Expanding Exceptional Model Mining on Unstructured Data

TU/e supervisor: dr. W. Duivesteijn

Introduction to Local Pattern Mining

In contrast with building a one-size-fits-none global model for the entire dataset, Local Pattern Mining is a data mining subfield that acknowledges that a dataset may encompass multiple groups that each behave in their own way. The goal of Local Pattern Mining is to discover interesting subsets in a dataset. The subsets must satisfy two conditions:

- they must be *interpretable*;
- they must display *exceptional behavior*.

By *interpretable*, we typically mean that we are only interested in subsets that can be defined as a conjunction of a few conditions on columns of the dataset at hand. Hence, we do not want to find "this group of ten datapoints is interesting", but instead:

$$\text{smoking} = \text{yes} \wedge \text{age} \leq 25 \rightsquigarrow \text{(some exceptional behavior)}$$

Such subsets are called *subgroups*. Returning only subgroups that can be defined in terms that the dataset owner can understand, ensures that we find things that are actionable: we can build a policy on this sort of knowledge.

Exceptional behavior can be defined in many many ways. A wild variety of Local Pattern Mining methods exist, typically distinguished by what choice they make here. For instance:

Frequent Itemset Mining **[1]** : defines it in an unsupervised manner as "high frequency";

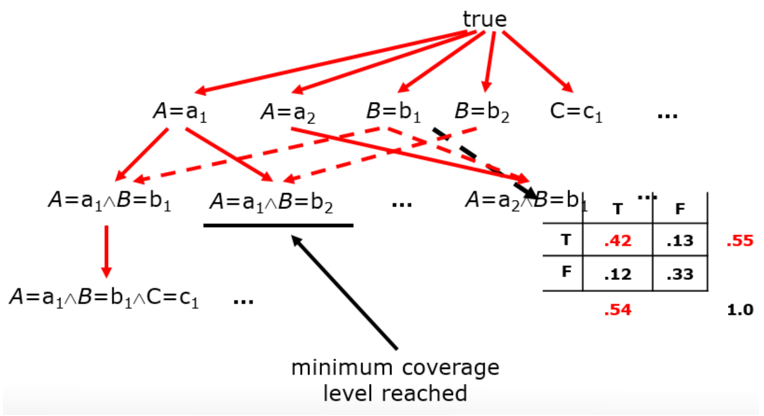
Subgroup Discovery **[5, 8, 4]** : defines it in a supervised manner as "unusual distribution of a single target columm";

Exceptional Model Mining **[6, 3]** : defines it in a supervised manner as "unusual interaction between several target columns".

Still, within those definitions, many parameters can be chosen and altered to end up with a sprawling cornucopia of interesting mining tasks and results.

Search Lattice

The subgroup search space can be represented as a lattice of candidate subgroups, where each level of the search consists of specializations of subgroups from the layer above:



When the dataset consists of many attributes with many distinct values, the number of subgroups that can be defined explodes, and the lattice above becomes ridiculously (intractably) wide. This often prohibits exhaustive search.

The Project

Both subgroup definitions and target interactions make all the sense in a world where your dataset comes in traditional, flat-table form. But lately, the world has been moving towards structured data: images, video, text, music, etcetera. It is apriori unclear what constitutes a meaningful subgroup description language, and which aspect of the data would even begin to take the role of a target attribute (let alone interesting interaction between multiple of those). The goal of this project is to generalize EMM to unstructured data; it is up to us to decide what that means.

Current Status

Two papers exist in this general direction. The first [2] strives to perform some form of Subgroup Discovery in some choice of unstructured data. The second [7] improves on the first, and expands to the more generic Exceptional Model Mining setting. The second publication outlines limitations of both existing papers; we are at the bleeding edge of science so there is *so* much choice of what to do next. We're in the pioneering stage; boldly go where noone has gone before!

Requirements

Since almost noone has done anything remotely close to this problem space, there will be little guidance in literature of what to do. You need to feel confident that you can handle blazing a trail that is currently still unmarked. When done well, I expect this project to lead to a publication at a top-level data mining or machine learning conference, or a journal.

References

[1] Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994: 487-499

[2] Ali Arab, Dev Arora, Jialin Lu, Martin Ester: Subgroup Discovery in Unstructured Data. CoRR abs/2207.07781 (2022)

[3] Wouter Duivesteijn, Ad Feelders, Arno J. Knobbe: Exceptional Model Mining - Supervised descriptive local pattern mining with complex target concepts. Data Min. Knowl. Discov. 30(1): 47-98 (2016)

[4] Francisco Herrera, Cristóbal J. Carmona, Pedro González, María José del Jesus: An overview on subgroup discovery: foundations and applications. Knowl. Inf. Syst. 29(3): 495-525 (2011)

[5] Willi Klösgen: Explora: A Multipattern and Multistrategy Discovery Assistant. Advances in Knowledge Discovery and Data Mining 1996: 249-271

[6] Dennis Leman, Ad Feelders, Arno J. Knobbe: Exceptional Model Mining. ECML/PKDD (2) 2008: 1-16

[7] Niels Schelleman: Exceptional In So Many Domains. MSc thesis, Technische Universiteit Eindhoven, 2023. See attachment to this project description.

[8] Stefan Wrobel: An Algorithm for Multi-relational Discovery of Subgroups. PKDD 1997: 78-87