Deriving Valuation Bases to Expand GP-Growth 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";

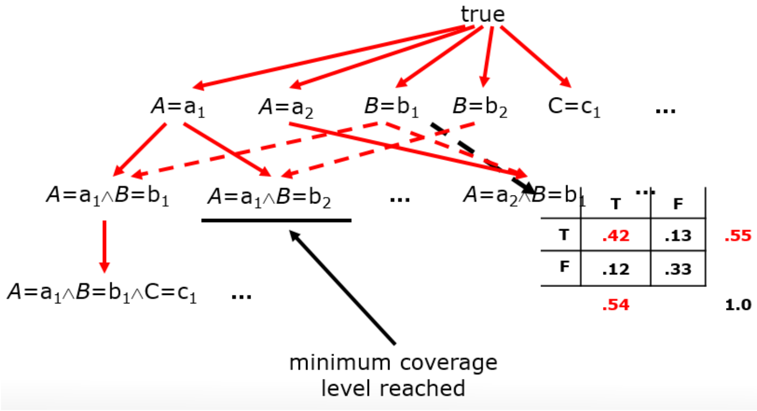
Subgroup Discovery **[6, 10, 5]** : defines it in a supervised manner as "unusual distribution of a single target columm";

Exceptional Model Mining **[7, 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.

For Frequent Itemset Mining (where interestingness = frequency), one can derive a clever manner to efficiently enumerate frequent itemsets without constructing the full lattice first. Instead, all relevant information is summarized in a tree that we construct ourselves. This is known as the FP-Growth algorithm [4].

The Project

The GP-Growth algorithm [8] is the FP-Growth equivalent for Exceptional Model Mining. It's a bit more complicated, since in EMM, interestingness is a little more complicated than frequency. Instead, GP-Growth demands the derivation of a *valuation basis*: for a type of unusual interaction between several target columns, we must be able to concisely summarize the information required to determine how exceptional a subgroup can be that lives lower than the current node in the search lattice.

For example, if we are currently looking at the node "$A = a_1$" in the lattice, and if our target interaction is the correlation between two numeric columns x and y, then the valuation basis needs to store four values: the sum of x-values in the subgroup $A = a_1$, the sum of y-values, the sum of x^2-values, and the sum of y^2-values. These values provide sufficient information to reconstruct the correlation between x and y in the subgroup. Moreover, when moving lower in the lattice, for instance to $A = a_1 \wedge B = b_1$, simple subtraction suffices to update the four values to the new subgroup. This provides access to efficient computation of the exceptionality of subgroups lower in the lattice.

The GP-Growth paper [8] provides two main results:

- a formal proof that GP-Growth works when a valuation basis can be found that performs the required computations in linear time and with sublinear memory requirements;
- the concrete derivation of valuation bases for four specific kinds of target interaction: variance, correlation, simple linear regression, decision table majority classifier; along with rationales why valuation bases cannot be defined for logistic regression and Bayesian networks.

This quite nicely and accurately covers the target model classes available in EMM literature in the summer of 2012, but eleven years have passed since. The goal of this project is:

- for every EMM model class available in literature:
  - either derive a condensed valuation basis for it and empirically show that it works;
  - or prove that it cannot be done.

The most promising initial candidates for a condensed valuation basis would be the rank correlation model class from [2] 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 valuation bases for GP-Growth have been derived since the initial paper [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] Lennart Downar, Wouter Duivesteijn: Exceptionally Monotone Models - The Rank Correlation Model Class for Exceptional Model Mining. ICDM 2015: 111-120

[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] Jiawei Han, Jian Pei, Yiwen Yin: Mining Frequent Patterns without Candidate Generation. SIGMOD Conference 2000: 1-12

[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] Florian Lemmerich, Martin Becker, Martin Atzmueller: Generic Pattern Trees for Exhaustive Exceptional Model Mining. ECML/PKDD (2) 2012: 277-292

[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