ROCsearch — An ROC-guided Search Strategy for Subgroup Discovery*

Marvin Meeng[1], Wouter Duivesteijn[2], and Arno Knobbe[1]

[1] LIACS, Leiden University, {m.meeng,a.j.knobbe}@liacs.leidenuniv.nl
[2] Fakultät für Informatik, LS VIII, Technische Universität Dortmund, wouter.duivesteijn@tu-dortmund.de

Subgroup Discovery (SD) aims to find coherent, easy-to-interpret subsets of the dataset at hand, where something exceptional is going on. Since the resulting subgroups are defined in terms of conditions on attributes of the dataset, this data mining task is ideally suited to be used by non-expert analysts. The typical SD approach uses a heuristic beam search, involving parameters that strongly influence the outcome. Unfortunately, these parameters are often hard to set properly for someone who is not a data mining expert; correct settings depend on properties of the dataset, and on the resulting search landscape. To remove this potential obstacle for casual SD users, we introduce ROCsearch [1], a new ROC-based beam search variant for Subgroup Discovery.

On each search level of the beam search, ROCsearch analyzes the intermediate results in ROC space to automatically determine a sensible search width for the next search level. Thus, beam search parameter setting is taken out of the domain expert's hands, lowering the threshold for using Subgroup Discovery. Also, ROCsearch automatically adapts its search behavior to the properties and resulting search landscape of the dataset at hand. Aside from these advantages, we also show that ROCsearch is an order of magnitude more efficient than traditional beam search, while its results are equivalent and on large datasets even better than traditional beam search results.

References

1. M. Meeng, W. Duivesteijn, A. Knobbe, ROCsearch – An ROC-guided Search Strategy for Subgroup Discovery, Proc. SDM, pp. 704–712, 2014.

* This research is supported in part by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C1.