Scalability, Search, and Sampling: From Smart Algorithms to Active Discovery

Stefan Wrobel

Otto-von-Guericke-Universität Magdeburg
School of Computer Science, IWS
Knowledge Discovery and Machine Learning Group
<http://kd.cs.uni-magdeburg.de>
P.O.Box 4120, Universitätsplatz 2
39016 Magdeburg, Germany
wrobel@iws.cs.uni-magdeburg.de

Abstract. The focus on *scalability* to very large datasets has been a distinguishing feature of the KDD endeavour right from the start of the area. In the present stage of its development, the field has begun to seriously approach the issue, and a number of different techniques for scaling up KDD algorithms have emerged. Traditionally, such techniques are concentrating on the *search* aspects of the problem, employing algorithmic techniques to avoid searching parts of the space or to speed up processing by exploiting properties of the underlying host systems. Such techniques guarantee perfect correctness of solutions, but can never reach sublinear complexity. In contrast, researchers have recently begun to take a fresh and principled look at stochastic *sampling* techniques which give only an approximate quality guarantee, but can make runtimes almost independent of the size of the database at hand. In the talk, we give an overview of both of these classes of approaches, focusing on individual examples from our own work for more detailed illustrations of how such techniques work. We briefly outline how *active learning* elements may enhance KDD approaches in the future.