Optimizing the Knowledge Discovery Process through Semantic Meta-Mining

Melanie Hilario

Computer Science Department
University of Geneva
Geneva, Switzerland

Abstract. I will describe a novel meta-learning approach to optimizing the knowledge discovery or data mining (DM) process. This approach has three features that distinguish it from its predecessors. First, previous meta-learning research has focused exclusively on improving the learning phase of the DM process. More specifically, the goal of meta-learning has typically been to select the most appropriate algorithm and/or parameter settings for a given learning task. We adopt a more process-oriented approach whereby meta-learning is applied to design choices at different stages of the complete data mining process or workflow (hence the term meta-mining). Second, meta-learning for algorithm or model selection has consisted mainly in mapping dataset properties to the observed performance of algorithms viewed as black boxes. While several generations of researchers have worked intensively on characterizing datasets, little has been done to understand the internal mechanisms of the algorithms used. At best, a few have considered perceptible features of algorithms like their ease of implementation or their robustness to noise, or the interpretability of the models they produce. In contrast, our meta-learning approach complements dataset descriptions with an in-depth analysis and characterization of algorithms - their underlying assumptions, optimization goals and strategies, together with the structure and complexity of the models and patterns they generate. Third, previous meta-learning approaches have been strictly (meta) data-driven. To make sense of the intricate relationships between tasks, data and algorithms at different stages of the data mining process, our meta-miner relies on extensive background knowledge concerning knowledge discovery itself. For this reason we have developed a data mining ontology, which defines the essential concepts and relations needed to represent and analyse data mining objects and processes. In addition, a DM knowledge base gathers assertions concerning data preprocessing and machine learning algorithms as well as their implementations in several open-source software packages. The DM ontology and knowledge base are domain-independent; they can be exploited in any application area to build databases describing domain-specific data analysis tasks, datasets and experiments. Aside from their direct utility in their respective target domains, such databases are the indispensable source of training and evaluation data for the meta-miner. These three features together lay the groundwork for semantic meta-mining, the process of mining DM meta-data on the basis of data mining expertise distilled in an ontology and knowledge base.