Expert Systems with Applications 37 (2010) 7656–7662

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Bayesian variable selection for binary response models and direct marketing forecasting

Geng Cui [a,*], Man Leung Wong [b], Guichang Zhang [c]

[a] Department of Marketing and International Business, Lingnan University, Tuen Mun, N.T., Hong Kong
[b] Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, N.T., Hong Kong
[c] Department of Economics, Ocean University of China, Qingdao, Shandong 266071, PR China

ARTICLE INFO

Keywords:
Bayesian variable selection
Binary response models
Distribution of priors
Direct marketing
Forecasting models

ABSTRACT

Selecting good variables to build forecasting models is a major challenge for direct marketing given the increasing amount and variety of data. This study adopts the Bayesian variable selection (BVS) using informative priors to select variables for binary response models and forecasting for direct marketing. The variable sets by forward selection and BVS are applied to logistic regression and Bayesian networks. The results of validation using a holdout dataset and the entire dataset suggest that BVS improves the performance of the logistic regression model over the forward selection and full variable sets while Bayesian networks achieve better results using BVS. Thus, Bayesian variable selection can help to select variables and build accurate models using innovative forecasting methods.

1. Introduction

The key objective for direct marketing forecasting is to identify potential customers from an existing database so that marketers can design accurate targeted marketing to increase sales and profitability. Meanwhile, today's businesses are capable of generating and collecting a huge amount of customer and transactional data in a relatively short period. Data reduction, and more specifically variable selection, is a major challenge in database marketing (Rossi & Allenby, 2003). Traditional methods of stepwise variable selection do not consider the interrelations among variables and may not identify the best subset for model building. Researchers need a more efficient method of variable selection to build accurate forecasting models and to take advantage of innovative modelling methods that have become increasingly viable.

Recently, the Bayesian method has been proposed as a semi-automatic method for variable selection and provides a feasible solution for exhaustive search (George, 2000). In comparison with the conventional statistical methods, Bayesian variable selection (BVS) is more beneficial for forecasting methods that are apt in handing nonlinearity and interactions among variables. However, how to execute efficient BVS remains a significant challenge. Moreover, whether the Bayesian approach to automatic variable selection can improve the accuracy of forecasting with real data warrant investigation (Rossi & Allenby, 2003). This study proposes Bayesian variable selection using informative priors to select variables to build direct marketing forecasting models. For computing analytically tractable priors and posterior model probabilities, we adopt the efficient algorithms of Chen, Ibrahim, and Yiannoutsos (1999) that require Gibbs samples from a single model. We first perform variable selection using both forward selection method and BVS. Then, we test the effect of the selected subsets on the forecast accuracy of both logistic regression and Bayesian networks on a holdout dataset and the entire dataset. The results of validation suggest that BVS improves the accuracy of forecast of logistic regression over the forward selection and the full variable sets. Bayesian networks, a model of joint probability distribution, achieve better results with the BVS set. These findings have meaningful implications for selecting variables to build forecasting models and direct marketing.

2. Direct marketing forecasting models

The primary objective of consumer response modelling in direct marketing is to identify those customers who are the most likely to respond. Researchers have developed many direct marketing response models using consumer data. One of the classic models, known as the *RFM* model, estimates the likelihood of consumer purchases from a direct marketing promotion using three variables: (1) *recency* of the last purchase, (2) the *frequency* of purchases over the past years, and (3) *money* or the monetary value of a customer's purchase history (Berger & Magliozzi, 1992). Models that include other variables such as consumer demographics and psychographics, credit histories, and purchase patterns can

* Corresponding author. Tel.: +852 2616 8245; fax: +852 2467 3049.
E-mail addresses: gcui@ln.edu.hk (G. Cui), mlwong@ln.edu.hk (M.L. Wong), zgc1976@ouc.edu.cn (G. Zhang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.04.077

help improve the accuracy of prediction and understanding of consumer responses. Statistical methods such as logistic regression and discriminant analysis have been popular tools in modelling consumer responses to direct marketing. Recently, researchers have developed other sophisticated methods such as beta-logistic models, tree-generating techniques, i.e., CART and CHAID, and the hierarchical Bayes model. Machine learning methods such as neural networks (ANNs) and Bayesian networks have also been applied to modelling consumer responses (Baesens, Viaene, van den Poel, Vanthienen, & Dedene, 2002; Cui, Wong, & Lui, 2006; Zahavi & Levin, 1997).

In the information age of today, the explosive growth of data represents one of the most significant challenges facing marketing researchers and managers, especially for data mining with large noisy databases. The capability of computer technologies and the Internet to collect and store data about consumers has far exceeded the ability of analysts to process them into usable, value-added information. Despite the improvements in modelling methods, variable selection remains one of the challenges for building forecasting models in direct marketing. How to distinguish relevant variables from noises have significant implications for building accurate forecasting models. Conventional methods of variable selection such as the stepwise approach may not select the best subset of variables (Miller, 1990). BVS can perform exhaustive search and provide a better subset of variables for model building using innovative methods to improve forecasting accuracy.

3. Variable selection

Much of the debate in marketing science involves the issue of variable selection (Punj & Stewart, 1983). Variable selection is an important issue because even one or two irrelevant variables may affect the performance of an otherwise viable model. The rationale for selecting variables may be based on an explicit theory or commonly agreed relevant dimensions, for instance using cluster analysis in market segmentation. Selecting variables on a theoretic basis is usually preferred. However, for most direct marketing models, there is often not sufficient theoretical guidance in selecting variables. Thus, variable selection is one of the most frequently encountered problems in direct marketing (Blattberg & Dolan, 1981).

The explosive growth of data represents one of the most significant challenges facing marketing researchers and managers in the information age. Today, researchers often have more variables than they need to build a good model, making variable selection an urgent issue in marketing research. Although many methods of variable selection exist for classification problems such as to predict whether a consumer will respond to a specific direct marketing promotion, researchers typically adopt a semi-parametric model such as logistic regression as a starting point. In the following sections, we focus on the methods of variable selection for the commonly used logistic regression model.

In general, exhaustive search is the only technique that can ensure finding the predictor variable subset with the best evaluation criterion. However, it is only a feasible technique when the number of predictor variables is less than 20. For more than 20 variables, exhaustive search methods may become computationally intractable. For a model with 25 predictor variables, for instance, exhaustive search must examine 33,554,431 subsets, i.e., all the possible combinations, and this number doubles for each additional predictor variable considered (Rogue Wave Software, 2009). Clearly, exhaustive search using the conventional methods is not always practical. Researchers typically settle for some other selection techniques as a compromise.

Most variable selection methods are based on evaluating the relationships between the dependent variable and the predictor variables. Variable selection for the class of binary classification or response models includes forward, backward and stepwise selections. Methods such as logistic regression apply the maximum likelihood estimation method after transforming the dependent into a logit variable. In this way, logistic regression estimates the probability of a certain event occurring. Forward and backward selection procedures are simple methods for variable section in logistic regression. In each case, the log-likelihood is tested for the model when a given variable is added to or dropped from the equation.

3.1. Forward, backward and stepwise selection

The forward method of variable selection starts with a null model or an empty set. Some preprocessing may be performed so that the predictor variables become nearly statistically independent. Variable selection using logistic regression uses a certain p-value as the entry criterion for any variables to be included. The usual criterion or the default value in most statistical software is the 0.05 significance level. Forward selection keeps on adding predictor variables but never deletes them, thus this technique is always computationally tractable. Forward selection may not find the subset with the highest evaluation criterion if predictor variables are not statistically independent or the model is not a linear combination of predictor variables. Although many researchers have reported good results with forward selection (Miller, 1990), this method is not guaranteed to find the subset with the highest evaluation criterion as it only compare a limited number of subsets.

Backward selection is similar to forward selection in computational properties but it compares more subsets. The starting subset in backward selection includes all the predictor variables, which are then deleted one at a time as long as this results in a subset with a higher evaluation criterion. In this case, starting the search with all the predictor variables helps taking interrelations among predictor variables into account. A major disadvantage of backward selection, however, is that one's confidence in the criterion values for subset evaluation tends to be lower than that for forward selection (Shtatland et al., 2000). This is especially true with small datasets. When the number of cases is close to the number of predictor variables, forward selection is the preferred option. Like forward selection, backward selection does not perform exhaustive search and may not find the subset with the highest evaluation criterion.

The stepwise procedure of variable selection combines the advantages of forward and backward selection. Stepwise selection usually starts with an empty set. A predictor variable may be added or dropped at any point in the search process. Thus, stepwise selection evaluates more subsets and tends to produce better subsets than the other two techniques, because the stepwise procedure may add a new variable that meets the criterion but also examines all other variables already included and excludes any variables that do not meet the criterion (Miller, 1990). In this sense, the stepwise procedure is a significant improvement over the simple forward or backward selection method. However, increased computing intensity is the price to pay for stepwise selection to find better subsets. Moreover, logistic regression may overestimate a variable's predictive power. To minimize this problem, researchers sometimes may apply more stringent criteria such as the significance level of 0.02, so that they can compare the alternative subsets in terms of their stability and performance.

While these methods are efficient and frequently used, they have several drawbacks (Miller, 1990). First, they suffer from the random variations in the data and may produce results that tend to be idiosyncratic and difficult to replicate. Consequently, they