# Exceptional Model Mining

Duivesteijn, W.

# English Summary

When given a large volume of raw data, the Computer Science subfield called *Data Mining* strives to extract information from the data; information that can be interpreted by whoever is using the data mining method at hand, and preferably used within the domain that person is interested in. With the advance of the internet in everyday life, and particularly the ubiquity of smartphones nowadays, gargantuan amounts of data are being generated by every person and company in the world. Hence the scientific community needs to develop methods that take on a seemingly uninspectable amount of data and squeeze out nuggets of information.

Identifying elements that behave differently from the norm 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 *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*: for instance, if we tell a pharmaceutical company that five given persons react badly to a certain type of medication, it is more difficult for them to act on the information than it would be if we could tell them that the group of smokers react badly to the medication.

"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 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 *Exceptional Model Mining*.

The first step of the EMM framework 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. 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, for which a classifier performs unusually, for which a Bayesian network on several nominal targets has a deviating structure, or for which a regression model has an exceptional parameter vector.

Scanning the descriptor space is computationally very intensive: we search for interesting subsets of a dataset, and one can imagine that for a large datasets, there are many candidate subsets. For the EMM instance with the regression model class, we have derived some upper bounds on the quality of a candidate subgroup, that can be computed without computing the parameter vector. Using the bounds, this last relatively expensive computation step can be omitted for up to 40% of the candidate subgroups, thus speeding up the whole process.

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. Additionally, we have developed a method to test whether such subgroups are false discoveries: solving a statistical problem roughly stating that, when we run many tests, we will eventually find something seemingly significant, purely by random effects. By generating a baseline of artificial false discoveries, and comparing the subgroups we find with the baseline, we can assess whether it is likely that our found subgroups are false discoveries too.

We have determined whether the results of one EMM instance (with the Bayesian network model) can additionally be used to improve the *prediction* of the targets when we are given new records of our dataset. Capturing the exceptional target interplay through EMM is shown capable of improving such a prediction in certain cases.