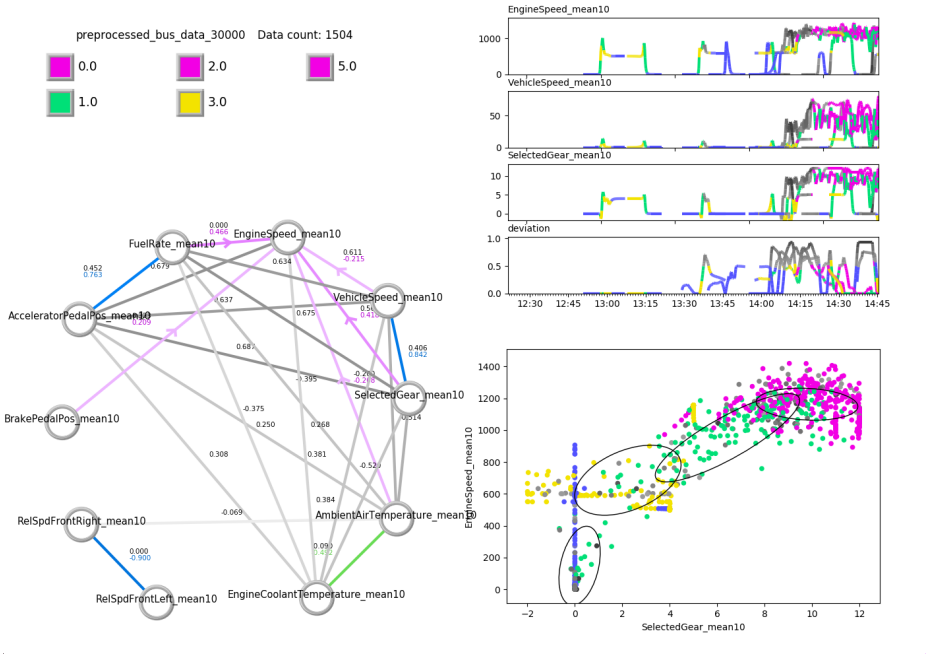

Fig. 2. Screenshot of the BIDAF tool.

References

1. Steinruecken, C., Smith, E., Janz, D., Lloyd, J., and Ghahramani, Z. (2019). The automatic statistician. In Automated Machine Learning (pp. 161-173). Springer, Cham.
2. BIDAF tool repository, https://github.com/RI-SE/BIDAF.