Exceptional factorisation - Matrix factorisation for exceptional model mining

Matthijs van Leeuwen
www.patternsthatmatter.org

Context

Situated in the research area of exploratory data mining, *exceptional model mining* (EMM) is a generalisation of *subgroup discovery* (SD). As such, it is concerned with the discovery of patterns, subsets of the data where the target attribute(s) show an interesting difference in distribution, compared to that of the entire dataset. Whereas SD only considers a single target attribute, EMM generalises this to complex targets, i.e., multiple target attributes or even social networks. The patterns, called subgroups, are usually found through combinatorial search, which can be either heuristic or exhaustive.

Matrix factorisation (MF), on the other hand, is a recent trend in data mining and machine learning that aims to factorise a large matrix into two (or more) smaller matrices. The idea is that these smaller matrices are easier to interpret, but they can also be used for prediction or recommendation tasks. The most common instance of MF is probably non-negative matrix factorisation, but other instances exist for, e.g., Boolean and integer matrices.

Goal

The goal of this project is to develop theory and algorithms that perform Exceptional Model Mining by means of Matrix Factorisation. For this, the EMM problem will have to be formulated as a MF problem. This is an interesting challenge on itself, at it is not straightforward 1) how to distinguish subgroup description from target, and 2) how to model the EMM problem as a global optimization problem (interestingness of subgroups is commonly quantified on a per-subgroup basis). By succeeding in doing this, however, it becomes possible to find non-redundant and compact sets of 'exceptional' subgroups, which can be directly derived from the factors.

Research questions

Example research questions include:

- How can the EMM problem be modelled as a MF problem?
- Which EMM instances can be covered by this formalisation?
- How do the results obtained using this formalisation compare to those obtained using existing methods? (Interestingness? Redundancy?)
- Is it possible to automatically determine a suitable number of patterns?

Realisation

1. Literature study.
2. Determine scope of the project, write research plan.
3. Develop theory and algorithm(s).
4. Implement algorithm(s).
5. Run experiments to evaluate algorithm(s).
6. Write thesis.

Student profile

Good understanding of data mining problems and algorithms; good implementation skills.

Relevant literature

[1] Dennis Leman, Ad Feelders, Arno J. Knobbe. *Exceptional Model Mining*. In: Proceedings of ECML/PKDD’08. 2008: 1-16
http://link.springer.com/chapter/10.1007%2F978-3-540-87481-2_1

[2] van Leeuwen, M. & Knobbe, A.J. *Diverse Subgroup Set Discovery*. In: Data Mining and Knowledge Discovery, special issue ECML PKDD'11, vol.25(2), pp 208-242, Springer, 2012.
http://patternsthatmatter.org/pubs/2012/diverse_subgroup_set_discovery-vanleeuwen,knobbe.pdf

[3] van Leeuwen, M. *Maximal Exceptions with Minimal Descriptions*. In: Data Mining and Knowledge Discovery, special issue ECML PKDD'10, vol.21(2), pp 259-276, Springer, 2010.
http://patternsthatmatter.org/pubs/2010/maximal_exceptions_with_minimal_descriptions-vanleeuwen.pdf

[4] Pauli Miettinen. *Generalized Matrix Factorizations as a Unifying Framework for Pattern Set Mining: Complexity Beyond Blocks*. In: Proceedings of ECML/PKDD’15, 2015.
http://link.springer.com/chapter/10.1007%2F978-3-319-23525-7_3

[5] Inderjit S. Dhillon & Suvrit Sra. *Generalized Nonnegative Matrix Approximations with Bregman Divergences*. In: Proceedings of NIPS’05, 2005.

[6] Tandon, Rashish & Suvrit Sra. *Sparse nonnegative matrix approximation: new formulations and algorithms*. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2010.