Exceptional Model Mining

Duivesteijn, W.

Chapter 10

Conclusions

We have introduced Exceptional Model Mining (EMM), a general framework to find subgroups of the data where something exceptional, something interesting is going on. These subgroups are not just any subset of the data: they must be coherent records in the dataset, covered by a succinct description in terms of conditions on attributes within the dataset. The attributes that can be used for such a description are strictly separated from the target attributes, which are used to evaluate the subgroups on. Hence, EMM can be seen as an extension of Subgroup Discovery (SD), incorporating a more complex target concept.

In traditional Subgroup Discovery the distribution of a single attribute is the target concept. In Exceptional Model Mining the target concept is a model over multiple attributes. We have discussed several model classes: correlation (Chapter 4), classification (Chapter 5), Bayesian network (Chapter 6), and linear regression models (Chapter 7). For each such model class we have developed quality measures: functions that extract relevant model characteristics, and from those characteristics compute a number quantifying how exceptional a description is. A description is considered exceptional when the model learned from the data covered by the description differs substantially, either from the model learned from the data belonging to its complement, or from the model learned from the overall dataset (for more on this choice, see Section 3.2.2). An Exceptional Model Mining run results in a succinct description of a subgroup, where for instance two targets are unusually correlated, or where a classifier performs

exceptionally good or bad, or where the conditional dependence relations between several targets deviate from the norm.

We have discussed experimental results for each of the introduced model classes. Among the most striking results are the coherent regions within Europe found on the *Mammals* data (see Section 6.2.2) with the Bayesian Network model, where animals depend on each other in a substantially different way, and the strong real-life evidence for the Giffen effect (see Section 7.2.2) found with the General Linear Regression model, where poor households in the Chinese provice Hunan display a positive price elasticity of demand for rice.

Since Exceptional Model Mining strives to find interesting subsets of the dataset, the search space is potentially exponentially large in the number of records in the dataset at hand. This leaves us exposed to the multiple comparisons problem: we are considering a large number of candidates for what essentially amounts to a statistical hypothesis, hence it is likely that by pure random effects, we will unjustly designate some of these candidates as passing the test. Such candidates are called false discoveries. We have demonstrated in Chapter 8 how we can turn the problematic existence of false discoveries into a valuable tool that allows us to solve multiple practical problems in Subgroup Discovery and Exceptional Model Mining. We employ a swap randomization technique to create a search space that is identical to the original search space, but with all connections with and between the targets severed. Running the original SD/EMM algorithm on this search space results in descriptions that can be seen as false discoveries. We build a global model, the Distribution of False Discoveries (DFD), over the qualities of these false discoveries. This enables us to compute a p-value, corresponding to the null hypothesis that a found description is generated by the DFD. Refuting this null hypothesis for a description we found through SD/EMM, implies that this description is unlikely to be a false discovery. Beyond assessing the significance of descriptions, DFD modeling can deliver a quantitative assessment which quality measures are better than others in distinguishing real from false discoveries, and allows us to compute a minimum threshold for each quality measure, that a description must exceed to be considered reliably exceptional.

Having introduced all these instances with their model classes and quality measures, a natural question arising is why Exceptional Model Mining is desirable. We have three answers to that question. For starters, the trivial reason to perform EMM is that we learn things about our data. Extracting pieces of information from a raw dataset is the core business of data mining, and it should not be thought of lightly if a method does merely that. As we have seen in the experimental sections of Chapters 4–7, each description one can find with EMM is such a coherent nugget of information. Those real-life nuggets are far more actionable for a domain expert than the raw data could ever be. Given that EMM is able to capture a richer concept of "interestingness" than conventional Subgroup Discovery, EMM can retrieve descriptions containing more information out of the data than was possible beforehand, as long as the domain expert and the data miner together can formulate a model for the particular concept of interestingness that they strive to find.

Beyond the trivial reason, EMM is a great tool for metalearning. For example, in Chapter 5 we introduced an EMM instance with a classification model as target concept. Hence this instance finds descriptions for which the classification is performed in a substantially different manner than overall, which could be interesting to the researcher. Additionally, one could mine explicitly on a metadataset crafted from the results of a classification run. Suppose one is interested in predicting a numeric variable, for instance the number of days a court case will take to resolve. Having trained and tested a classifier, we end up with a metadataset of court cases, each with the real number of days and the predicted number of days. We can now use these real and predicted numbers as the two targets in an EMM run, for instance using the correlation model from Chapter 4. This EMM run will result in coherent subsets of the data for which the predictions of our classifier are particularly good or bad, which is potentially very useful information for further development or finetuning of the classification algorithm.

Lastly, the descriptions found though EMM may be directly applicable in a setting that is less exploratory and more oriented towards a concrete goal. The EMM instance with a Bayesian network model as target concept, which we discussed in Chapter 6, is a good example. While the original goal of the EMM instance is simply to find descriptions for which the conditional

dependence relations between the targets are unusual, the descriptions have demonstrated their capability to improve multi-label SVM classifiers in Chapter 9, though it does not work as well for decision trees. The main idea is that every description can be seen as a binary attribute of the dataset, indicating whether the record is covered by the description. These binary attributes highlight regions in the dataset where the labels interact in an unusual manner, so employing them in the learning phase may improve a multi-label classifier. Even though predictiveness was not considered at all when the descriptions were found, the classifier performance of SVM methods improved when these additional attributes were available.

As was shortly indicated in Section 3.2.2, efficiency can be an issue when running Exceptional Model Mining. Even with relatively modest parameter settings of the beam search and a reasonably-sized dataset, it is not uncommon to consider a number of descriptions that runs in the hundreds of thousands. For each of these descriptions, a model most be learned from data, and the dissimilarity of two models must be assessed to assign a quality to the subgroup. If either learning the model or assessing the dissimilarity is computationally too expensive, we end up with an intractable algorithm.

When the chosen model class is not too complex (e.g. correlation, the alternative simple linear regression model from Section 7.4, classification), the problem is scarcely more serious than for traditional Subgroup Discovery. For the general linear regression model efficient fitting algorithms exist, and based on upper bounds on eigenvalues and error terms, there is a scheme to prune descriptions on which it is relatively difficult to learn the model [23]. For the Bayesian network model however, the outlook is much bleaker. Without assumptions or heuristics, learning a Bayesian network from data is exponential in the number of vertices in the network [47]. Even with strong restrictions on the network structure, the problem remains superlinear [34]. Hence, for each of the hundreds of thousands of descriptions we learn a model at a high computational cost. We think that the Bayesian network model complexity is on the borderline of what can reasonably be incorporated into the Exceptional Model Mining framework.

To alleviate the efficiency issue, there are a few straightforward steps a researcher can take. When a parallel single-pass algorithm with sublinear memory requirements exists to learn the model from data, we can use the GP-Growth algorithm [72] to prune the search space. Also, choosing to compare the model for a description to the model for the whole dataset, rather than the model for the complement of the description, divides the number of models to be learned by two, as discussed in Section 3.2.2. If all else fails, since we usually resort to heuristic search in EMM, we can set the parameters bounding the search (such as the beam width w discussed in Section 3.1) tighter to reduce the number of descriptions to be evaluated, at the cost of an increased chance that exceptional descriptions remain undiscovered.

Exceptional Model Mining is in many respects a white box system. When employing an EMM instance on a particular domain, it is fairly simple to convey to a domain expert what kind of exceptionality is being sought after (by means of agreeing on the model class). The resulting descriptions are conjunctions of a few conditions on single attributes, which should be simple to interpret for the expert. Depending on the model class, a domain expert may also be able to properly investigate the discrepancies in fitted models. For instance, in the case of a correlation or regression model, this may enrich the expert's understanding of the result, but in the case of a Bayesian network fitted on a hundred animals it probably will not. We expect that deploying existing EMM instances in, or developing new EMM instances for, other fields, could lead to many fruitful collaborations between data miners and experts in those fields.