Deriving Upper Confidence Bounds to Expand Monte-Carlo Tree Search to More EMM Model Classes

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";

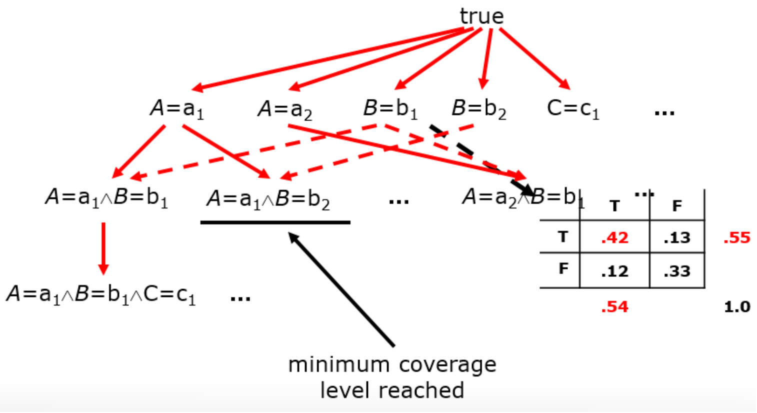
Subgroup Discovery **[6, 10, 5]** : defines it in a supervised manner as "unusual distribution of a single target columm";

Exceptional Model Mining **[7, 4]** : 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

Two groups of French researchers have developed a variant of the famous Monte-Carlo Tree Search (MCTS) algorithm, to efficiently traverse this search lattice. Their papers can be found here [2, 8]. However, both works seek subgroups that discriminate a class label and as such, evaluate candidate subgroups with Weighted Relative Accuracy (WRAcc) as a quality measure. This makes the algorithms applicable on Subgroup Discovery, but not necessarily on Exceptional Model Mining.

The project here is to expand MCTS to Exceptional Model Mining. When exceptionality is no longer gauged in terms of an unusual distribution of a single target, but as an interaction between multiple targets, it becomes unclear:

- what type of Upper Confidence Bounds we would need in the Select phase of MCTS;
- what aggregation functions would be appropriate in the Update phase of MCTS, to do justice to the specific quality measures of EMM.

For different kinds of EMM target interactions, it is quite likely that these questions require different answers. The most promising initial candidate EMM model classes to investigate are the relatively simple ones introduced in [7], the rank correlation model class from [3], and the model class introduced in the first paper on Exceptional Preferences Mining [9]. Part of the project is to identify promising candidate model classes yourself.

Current Status

To the best of my knowledge, no other papers exist on this topic beyond [2, 8]. However, it's probably wise to survey the literature by the authors of that paper, to make sure that we're not missing anything.

Requirements

The main task of a student in this project is to derive and prove things. So, some affinity with math is desired. 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] Guillaume Bosc, Jean-Franççis Boulicaut, Chedy Raïssi, Mehdi Kaytoue: Anytime discovery of a diverse set of patterns with Monte Carlo tree search. Data Min. Knowl. Discov. 32(3): 604-650 (2018)

[3] Lennart Downar, Wouter Duivesteijn: Exceptionally Monotone Models - The Rank Correlation Model Class for Exceptional Model Mining. ICDM 2015: 111-120

[4] 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)

[5] 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)

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

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

[8] Romain Mathonat, Diana Nurbakova, Jean-Franççis Boulicaut, Mehdi Kaytoue: Anytime mining of sequential discriminative patterns in labeled sequences. Knowl. Inf. Syst. 63(2): 439-476 (2021)

[9] Cláudio Rebelo de Sá, Wouter Duivesteijn, Carlos Soares, Arno J. Knobbe: Exceptional Preferences Mining. DS 2016: 3-18

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