Exceptional Model Mining

Duivesteijn, W.

Chapter 1

Introduction

In their seminal 1996 paper [30], Fayyad, Piatetsky-Shapiro, and Smyth outlined their view on data mining, and what they called *KDD*, the then-emerging field of *Knowledge Discovery in Databases*. The basic problem that KDD strives to solve is the following: when presented with a set of raw data (which is usually too voluminous to inspect manually), distill information out of that dataset that is more compact, more abstract, or more useful. The authors wrote: "KDD is an attempt to address a problem that the digital information era made a fact of life for all of us: data overload." Since then, the internet has evolved from an additional source of information that we would occasionally dial into, to an always available vital necessity. Add to that the recent smartphone penetration into everyone's daily life, and we see that every person and company in the world generates more and more data. Hence the need for KDD methods has become evermore pressing.

Fayyad et al. divide the KDD process into nine stages, the seventh of which is *Data Mining*. After understanding the application domain, creating a dataset, cleaning and projecting the dataset, hypothesis selection and a few other preparatory steps, we arrive at the stage where we can search within a given dataset for "patterns of interest in a particular representational form or a set of such representations", before going to subsequent stages where patterns are interpreted and acted upon. In this dissertation we are mainly occupied with a subfield of data mining (the seventh stage of KDD), with some additional pattern interpretation (the eighth stage of KDD).

In the data mining phase, a given dataset is assumed. One can distinguish several methods to mine the dataset. The following were discussed by Fayyad et al.

Classification: mapping records of the dataset into one or several classes;

Regression: mapping records of the dataset to a real-valued prediction variable;

Clustering: identifying a finite set of categories to describe the dataset;

Summarization: finding a compact description for a subset of the dataset;

Dependency Modeling: finding a model that describes significant dependencies between variables;

Change and Deviation Detection: discovering substantial deviations in the data from the normative, or from previously measured values.

The data mining task we consider in this dissertation combines aspects of the last three methods, and has an application in the first.

The goal of *Local Pattern Mining* (LPM) is to find subsets of the dataset at hand, that are *interesting* in some sense. The goal is not to partition the dataset, and not to classify the dataset. We rather strive to pinpoint multiple (potentially overlapping) interesting subsets at the same time. The interestingness of a subset is gauged without considering the (lack of) coherence of records in the complement of the dataset, and without considering to what extent its interestingness is already represented by other found subsets: subsets are judged purely on their own merit. In LPM we are not quite interested in just any subset of the dataset; we are usually striving to find *subgroups*: subsets of the dataset that can be succinctly described in terms of conditions on attributes of the dataset. In this respect, LPM resembles the **Summarization** method introduced above. Originally, LPM was introduced as an *unsupervised* task where the *interestingness* was measured in terms of an unusually high frequency of a co-occurrence. In terms of such an interestingness definition, LPM resembles the **Deviation Detection** method introduced above.

The simplest form of *supervised* Local Pattern Mining is *Subgroup Discovery* (SD). In this task, one nominal attribute of the dataset is designated as the *target*. SD then strives to find subgroups of the dataset, for which

this target has an unusual distribution. Exceptionality of the distribution is usually gauged in terms of the relative frequencies of target values within the subgroup, compared to these frequencies on the whole dataset, and in terms of the size of the subgroup.

Unsupervised Local Pattern Mining (finding subgroups based on high frequency) and Subgroup Discovery (finding subgroups based on the distribution of one target) are interesting tasks. However, they do not encompass all possible forms of "interesting" subgroups of the dataset. In this dissertation we introduce the *Exceptional Model Mining* (EMM) framework, to accomodate a more general form of interestingness. In the EMM framework, the attributes of the dataset are partitioned into two: one part (the *descriptors*) is used to *define* subgroups on, and one part (the *targets*) is used to *evaluate* subgroups on. The concept of interest in subgroups is captured by learning, from (a subset of) the dataset, a *model* fitted on the targets. The goal of EMM in general is to find subgroups for which the model learned from the records belonging to the subgroup, has parameters that deviate substantially from the parameters of the model learned from the whole dataset. Alternatively, one can compare with the model learned from the complement of the subgroup; this choice will be discussed in detail in Section 3.2.2. EMM is instantiated by selecting two things: a *model class*, which indicates the type of interplay between targets we strive to find deviations for, and a *quality measure*, which quantifies the dissimilarity between two models from the model class. Striving to find unusual interplay between several targets, is where EMM resembles the **Dependency Modeling** method introduced by Fayyad et al.

To illustrate the difference between these Local Pattern Mining tasks, consider the following examples of subgroups one can find with them. In unsupervised LPM, there is no designated target attribute. One could find the subgroup of customers of a supermarket, that simultaneously buy coffee and milk. In Subgroup Discovery, suppose that the target is whether a person develops lung cancer. One could find the subgroup of smokers, whose lung cancer incidence is above average. In Exceptional Model Mining, suppose that the price of a house and its associated lot size are the two targets. One could then find the subgroup of inner city houses, for which the correlation between the two targets is substantially weaker than for the average house.

1.1 Overview

This dissertation consists of ten chapters, of which this introduction is **the first**. In this section, we shortly outline the remaining chapters, discussing the previous publications on which they are based, and giving the appropriate credits to (co-)authors.

In **Chapter 2: Motivation and Preliminaries**, we give motivating examples for Exceptional Model Mining, and introduce some notation. The examples have been discussed before in publications [23] and [25].

In **Chapter 3: The Exceptional Model Mining Framework**, we introduce the general Exceptional Model Mining framework. The EMM concept, including the introduction of the refinement operator, has appeared before in a paper by D. Leman, A. Feelders, and A. Knobbe, published in the proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2008) [71]. The remainder of Chapter 3, discussing our choices for the refinement operator and description language, algorithm and complexity analysis, how to define an EMM instance, related work, and the used software, is new.

The four subsequent chapters all introduce one choice of model class for EMM. None of these chapters explicitly discuss related work; since they instantiate the general framework of Chapter 3, we discuss all relevant related work there.

In **Chapter 4: Deviating Interactions – Correlation Model**, we discuss the EMM instance with the correlation between two numeric targets as model class. The original idea for this model class was first published by D. Leman, A. Feelders, and A. Knobbe, in the proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML / PKDD 2008) [71]. In Chapter 4, we reinterpret their work, and put it in the more general EMM context.

In **Chapter 5: Deviating Predictive Performance – Classification Model**, we discuss the EMM instance with a classifier on several unrestricted targets and one discrete output target as model class. Again, the original idea for this model class was first published by D. Leman, A. Feelders, and A. Knobbe [71], but the interpretation and EMM contextualization are new.

In **Chapter 6: Unusual Conditional Interactions – Bayesian Network Model**, we discuss the EMM instance with a Bayesian network on several nominal targets as model class. This work was published by W. Duivesteijn, A. Knobbe, A. Feelders, and M. van Leeuwen, in the proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010) [25].

In **Chapter 7: Different Slopes for Different Folks – Regression Model**, we discuss the EMM instance with a linear regression model on multiple targets as model class. In addition to the standard content of an EMM instance chapter, this chapter also contains a discussion of pruning the EMM search space with bounds on the developed quality measure. This work was published by W. Duivesteijn, A. Feelders, and A. Knobbe, in the proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2012) [23]. The idea for the simpler, alternative model class described in Section 7.4 was first published by D. Leman, A. Feelders, and A. Knobbe [71]; its interpretation and contextualization are new.

Having discussed Exceptional Model Mining instances, the following two chapters are dedicated to a related and an extended task. Contrary to the preceding four chapters, these following two chapters do come with their own related work discussions.

In **Chapter 8: Exploiting False Discoveries – Validating Found Descriptions**, we develop a method to determine the statistical significance of the outcome of supervised Local Pattern Mining tasks, such as Exceptional Model Mining. The quality of found descriptions is gauged against a model built over artificially generated false discoveries, to refute the hypothesis that a found description is also a false discovery. This method is additionally used to objectively compare different quality measures for the same task, by virtue of their capability to distinguish true from false discoveries. This work was published by W. Duivesteijn and A. Knobbe, in the proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2011) [24].

In **Chapter 9: Multi-label LeGo – Enhancing Multi-label Classifiers with Local Patterns**, we explore the additional value of descriptions found through EMM for the improvement of a global model. The descriptions found with the EMM instance in Chapter 6, with the Bayesian network model as target concept, highlight regions in the dataset where interplay between the targets is unusual. The ability to capture such interplay between labels is

what elevates a multi-label classifier over multiple single-label classifiers. Hence, employing the descriptions as binary attributes for a multi-label classifier should improve classifier performance. In this chapter we discuss the extent to which this *LeGo approach* [37, 57] indeed improves performance. This work was published by W. Duivesteijn, E. Loza Mencía, J. Fürnkranz, and A. Knobbe, in the proceedings of the 11$^{\text{th}}$ International Symposium on Intelligent Data Analysis (IDA 2012) [26]. An extended version was published by the same authors as a technical report of the Technische Universität Darmstadt [27]. This being joint work involving another PhD student, two reinterpretations and contextualizations of publications [26] and [27] are available. The one is Chapter 9 of this dissertation, and the other has appeared as a chapter in the Ph.D. dissertation of E. Loza Mencía [73].

In **Chapter 10: Conclusions**, we draw general conclusions from all preceding chapters. We discuss rationales why Exceptional Model Mining is not only a desireable, but also a practically useful framework to have.