Exceptional Model Mining

Duivesteijn, W.

Chapter 2

Motivation and Preliminaries

Finding 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: simply identifying the peculiarly-behaving records. The characteristic feature of local pattern mining techniques that separates them from such outlier detection methods, is that in local pattern mining, we are not just looking for any outlying record or set of records in the data. Instead, we are looking for subgroups: coherent subsets for which we can formulate a concise description in terms of conditions on attributes of the data. The existence of such descriptions makes the subgroups more actionable: if we can tell a drug manufacturer that ten of his patients react badly to a certain type of medication, this doesn't help him much, but if we can tell him instead that the group of smokers under the age of thirty react badly, this gives the manufacturer a clear indication in which direction to find a solution to his problem.

When the target concept in a dataset can no longer be captured by one particular attribute, but we still want to find exceptional subgroups in the dataset, we find a need for Exceptional Model Mining. As an example of a relatively complex target concept, consider the research performed by Robert T. Paine in 1963 and 1964 in Makah Bay, Washington [86]. It concerns the carnivore starfish *Pisaster ochraceus* whose presence kept a marine ecosystem consisting of 15 species stable. In this system, the sponge *Haliclona* was browsed upon by the nudibranch *Anisodoris*. When *Pisaster* was artificially removed, the bivalve *Mytilus californianus* and

the barnacles *Balanus glandula* and *Mitella polymerus* rapidly grew and crowded out other species. In total, only 8 species remained. Also, the sponge-nudibranch food chain was displaced, and the anemone population was reduced in density. Counterintuitively, when present, *Pisaster* did not eat any of these last three species.

In the studied ecosystem, *Pisaster* was the top carnivore: it consumed other species, but no other species consumed him, and *Pisaster* was the only species in the system for which both these statements held. This made Paine et al.'s research very relevant from a biological point of view; up until that point, it was generally assumed that removing the top carnivore from an ecosystem would increase diversity, but the *Pisaster* experiment proved that that was not necessarily the case.

Paine remarks that the food chains are strongly influenced by *Pisaster*, but by an indirect process. When dealing with a dataset detailing the presence of individual species, existing methods can probably detect simple patterns in the ecosystem, such as the growth of *Mytilus*, *Balanus* and *Mitella* and the decline in the number of species when *Pisaster* is removed. However, the more indirect influence of *Pisaster* on processes such as a food chain it is not direcly related to, for instance between *Haliclona* and *Anisodoris*, cannot be found by looking at single species or even correlations between pairs of species: the (in-)dependence between *Haliclona* and *Anisodoris* is conditional on the presence of *Pisaster*.

Paine models the food chains in the ecosystem as a Bayesian network. In order to find subgroups where the food chains between species are substantially different from the norm, we need to be able to detect the indirect processes that can be captured with a Bayesian network. Using an Exceptional Model Mining instance, we can for instance find subgroups defined by environmental parameters in which complex food chains are displaced. The ability to cope with Bayesian networks makes the same EMM instance applicable to datasets from such diverse fields as information retrieval [9], traffic accident reconstruction [18], medical expert systems [20], gene expression in computational biology [33], and financial operational risk [82].

Another EMM instance could for example be used to find evidence for the *Giffen effect* in data. This effect can be seen as a form of Simpson's

paradox for regression models. The economic law of demand states that, all else equal, if the price of a good increases, the demand for the product will decrease. Sir Robert Giffen described conditions under which this law does not hold [77]. The classic example concerns extremely poor households, who mainly consume cheap staple food, and relatively rich households in the same neighborhood, who can afford to enrich their meals with a luxury food. In this situation, when the price of the staple food increases, there will be a point where the relatively rich households can no longer afford the luxury food. These people need to uphold their calorie intake. Hence, they react by consuming more of the cheapest food available to them, which is the staple food whose price just increased. For the relatively rich households in this poor neighborhood, an increase in the price of the staple food, will lead to an increase in the demand for the staple food. Notice that this relation does not hold for the extremely poor households: they consume only the staple food to begin with, so when the price increases they can simply afford less of it.

For a long time, the Giffen effect was a controversial theory in Economics, since no real-life dataset featuring the effect was available. In 2008, more than a century after the theorem was formulated for the first time, Nolan and Jensen published a paper [53] containing the first real-world dataset containing the Giffen effect, for rice in Hunan, China. Their field study entailed distributing vouchers among randomly drawn households, with which the recipients could buy rice at a lower price. The authors monitored the price of and the demand for rice before, during, and after the voucher programme, as well as a plethora of alternative factors that could influence demand. The relation between the demand for rice and the influencing factors (including the price of rice) was captured by a regression model. Nolan and Jensen observed that the households consuming less than 80% of their calorie intake through rice, i.e. the relatively rich households in this poor neighborhood, displayed the Giffen effect, while the other households did not.

The group of relatively rich households in a poor neighborhood is a subgroup. The subgroup displays an unusual interaction between multiple targets, as captured by the regression model. Hence, subgroups displaying the Giffen effect can be automatically detected by an Exceptional Model Mining instance, mining for an unusual slope of a regression line.

2.1 Preliminaries

Having motivated Exceptional Model Mining in the previous section, we will formally introduce the framework in the next chapter. To that end, we first introduce some definitions and notations that will be used throughout the remainder of this dissertation. Any symbol introduced in this section may pop up at any given moment; we assume its meaning to be understood by the reader from this point on.

We assume a dataset Ω to be a bag of N *records* $r \in \Omega$ of the form

$$r = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m)$$

where k and m are positive integers. We call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $\ell_1, \ldots, \ell_m$ the *target attributes* or *targets* of r. The descriptors are taken from an unrestricted domain $\mathcal{A}$. In later chapters we will learn models from a selected *model class* over the targets; restrictions on the type of each target may be imposed by the choice of model class. We refer to (elements of) the i^{th} record by superscript i.

For our definition of subgroups we need to define *descriptions*. In practice, descriptions will usually be taken from a description language $\mathcal{D}$, to be chosen by the user. We will leave this concept abstract for now; a particular choice we make for $\mathcal{D}$ will be discussed in Section 3.1.1. However, mathematically, we will define descriptions as functions $D : \mathcal{A} \to \{0, 1\}$. A description D *covers* a record r^i if and only if $D\left(a_1^i, \ldots, a_k^i\right) = 1$.

Definition (Subgroup). A *subgroup* corresponding to a description D is the bag of records $G_D \subseteq \Omega$ that D covers, i.e.

$$G_D = \left\{ r^i \in \Omega \;\middle|\; D\left(a_1^i, \ldots, a_k^i\right) = 1 \right\}$$

From now on we omit the D if no confusion can arise, and refer to a subgroup as G. We will freely associate subgroups with their descriptions and vice versa. Also, the 'patterns' in the commonly used term 'Local Pattern Mining' are equivalent to our descriptions, and hence imply subgroups. These terms will all be used interchangably when a clear separation between the concepts is not necessary. Whenever it is clear that we have a particular subgroup G in mind, we will write n for the number of records in

that subgroup: $n = |G|$, to which we will also refer as the *coverage* of the description. The complement of a subgroup is denoted by G^C, and for its number of records we write n^C. Hence, $G^C = \Omega \backslash G$, and $n^C = N - n$.

In order to objectively evaluate a candidate description in a given dataset, we need to define a *quality measure*. For each description D in the description language $\mathcal{D}$, this function quantifies the extent to which the subgroup G_D induced by the description deviates from the norm.

Definition (Quality Measure). A *quality measure* is a function $\varphi : \mathcal{D} \to \mathbb{R}$ that assigns a unique numeric value to a description D.

Since descriptions imply subgroups, we will occasionally write $\varphi(G)$ and refer to the quality of a subgroup.