SUCCESSFUL DATA MINING IN PRACTICE

Richard de Veaux, Williams College, Dept of Math and Stat, Bronfman Science Center, Hoxsey St., Williamstown MA, (413) 884-2809, deveaux@williams.edu

ABSTRACT

The sheer volume and complexity of data collected or available to most organizations has created an imposing barrier to its effective use. These challenges have propelled data mining to the forefront of making profitable and effective use of data. Data mining is a process that uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that may be used to make accurate predictions.

While the most widespread application of data mining are in CRM (customer relationship management) some of the other important applications include fraud detection and identifying good credit risks among a growing number of innovative applications.

After the data have been collected and prepared (which can take up most of the resources of a data mining project), the first and simplest analytical step in data mining is to **describe** the data — for example, summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look at the distribution of values of the fields in your data. Unfortunately, the standard exploratory data techniques of graphing and summarizing each variable take too long when dealing with hundreds of candidate predictors. Making scatter plots of each pair is even less feasible.

Thus, data description alone cannot provide an action plan. You must often build an exploratory predictive model based on the data, and then return to the graphical phase after using both domain based knowledge and exploratory models. In classical data analysis, the exploratory phase usually precedes the model selection phase. It's seen as a necessary preliminary for understanding the data before modeling it.

But in data mining, sometimes we start with a preliminary model just to narrow down the set of potential predictors. This exploratory data modeling (EDM) seems to be at odds with standard statistical practice, but, in fact, it's simply using models as a new exploratory tool.

In this workshop, we provide a brief tour of the current state of data mining algorithms and use several case studies to explain how EDM can be used to narrow the search for a predictive model and to increase the chances of producing useful and meaningful results.