Interactive discovery of non-redundant pattern sets

Promotor: Luc De Raedt, Matthijs van Leeuwen
Daily advisor: Matthijs van Leeuwen

The aim of this thesis is to design an algorithm that allows a data miner to interactively discover non-redundant sets of interesting patterns from large datasets.

In general, the goal of data mining is to discover interesting knowledge from large databases, often in the form of patterns. It is a challenging field from multiple perspectives: How do you formally define the problem at hand? How do you algorithmically solve it? And what experiments do you perform to empirically show that your solution is useful for real world applications?

Subgroup discovery is a research area within data mining and aims at finding subgroups, i.e. regions in a database that substantially deviate from the overall database with respect to a specified target. If we consider e.g. a database containing real estate information, the task could be to identify and characterize groups of houses that are relatively expensive. Or, an example from bioinformatics, what environmental and hereditary factors contribute to higher risks for developing certain diseases?

Subgroup discovery is an established paradigm and a bunch of discovery algorithms exist. Some of these are heuristic, others employ exhaustive search. All kinds of data can be used, from categorical to numerical, and from structured to multi-relational data. However, the generic approach suffers from 2 major problems:

1) Results typically display a tremendous amount of redundancy. That is, the top-k patterns are often almost identical, making them hardly interesting to a domain expert. The cause of this problem lies in the fact that subgroups are assessed only individually, not together. Clearly, we should be looking for sets of patterns instead; subgroup sets that are both interesting and non-redundant.

2) Background knowledge of a domain expert is seldom used, neither to exclude results nor to guide the search process. That is, the results of a search could be completely uninteresting to the expert, because they only contain `common knowledge'. We therefore propose to incorporate background knowledge in the search process through user interaction, allowing the user to effectively steer the search towards parts of the hypothesis space that he or she is interested in.

In this thesis project you are invited to tackle the two problems stated above. Both redundancy [1] and user interaction [2] are hot topics in the research community, but neither of the individual problems is solved yet (and certainly not together). You are invited to build upon recent work and inject your own ideas to achieve true interactive and non-redundant subgroup discovery.

Profile: student with interest in machine learning or data mining; good understanding of algorithms and good implementation skills.
