

The rise of data mining: opportunity or threat?

Andrew Kerkes and Scott MacLean discuss whether the rise of data mining is an opportunity or threat.

Andrew Kerekes and Scott MacLean argue that data mining may yet be the unsung hero of analysis, with possibilities still unrecognised by many in the research industry – and many of the key software applications have been invented in the Asia Pacific region by Australian and Kiwi developers.

What is data mining? Data mining has been around in a business setting at least since the early 1960s when Fair and Isaac started applying statistical analysis to understand how credit risk could be predicted based on retrospective analysis of 'application form' and behavioural data. The migration of data mining techniques into other 'quantitative' domains of business was inevitable and practitioners in market research were soon amongst those applying more advanced analysis to extract insight from large data sets. More recently, data mining has come to mean the automated detection of relevant patterns in a database to predict customer behaviour.

The rise of data mining software has been surrounded by hyperbole with many vendors proclaiming that advanced analytic techniques would soon be available to any business user. The demise of the analyst seemed apparent... Fortunately (for analysts) human knowledge, judgement, and yes, even intuition, remain critical determinants in working out the differences between relevant and irrelevant patterns in data, even those produced by newer analytic 'data mining' techniques. The converse happened; business needed more analysts if they were to derive value from their data mining efforts.

Marketing and research have always been connected with helping businesses increase the 'lifetime value' of their customers. Consumer insight has been used to attract more customers, retain existing customers for longer, achieve higher market share, and to target advertising and promotions, all to increase the revenue and profit that businesses can achieve from customers.

Clients increasingly expect market research providers to collaborate in marketing consultancy as well as provide research capabilities. There is a clear connection between the business and marketing competencies that research consultants have and their ability to scope research that delivers quality insights. It is imperative that market research consultants have what might be called 'domain' knowledge; knowledge that includes understanding clients' businesses, the markets they operate in, as well as the knowledge and techniques associated with collecting and analysing data to provide meaningful customer insights.

Researchers with comprehensive domain knowledge are well-placed to become the trusted partners of clients and will be integral to driving business success. Consultants need also to understand the importance of a timely connection between insight and action and how this can be fundamental to achieving success in a rapidly changing and evolving market. Data mining accelerates the potential

delivery of insights and, through the integration with operational platforms, the acceleration of tactical deployment.

The requirements to achieve business benefits from data mining are, however, in many ways no different to those required to achieve benefits from more traditional consumer research. The need for marketing 'domain' knowledge, consulting and analysis skills, and the ability to carefully interpret and judge the results of analysis remains, independent of the source data or analytical techniques applied. To us, data mining (and machine learning), carefully applied, may yet be the unsung hero of analysis, with possibilities still unrecognised by many in the research industry. The competencies that have been developed by market researchers are equally useful in both 'independent' survey 'data sets' and marketing databases (i.e. usually within 'data warehouses' that contain 'transaction' data). The next Holy Grail will be to integrate the two types of data - 'survey' and 'transaction' data - and analysts who understand both domains can expect to become increasingly sought-after.

The tools

There is a plethora of data mining software routines out there, many of them relatively low-cost, and some of the best having been developed by Australian and/or New Zealand analysts and academics, such as See5 by Sydney-based Ross Quinlan and Weka, from the University of Waikato in New Zealand. In addition, the R package (again, originally developed in NZ) now has hundreds of contributed routines and modules from around the world, many of them dealing with data mining or similar applications (e.g. market basket analysis).

The Weka suite, in particular, is one of our favourites, because once you get past the somewhat different terminology that machine learning specialists utilise (e.g. 'instance' instead of 'case' or 'respondent'), and the initially confusing means of selecting predictors and specifying algorithms, the power of these approaches becomes readily apparent. See5 is also a sterling performer, in terms of being able to rapidly develop complex sets of decision rules for the allocation of 'instances' into pre-defined categories. (A more elementary version of See5 is also contained in the Weka suite, along with dozens of other routines).

Machine learning (largely) eschews the niceties of statistical assumptions, and essentially lets you work with what you have, often with an astounding degree of success. In other words, predictive relationships can be developed when you might have thought that none would exist, and where no amount of clever cross-tab work will help you, or where the dreaded issue of multi-collinearity between predictors will trip you up.

And even missing values have a role to play – instead of substituting with a mean/median/'whatever', data mining approaches can permit us to treat missing values as being on par with valid values.

Sounds too good to be true? Well, yes, it certainly can be. And as with all data analysis, the results are only as good as your data file, and it is even more important (if that can be possible) to spend the time cleaning your original file, and understanding just what you have in there, before you leap into the clever stuff.

But once you do, with a bit of effort (often quite a lot of effort) you can develop an astounding level of knowledge about your data and, hence, about your respondents or customers.

Here is the basis of the many of these approaches:

- Each case in the dataset is assumed to belong to one of a small number of mutually exclusive classes
- Properties of every case that may be relevant to its class are provided, although some cases may have unknown (i.e. 'missing') or non-applicable values for some attributes
- The software can deal with any number of attributes and is therefore suited to large databases containing potentially 100s (or even 1000s) of predictors
- The data mining task is to find how to predict a case's class from the values of the other attributes
- This is done by constructing a classifier that makes this prediction, expressed as decision trees or, alternatively, as sets of rules.

For example, a case study published on www.rulequest.com uses data concerning telecoms churn. These data were artificial but were said to be 'based on claims similar to real world'. Each case is described by just 16 numeric and three nominal attributes. The data were divided into a training set of 3,333 cases and a separate test set containing 1,667 cases. A 19-rule classifier was found from the training data, seven rules for no churn and 12 for churn, which had an accuracy of 95 per cent on the unseen test cases.

To take just a few more examples, here are some successful data mining investigations we have been involved in:

- In a business to business services marketing context, developing a single decision tree to correctly allocate respondents to segments with a better than 90 per cent accuracy rate
- Identification of key drivers of interest in buying a new health product, via an algorithm applied to a stacked file of over 12,000 cases, which precisely reflected findings derived from logit modelling conducted in parallel.

The leap for market research professionals will be to increasingly understand how insights are more directly linked to business processes that might be embedded in customer relationship management, campaign management or even supply chain functions. The number of analytic techniques that can be applied to data is growing (as they always have), the source of the data used for analysis is growing, and the need for integrated data and integrated insight is growing.

Consumer research professionals have a growing repertoire that their clients will be asking them to draw insights from; the more innovative research providers are already positioning themselves for the changes that are taking place by drawing on the considerable research and analysis skills their teams have developed and bringing in skills from the operational marketing sphere. With increasing competition in most consumer and business-to-business markets, more products tailored to smaller customer segments, and more tactical capability, the future looks bright for the right research consultants.