**Rough Set Approach to Scalable Similarity Measures – Machine Learning Applications**

Ability to operate with similarities is crucial for many machine learning scenarios. For example, in interactive data labeling (powered by active learning), it is useful to focus on the not-yet-labeled objects which are (1) not similar to the already-labeled ones or (2) similar to the labeled objects with respect to which a current machine learning model is particularly uncertain. As another example, when one wants to diagnose a possible cause of mistake of a machine learning model with respect to a given current object, it is a good idea to analyze similar objects in the model’s performance history.

In this talk, we consider such scenarios from the perspective of making the underlying similarity computations scalable and insightful. We discuss one particular idea of expressing similarities, which is based on the ensembles of decision reducts – one of the fundamental notions in the theory of rough sets. We discuss different variations of decision reducts (such as approximate reducts and bireducts) which are usually utilized to learn simple decision models from the data. Additionally – which is our special interest in this talk – they form data partitions based on easily interpretable subsets of attributes. And accordingly – somewhat analogously to the LSH methods (which are not particularly insightful, however) – their ensembles form the data-partition-based similarities.

Acknowledgment: This research was co-funded by Smart Growth Operational Programme 2014-2020, financed by European Regional Development Fund, in frame of project POIR.01.01.01-00-0213/19, operated by National Centre for Research and Development in Poland.