Toward Economic Machine Learning and Utility-based Data Mining

Foster Provost
Stern School of Business
New York University
44 W. 4th St. New York, NY 10012

fprovost@stern.nyu.edu

ABSTRACT

Data mining requires certain information—for example, supervised learning requires training data. Some prior research has recognized that this information often does not simply present itself for free, but involves various acquisition costs. In addition, applying the learned models involves costs and benefits. I introduce a general economic setting that includes as special cases the settings of many different streams of prior research, such as cost-sensitive learning, traditional active learning, semi-supervised learning, active feature acquisition, progressive sampling, and budgeted learning, which are interwoven inextricably. For data mining in the general setting I suggest a strategy of maximum expected-utility data acquisition. Finally, I discuss how there are many open research issues that must be addressed. As a simple example, we must be able to deal with the seemingly straightforward problem of handling missing values in induction and inference.

See http://pages.stern.nyu.edu/~fprovost/ for more details (forthcoming).

UBDM '05 , August 21, 2005, Chicago, Illinois, USA.