Duality in Subgroup Discovery: Solving the Exhaustive Search Problem Using Formal Concept Analysis and Mixed Integer Linear Programming

In the present paper we analyze the exhaustive subgroup discovery problem for a binary target variable in the language of formal concept analysis. In particular we use here the duality between formal concept extents and formal concept intents. We model subgroups with binary variables that simultaneously describe both the extent and the intent of a formal concept. This allows to solve the exhaustive search problem by formulating it as a mixed integer linear (MILP) problem. After formulating this MILP approach, we furthermore describe several computational techniques that can then be additionally applied. Some of these tricks are e.g.:

i) Dropping of integrality constraints: Due to the structure of the MILP formulation, certain integrality constraints of the problem (which in its original form is a purely binary program), can be dropped. Here there are several possibilities.

ii) For data sets with duplicated data points, weighting techniques can help to reduce the problem.

iii) A further analysis shows that one does not need to maximize the quality function over the whole system of all concepts. It suffices to optimize over the smaller system of all concepts that are generated by objects a positive target value.

iv) Duplicated and reducible attributes of the context can simply be dropped.

v) A priori known valid formal implications between objects and attributes (or both) can be additionally implemented as inequality constraints to reduce the gap between the relaxed and the non-relaxed MILP formulation. This is especially very interesting for situations where one has a very specific conceptual scaling of certain attributes (i.e., nominal, hierarchically nominal, interordinal attributes or attributes that represent ranking data). For the special case of formal implications with one-element premises one can implement only the transitive reduction of the corresponding transitive implicational relation.

vi) If one has an a priori demand for a minimum value of the quality function (this is for example the case in the situation where one wants to check statistical significance of the subgroup discovery result by means of a permutation test), then one can also make use of a priori types of optimistic estimates that would additionally reduce the search space. (The concrete implementation of these optimistic estimates can be done by SOS-type (or big-M-type) constraints).

Finally we apply our MILP approach to several data sets. It turns out that the above tricks lead to a reduction of computational time up to factors of around 30 for our data examples. Additionally, we also compare our approach to existing algorithms.