Beyond Statistical Procedures for Predictive Modelling - Data Mining Algorithms and Support for University Research at QUT

Ray Duplock and Neil Kelson

High Performance Computing and Research Support (HPC&RS), Information Technology Services, Queensland University of Technology
Contact: r.duplock@qut.edu.au

Introduction – An Urgent Need for Change

In a seminal data mining article, Leo Breiman (2001) argues that to develop effective predictive classification and regression models, we need to move away from the sole dependency on statistical algorithms and embrace a wider toolkit of modeling algorithms that include data mining procedures. Currently there are two separate modeling cultures - a data modeling culture (statistical culture) and an algorithmic modeling culture (data mining culture). Sadly, many statisticians and quantitative data analysts, including researchers, rely solely on statistical procedures when undertaking data modeling tasks; the sole reliance on these procedures has lead to the development of irrelevant theory, questionably conclusions, and have kept statisticians from working on a large range of interesting current problems in both the research and the commercial domains (Breiman, 2001, p.199).

Data Mining algorithms originated from the work of industrial statisticians, computer scientists, engineers, biochemists, biologists, psychologists and neuroscientists. Researchers from these domains have been developing these algorithms for the last 30 years in response to intractable analysis problems that have proved impossible to solve using traditional statistical routines. Data Mining algorithms have proved to be fruitful in fitting models to complex data sets, particularly in the areas of speech recognition, image recognition, nonlinear time series predictions, handwriting recognition, and the analysis of large datasets that display the "curse of dimensionality" – a growing problem in many areas that amass large amounts of data such as astronomy and astrophysics, genetic and biochemistry studies, and chemical and scientific instrumental data. Data Mining algorithms are entrenched in commercial organisations engaged in predictive modeling because of the greater predictive accuracy these algorithms offer over statistical procedures (e.g., a CART model is between 10-15% more accurate than a linear regression model and a CART model will usually outcompete a logistic regression procedure in classification analysis; Steinberg and Colla, 1997, p.12). Data Mining algorithms also offer a number of more practical benefits over statistical procedures: they are easier to use, they are distribution free (nonparametric), they are largely assumption free, they handle nonlinear relationships more effectively , can handle many predictive variables (up to several thousand at a time) and are more robust to outliers and missing data.

So with all these advantages, why have our universities got an almost universal myopic reliance on statistical procedures – even the use of statistical procedures that are two and three generations old are preferred? Why have we been so reluctant and slow to adopt these new innovations that have offered much in the commercial domain and in some limited academic research areas? This poster will look at some innovation impediments within our organisation and across the university sector.

Types of Data Mining Routines, Data Mining and Modeling Myths

Data Ming algorithms include such things as Neural Network models; Classification and Regression models (CART), Tree Ensemble methods such as TreeNet and Random Forests, Genetic Algorithms, Support Vector Machine algorithms, Nonlinear Spline modeling algorithms and Swarm algorithms. Although data mining algorithms started to be developed in the early 1970s, the two first useful and successful algorithms, CART and Multilayer Perceptron networks, were developed in the early 1980s – about 30 years ago! Many myths, fallacies and some gross misunderstandings have grown up about data mining, data modeling and data issues that have contributed to hinder the uptake of these procedures. Several of these myths include:

- Data Mining algorithms are only useful for large datasets
- Data Mining algorithms are only useful for exploratory data analysis
- Data Mining algorithms develop models that can easily overfit your data and as a result are undesirable
- Occam's Razor implies that a simple model is better than a more complex model and as a result a simple model will apply better to unseen data (favoring statistical models)
- Dimensionality is also a 'curse' and you better reduce your variable set to a few 'important' predictor variable else your model will be ineffective
- When you analyse data there is always one 'correct' model – your generated statistical or data mining model is 'correct'

Although space prohibits an extensive discussion of each of these points, briefly:

- Data Mining algorithms ***can be used successfully on small datasets and large datasets*** (Breiman 2001, p.199)
- Data Mining algorithms are very useful in theory testing and theory construction (Marshall and English 2000; Jaccard and Jacoby, 2010, p.309-311; Read et.al., 2010)
- Data Mining techniques have complex tests to ensure the model does not overfit data. ***The application of cross-validation in data mining procedures ensures the resultant models do not overfit large or small datasets*** (Breiman et.al., 1984, p.75-81; Hastie et.al., 2009, p.241-245)
- Unfortunately the 14th century Occam's razor myth is slow to die. Simple models seldom fit complex datasets – ***complex natural processes need complex models to explain their outcomes and to make useful predictions*** (Webb, 1996; Breiman, 2001, p.206-207)
- Relying on some statistical procedures that are 100+ years old will display many 'curses', some are disguised as data 'curses'. For years we have been told that a statistical model is only reliable if it has around 5-7 independent variables (or predictor variables) and if you have more than this, you need to engage in an appropriate data reduction procedure such as a principle component analysis or common factor analysis procedures (this advise is indeed correct). The sad reality is dimensionality is a curse since statistical modeling algorithms cannot handle more than about 5-7 variables. ***Data Mining algorithms can handle 100s, 1000s of predictor – there is no need for data reduction. Reducing dimensionality reduces the amount of important variable information you have particularly in important variable combinations and interactions*** (Breiman 2001, p.208).
- There is seldom one correct model. Applying a multiple of data mining routines to a dataset will generate a multiple of models , some of which will be similar in predictive accuracy, but each may be telling a different story in terms of variables of importance, model characteristics and variable interactions. A variety of statistical models can also be generated on the same dataset, again with similar accuracy and model characteristics. ***There is seldom one correct statistical or data mining model for a dataset.***

Beside these ill founded myths and beliefs that have beset data mining, a number of other key factors have restricted the diffusion of these innovations with our university. These issues are briefly described below.

Top Down Innovation Requests, Data Mining Software Access and Seminars and Workshops about these Innovations

Other important issues restricting the diffusion of these innovations are:

- Most researchers are not aware of these procedures and their primary advisors in their area (local statisticians) have a vested interest to maintain a focus in statistical routines . However, given the multitude of research articles that have used data mining procedures in all academic domains and in a wide range of academic journals – these undoubtedly would/should be read by this audience and one wonders why their curiosity and focus has not been heightened. At least you would expect they would be aware of these procedures and have a vague idea of their benefits and uses? But discussions with many researchers over the last 20 years has yielded a surprising level of ignorance in this area?
- Acquiring funding to purchase good data mining software is more than a Schwarzeneggerian feat within a university environment. Without a lot of support from key academic researchers, this is an exercise in futility. And it's impossible to get this level of support for a new innovation – a difficult chicken and egg scenario.
- Good data mining software is increasingly proprietary and it's expensive. We can be beguiled by some good free data mining software available in Weka and R, although these have limitations. Salford Systems data mining suit, SAS Enterprise Miner, Matlab data mining features are not cheap – they are expensive tools. However, universities need to realise that to do good research is not an inexpensive enterprise, and the costs of data mining software to guide and support research is an expensive but necessary outlay
- Data mining companies can and must do more to support the diffusion of their products into our sector. Some do, but a lot more can and needs to be done. For these companies to lament that these innovations are slow to diffuse into our sector, they need to look no further than their own university pricing structures -- making realistic improvements in this area will go a long way to alleviating this condition
- The HPC&RS group will be running a series of innovative and original data mining seminars and workshop starting in October this year. Three-3 hour seminars and three-3 hour workshops covering data mining classification and regression procedures will be run free of charge for our researchers and postgraduate students. To our knowledge these seminars and workshops are the first of their kind and will show *how data mining algorithms can be used in lieu of statistical procedures or to complement statistical procedures for classification and regression modeling tasks, and data clustering procedures; for theory construction or theory testing ; and in small and large datasets*. And it is with much pleasure that we would like to advise that after personal discussions with the Salford Systems' US account manager, that not even top US universities are trying to engage their researchers in this way with data mining initiatives – an innovative initiative even from their perspective. It's great to see an Australian University taking the lead in this innovative area!

Data Mining and statistical procedures are required to build and test substantial research theory. Better theory benefits all humanity: these innovations are too important to ignore!

References

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Pacific Grove: Wadsworth.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*, 199-215.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer: New York.

Jaccard, J., & Jacoby, J. (2010). *Theory Construction and Model-Building Skills: A Practical Guide for Social Scientists.* New York: Guilford Press.

Read, S.J., Monroe, B.M., Brownstein, A.L., Yang, Y., Chopra, G., & Miller, L.C. (2010). A Neural Network Model of the Structure and Dynamics of Human Personality. *Psychological Review*, *117(1)*, 61-92

Marshall, D. B., & English, D. J. (2000). Neural Network Modeling of Risk Assessment in Child Protective Services. *Psychological Methods*, *5(1),*102-124.

Steinberg, D., & Colla, P. (1997). *CART -- Classification and Regression Trees.* San Diego, CA: Salford Systems.

Webb, G.I. (1996). Further Experimental Evidence against the Utility of Occam's Razor. *Journal of Artificial Intelligence Research*, *4*, 397-417.