Probably Approximately Correct Implication Bases

Tom Hanika

Department of Mathematics
University of Kassel

Resumo: Extracting theory, i.e., implicational knowledge, from a data set is one of the main goals in modern machine learning. Due to complexity matters, this process is usually restricted to a subset of rules that is sound and complete with respect to that theory, called basis. However, computing knowledge bases for data sets, even those that are minimal in size, such as the so-called canonical basis, is in general an infeasible problem.

To circumvent this computational limitation, we revisit the notion of probably approximately correct (PAC) implication bases from the literature and present a first formulation in the language of Formal Concept Analysis (FCA). The goal here is to investigate whether such bases represent a suitable substitute for exact implication bases in practical use cases. To this end, we quantitatively examine the behavior of probably approximately correct implication bases on artificial and real-world data sets and compare their precision and recall with respect to their corresponding exact implication bases. Using a small example, we also provide evidence suggesting that implications from PAC bases can still represent meaningful knowledge from a given data set.

An alternate approach for efficiently computing the canonical basis is exploration from FCA. This procedure solves the problem for learning the Horn envelope of an arbitrary domain using an expert, or an oracle, capable of answering certain types of queries about this domain. However, the number of queries it may ask is exponential in the size of the resulting Horn formula in the worst case. We therefore extend our PAC approach. For this we revisit a well-known polynomial-time algorithm for learning Horn formulae with membership and equivalence queries and modify it to obtain a polynomial-time probably approximately correct algorithm for learning the Horn envelope of an arbitrary domain. Building on this, we introduce the notion of strong approximation, which reduces the number of queries to the oracle. Finally, we compare the results of the two approaches on artificial and real world data sets and show applications to collaborative settings.