Exceptional Model Mining – Describing Deviations in Datasets

Wouter Duivesteijn WOUTERD@LIACS.NL
Arno Knobbe KNOBBE@LIACS.NL
LIACS, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, the Netherlands

Keywords: Local Pattern Mining, Subgroup Discovery, Exceptional Model Mining

Identifying elements that behave differently from the norm in a dataset is a task of paramount importance. Most data mining research in this direction focuses on *detecting* outliers. In Local Pattern Mining, however, we are not just looking for any deviating record or set of records in the data. Instead, we are looking for deviating *subgroups*: coherent subsets that can be *described* in terms of a few conditions on attributes of the data. The existence of such descriptions makes the resulting deviating subgroups more actionable.

'Behaving differently from the norm' can be defined in many ways. Traditionally such exceptionality is measured in terms of frequency (Frequent Itemset Mining), or in terms of a deviating distribution of one designated target attribute (Subgroup Discovery). These concepts do not encompass all forms of deviation we may be interested in. To accomodate a more general form of interestingness, we developed the Exceptional Model Mining framework (Leman et al., 2008; Duivesteijn et al., 2010; Duivesteijn et al., 2012).

The first step of the EMM framework (see Figure 1) is partitioning the attributes in two: one set to *define* subgroups on (the *descriptors*), and one set to *evaluate* the subgroups on (the *targets*). Then a *model class* is selected over the targets, and a *quality measure* over this model class is designed. Finally, the already existing Subgroup Discovery methodology is used to scan the descriptor space for subgroups that perform well according to the quality measure.

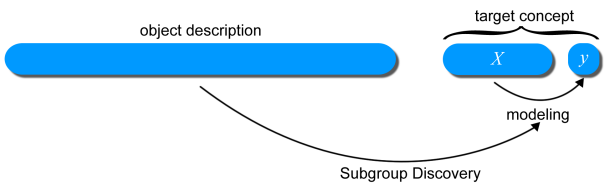

Figure 1. The Exceptional Model Mining Framework

The model class represents interplay between the targets, and the quality measure gauges the exceptionality of model parameters. For instance, we can find subgroups for which two targets are unusually correlated (Leman et al., 2008), subgroups where a classifier performs unusually (Leman et al., 2008), subgroups where a Bayesian network on several nominal targets has a deviating structure (Duivesteijn et al., 2010), and subgroups where a regression model has an exceptional parameter vector (Duivesteijn et al., 2012).

Using EMM instances, we have found subgroups concerning meteorological conditions coinciding with food chain displacement, subgroups defying the economical law of demand, subgroups showcasing the dampening effect of collective bargaining on the distribution of salaries, etcetera. Subgroup significance is tested against a Distribution of False Discoveries, and with the regression model class some subgroups can be pruned without computing the parameter vector.

Acknowledgments

This research is financially supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.065.822.

References

Duivesteijn, W., Feelders, A., & Knobbe, A. J. (2012). Different slopes for different folks – mining for exceptional regression models with cook's distance. *KDD* (pp. 868–876).

Duivesteijn, W., Knobbe, A. J., Feelders, A., & van Leeuwen, M. (2010). Subgroup discovery meets bayesian networks – an exceptional model mining approach. *ICDM* (pp. 158–167).

Leman, D., Feelders, A., & Knobbe, A. J. (2008). Exceptional model mining. *ECML/PKDD (2)* (pp. 1–16).