Pimp my BUS: Improving a Minicluster-based Deterministic Pattern Sampling Algorithm for Exceptional Model Mining

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";

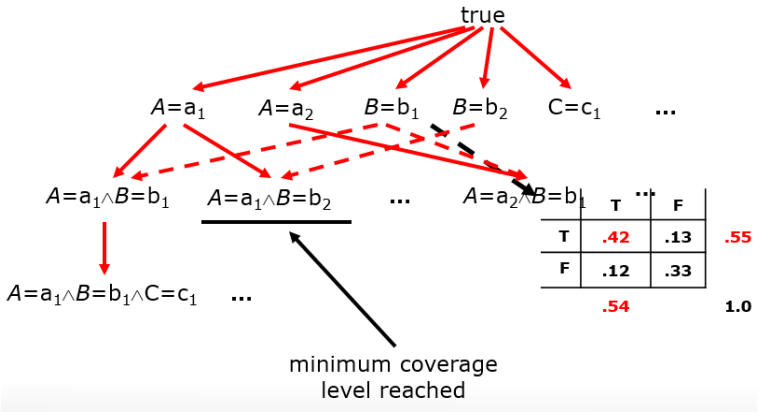
Subgroup Discovery **[5, 8, 4]** : defines it in a supervised manner as "unusual distribution of a single target columm";

Exceptional Model Mining **[6, 2]** : 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.

A canonical way to traverse this search lattice is through *Beam Search* [2, Algorithm 1]: a top-down algorithm where we start with the most general subgroup there is: no conditions, so the subgroup encompasses all records in the dataset. We then generate all possible refinements (conjoining to the current condition list a single condition on a single attribute) of this seed subgroup as candidates for the next level: their qualities are determined, and the best w (to be set by the user) are kept as the *beam* for the next level. On the next level each subgroup in the beam is refined into a new set of candidates, and this process repeats until the maximum depth d (to be set by the user) is reached. Hence, Beam Search starts from the top, and makes a level-wise downward traversal through the lattice until a certain maximum depth is reached; each level beyond depth 1 is only partially traversed.

All Aboard the BUS!

A brand new way to traverse this search lattice is through *Bottom-Up Search* [3, Algorithm 2]: a deterministic pattern sampling algorithm where we start from the most specific subsets we see as promising: mini-clusters of $2, 3, \ldots, z$ observations that lie closest together in target space. We then generate all possible subgroup definitions that encompass all records in a mini-cluster (as well as as many other records that happen to also fall within this defintion; we're generating candidate hypotheses here). This constitutes one bottom-up jump from the lowest level of the lattice to a higher-up selection of the search lattice (quite likely scattered across multiple search levels). We determine the quality of all subgroups thusly generated, and the best k are reported as the best subgroups.

The Project: Pimp My BUS

BUS tends to be more efficient than Beam Search at no substantial loss of subgroup quality. This is nice. The goal of this project is to extend and improve it, for instance along the following axes:

- BUS has been evaluated on two kinds of target interactions: exceptional trends and exceptional preferences. Many more exist. Can we find target interactions where it doesn't work as well?
- we know that BUS doesn't always work best, but we have no idea why. Can we do some metalearning to find context under which BUS particularly strongly outperforms Beam Search or vice versa?
- the current version of BUS generates one collection of candidate subgroups and evaluates those. It might be better if BUS is allowed to make some local traversals in the search lattice: from the candidate

subgroups BUS defines, can we refine/remove single conditions and would that make the algorithm perform better?

- would non-convex miniclusters provide better seeds for BUS?

Current Status

As a sanity check, we should have a look how distinct these ideas are from the ones introduced by Umek and Zupan [7]. Other than that, the attached MSc thesis [3] is the only work in this direction, so the path to extensions is rather straightforward.

Requirements

If you have any clever ideas about search algorithms, this would be valuable expertise to bring into this project. Other than that, a standard level of data mining background and a natural curiosity that drives you to improve things would suffice to excel in this project. When done well, I expect this project to lead to a publication at a top-level data mining or machine learning conference, or perhaps a journal.

References

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

[2] 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)

[3] Bart Engelen: Bottom-Up Search: A Distance-Based Search Strategy for Supervised Local Pattern Mining on Multi-Dimensional Target Spaces. MSc thesis, Technische Universiteit Eindhoven, 2023. See attachment to this project description.

[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] Lan Umek, Blaz Zupan: Subgroup discovery in data sets with multi-dimensional responses. Intell. Data Anal. 15(4): 533-549 (2011)

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