A Short Survey of Exceptional Model Mining — Exploring Unusual Interactions Between Multiple Targets

Wouter Duivesteijn

Fakultät Informatik, LS VIII, Technische Universität Dortmund, Germany*
wouter.duivesteijn@tu-dortmund.de

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 interpretable, hence 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* (EMM) framework [1, 2]. The key characteristic of EMM is that subgroups are sought on which multiple targets *interact* in an unusual way.

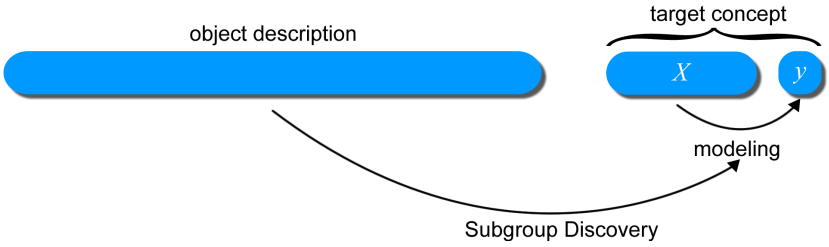

Fig. 1. The Exceptional Model Mining Framework

In the EMM framework (see Figure 1), the set of attributes is partitioned in two: one set to *define* subgroups (the *descriptors*), and one set to *evaluate* the subgroups (the *targets*). First, a *model class* is selected over the targets; this model class represents the type of interaction that we are interested in. Then, a

* This research is supported in part by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C1.

quality measure is designed; this measure quantifies how unusual a subgroup is by comparing the relevant model parameters. Finally, the already existing Subgroup Discovery methodology is used to scan the descriptor space for subgroups that perform well according to the quality measure.

The types of target interaction investigated so far are: unusual correlation between two targets [1], unusual classifier performance [1] on any number of targets, deviating structure of a Bayesian network on several nominal targets [3], and an exceptional parameter vector for an arbitrarily large regression model [4]. With these model classes, an eclectic range of subgroups has been discovered, including:

- houses in Windsor, Canada, with a driveway, a recreation room, and at least two bathrooms, unusually display no correlation between lot size and sales price [1];
- in an Affymetrix microarray dataset detailing neuroblastoma (a form of cancer) patients, a Decision Table Majority classifier performs exceptionally accurately on the people with a strongly expressed gene 'RAP1 interacting factor homolog (yeast)' [1];
- the Bayesian network of emotions that people feel when listening to music becomes sparser when that music has particularly bright sounds [3];
- in the Nordic and mountainous regions of Europe, the coexistence network of mammals is substantially different from the norm [3];
- for the relatively rich people in a poor neighborhood of Hunan, China, an increase of the price of rice leads to an increase of the demand for rice [4];
- collective bargaining agreements flatten the wage distribution with respect to years spent in school and years of work experience [4].

Despite being designed to highlight exceptionalities rather than capture the norm in a dataset, the subgroups found with the Bayesian network model class are also able to improve multi-label SVM classifier performance [5, 6].

References

1. D. Leman, A. Feelders, A.J. Knobbe, Exceptional Model Mining, Proc. ECML/PKDD (2), pp. 1–16, 2008.
2. W. Duivesteijn, Exceptional Model Mining, PhD thesis, Leiden University, 2013.
3. W. Duivesteijn, A. Knobbe, A. Feelders, M. van Leeuwen, Subgroup Discovery meets Bayesian Networks – An Exceptional Model Mining Approach, Proc. ICDM, pp. 158–167, 2010.
4. W. Duivesteijn, A. Feelders, A. Knobbe, Different Slopes for Different Folks – Mining for Exceptional Regression Models with Cook's Distance, Proc. KDD, pp. 868–876, 2012.
5. W. Duivesteijn, E. Loza Mencía, J. Fürnkranz, A. Knobbe, Multi-label LeGo – Enhancing Multi-label Classifiers with Local Patterns, Proc. IDA, pp. 114-125, 2012.
6. W. Duivesteijn, E. Loza Mencía, J. Fürnkranz, A. Knobbe, Multi-label LeGo – Enhancing Multi-label Classifiers with Local Patterns, Technical Report, Knowledge Engineering Group, Technische Universität Darmstadt, TUD-KE-2012-02, 2012.