1

Model Selection in Knowledge Discovery and Data Mining

Tu Bao Ho
Japan Advanced Institute of Science and Technology

Trong Dung Nguyen
Japan Advanced Institute of Science and Technology

Abstract. The process of knowledge discovery in databases inherently consists of several steps that are necessarily iterative and interactive. In each application, to go through this process, the user has to exploit different algorithms and their settings that usually yield different discovered models. The selection of appropriate discovered models or algorithms to achieve such models, referred to as model selection—which requires meta-knowledge on algorithms/models and model performance metrics—is generally a difficult task for the user. Taking this difficulty, into account, we consider that the ease of model selection is crucial in the success of real-life knowledge discovery activities. Different from most related work that aims at automatic model selection, in our view, model selection should be a semiautomatic work requiring an effective collaboration between the user and the discovery system. For such collaboration, our solution is to give the user the ability to try various alternatives easily and to compare competing models quantitatively by performance metrics and qualitatively by effective visualization. This chapter presents our research on model selection in the development of a knowledge discovery system called D2MS. The chapter first addresses the motivation of model selection in knowledge discovery and related work, an overview of D2MS, and its solution to model selection and visualization. It then presents the usefulness of D2MS model selection in two case studies of discovering medical knowledge from meningitis and stomach cancer data using three methods, i.e., decision trees, conceptual clustering, and rule induction.

0-8493-1121-7/03/$0.00+$1.50