Incorporating cost in Bayesian Variable Selection, with application to cost-effective measurement of quality of health care

D. Fouskakis,
Department of Mathematics, National Technical University of Athens, Greece;
e-mail: fouskakis@math.ntua.gr

I. Ntzoufras
Department of Statistics, Athens University of Economics and Business, Greece;
e-mail: ntzoufras@aueb.gr

D. Draper
Department of Applied Mathematics and Statistics, University of California, Santa Cruz, USA;
e-mail: draper@ams.ucsc.edu

Summary

In the field of quality of health care measurement, one approach to assessing patient sickness at admission involves a logistic regression of mortality within 30 days of admission on a fairly large number of sickness indicators (on the order of 100) to construct a sickness scale, employing classical variable selection methods to find an "optimal" subset of 10–20 indicators. Such "benefit-only" methods ignore the considerable differences among the sickness indicators in cost of data collection, an issue that is crucial when admission sickness is used to drive programs (now implemented or under consideration in several countries, including the U.S. and U.K.) that attempt to identify substandard hospitals by comparing observed and expected mortality rates (given admission sickness). When both data-collection cost and accuracy of prediction of 30-day mortality are considered, a large variable-selection problem arises in which costly variables that do not predict well enough should be omitted from the final scale.

In this work we develop a method for solving this problem based on posterior model odds, arising from a prior distribution that accounts for the cost of each variable and results in a set of posterior model probabilities which corresponds to a generalized cost-adjusted version of the Bayesian information criterion (BIC). We use reversible-jump Markov chain Monte Carlo (RJMCMC) methods to search the model space. The proposed method provides a principled approach to performing a cost-benefit trade-off that avoids ambiguities in identification of an appropriate utility structure. Our cost-benefit approach results in a set of models with a noticeable reduction in cost and dimensionality, and only a minor decrease in predictive performance, when compared with models arising from benefit-only analyses.

Additionally to the above, we present a population based reversible jump MCMC algorithm which efficiently searches the model space when an overall limit on the total data collection cost of each subset is enforced. The method is presented and compared with simpler MCMC schemes.

All results are phrased in the language of health policy but apply with equal force to other quality assessment settings with dichotomous outcomes, such as the the study of retention rates in the workplace and the creation of cost-effective credit scores in business.

The presentation is available at `http://stat-athens.aueb.gr/~jbn/current.htm`.

Related publications

1. Fouskakis, D., Ntzoufras, I. and Draper, I. (2009). Bayesian variable selection using a cost-penalised approach, with application to cost-effective measurement of quality of health care. *Annals of Applied Statistics* **3**, 663–690.

2. Fouskakis, D., Ntzoufras, I. and Draper, I. (2009). Population-Based Reversible Jump MCMC for Bayesian Variable Selection and Evaluation Under Cost-Limit Restrictions. *Journal of the Royal Statistical Society C (Applied Statistics)* **58**, 383–403.