KDD-99 Panel on Last 10 and Next 10 years

Ross Quinlan

RuleQuest Research

30 Athena Ave

St. Ives NSW 2075, Australia

quinlan@rulequest.com

Looking back at the position statements a decade ago, I can certainly identify some individual good and bad predictions. Larry Kershberg mentions the use of data mining "to produce highly selective marketing mailings" (a major current application); Pat Langley refers to "qualitative laws that relate different classes", anticipating association rules by some years; and I put my foot in it talking of "realistic databases, i.e. $\geq 10^5$ records" (a morsel by today's standards).

What is more damning, though, is our collective omission of many vital aspects of data mining as it is practiced today. Here are two examples:

- Getting the data ready for mining. Many authorities have noted that establishing a data warehouse and cleansing the data often account for most of the effort (and cost) of a data mining project. Ten years ago, we assumed that the data were no problem -- all we had to worry about were the algorithms for analysing it.

- The size of the data mining industry. We (or at least I) didn't anticipate the level of activity now (or the number of people who would be drawn to KDD conferences). This scale is important because it enables specialization in vertical markets, such as the top-end products targeting CRM and electronic commerce, whereas we thought more in terms of generic (horizontal) approaches to shared problems.

Sadly, some of the problems that the KDD-1989 Panel identified are still with us. We have yet to develop techniques for mining relational data that scale well enough to be useful -- we often have to construct those pernicious "flat files" via aggregation and summarization, not because we like throwing away information, but because our KDD systems force us to. We still lack strong methods for understanding and explaining time-varying data, even though the underlying forces that generate transaction data (for instance) are clearly not static.

I'm reluctant to make predictions that Gregory can hang around my neck (like the proverbial albatross) at some future meeting. However, here are a couple of questions that the audience might like to address:

- Will data mining continue to be carried out mainly in the corporate and scientific worlds, or will it migrate to the consumer level?

- We are all aware of the information resource provided by the WWW; a tool that would assist formulation of knowledge from it (such as summaries, alerts, comparative reviews) would have a great potential market.

- Will KDD research move further away from its roots in AI and ML, seeking closer ties with (say) statistics and database? This would be a pity, in my view, because I would like to see data mining produce more descriptive outputs rather than "just" qualitative models.

About the author:

J. Ross Quinlan is an Adjunct Professor at The University of New South Wales, Australia, and Director of RuleQuest Research, a small data-mining software company.

He is author of many important publications in Machine Learning and related areas, and the developer of the landmark C4.5 family of decision-tree classifiers.