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

may generate biased regression coefficients and lead to overly optimistic estimates (Austin & Tu, 2004). Secondly, the choice of value for entry and exit for variables (e.g., 0.05) is perhaps the most controversial aspect of stepwise regression for variable selection. Depending on the sample size and the specific value, it may include either too many variables for meaningful interpretation or too few variables to build a reliable model. Despite the criticism of the arbitrariness of this method, stepwise logistic regression has been widely used due to the lack of better alternatives (Blattberg & Dolan, 1981).

To reduce the sensitivity of variable selection procedures to sample variations, researchers have proposed other selection criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The purpose for these information criteria is to penalize model complexity – the number of predictor variables used in the equation. The AIC criterion is asymptotically equivalent to the cross-validation criterion. For the BIC, the probability of choosing an incorrect model or subset approaches zero as the sample size increases. In this sense, AIC and BIC are equivalent to the Bayesian method. Under some conditions they can serve as a bridge between the frequentist approach and the Bayesian method. However, both AIC and BIC originate from a purely frequentist framework. They can only emulate the Bayesian approach when the priors are as important as the likelihood (i.e., the new data) or when the priors are of little importance (Shtatland et al., 2000).

3.2. Bayesian variable selection

Conventional methods that select variables by evaluating the relationships between the dependent variable and the predictor variables are mostly based on a conditional distribution model that does not consider the interrelations among the predictors, although interactions can be added manually. In recent years, the Bayesian method of automatic variable selection has been proposed for regression models and has been the subject of substantial research (George, 2000). The Bayesian approach to variable selection is straightforward in principle. The posterior probability distribution is the product of the prior distribution placed on the model and the likelihood, i.e., parameters based on the data. Thus, one needs first to identify a distribution of priors for the model. Following that, one can quantify the prior uncertainties via probabilities for each model under consideration, specify a prior distribution for each of the parameters in each model, and then use the Bayes theorem to generate the posterior model probabilities that are proportional to the product of the prior probability and the likelihood. However, Bayesian variable selection is difficult to carry out because of the difficulties in specifying the prior distributions for the regression parameters for all possible models and specifying a prior distribution on the model space, and the computational burden associated with these processes (Cheng et al., 2002). Fortunately, fully Bayesian approaches to variable selection are now feasible due to recent advances in computing technology and the development of efficient computing algorithms (Cui & George, 2008; Rossi & Allenby, 2003). Some of the proposed BVS procedures include those proposed by George and McCulloch (1993, 1997) and George, McCulloch, and Tsay (1995).

Similar to other methods, Bayesian method examines the effect of variable selection on model performance either by a holdout dataset or by measures such as the AIC. In this case, the optimum subset is the one that has the largest posterior probability out of all the available subsets. The biggest advantage of Bayesian variable selection is that it provides a feasible method for exhaustive search. However, the existing methods of Bayesian variable selection have been designed for linear additive models. Previous methods took the priors for the regression coefficients to be normal scale mixtures, which are well suited for very large problems and are mainly designed for tuning the convergence of the Gibbs sampler (George et al., 1995). Using reference or uniform priors giving equal probabilities to all models may not provide satisfactory solutions in many situations. In some cases, elicitation of informative priors or the quantification of real prior information may provide better solutions. However, most Bayesian variable selection methods have been developed for linear regression models. For direct marketing forecasting, Bayesian variable selection needs to be adapted for binary classification methods such as the semi-parametric logistic regression model and other innovative methods that can handle interrelations among variables more efficiently in order to realize its full potential for improving classification accuracy.

4. Bayesian variable selection for direct marketing

This study applies Bayesian variable selection using informative priors for direct marketing forecast. To model consumer responses to direct marketing, the problem boils down to binary classification, i.e., to predict whether a consumer will buy or not buy. To apply BVS to a binary response model such as logistic regression, researchers need to accomplish three tasks: (1) specify the prior distributions for the regression parameters for all possible models, (2) specify a prior distribution on the model space, and (3) compute the marginal and posterior probabilities. To solve these problems, we adopt the algorithms proposed by Chen et al. (1999), who have addressed these issues for the logistic regression model and derived the theoretical and computational properties of the priors. The computational algorithms only require Gibbs samples from the full model to facilitate the computation of the priors and posterior model probabilities for all possible subsets.

4.1. Specification of the priors

In this study, we rely on informative priors to aid the Bayesian variable selection process. Specifying meaningful prior distributions for the parameters in each model is difficult, because it requires contextual interpretations of a large number of parameters. A need arises then to identify a method that derives useful automated specifications. To solve this problem for a linear model, Chen et al. (1999) recommend specifying a prior prediction y_0 for the response vector, and a scalar a_0 that quantifies one's assignment of information contributing to this estimate relative to the information to be collected in the experiment. Then, y_0 and a_0, along with the design matrix for model m, are used as prior information to specify an automated parametric informative prior for the regression coefficients $\beta^{(m)}$. The motivation behind this approach is that investigators often have prior information from similar past studies measuring the same response variable and covariates as the current study.

Since direct marketers often have access to historical data of a company's promotion campaigns and customer purchases, researchers should be able to generate useful informative priors for variable selection to build forecasting models using the Bayesian approach. First, direct marketers regularly send solicitations to potential consumers, and they collect similar data year after year. Such data are naturally re-occurring and are updated on a regular basis. Second, direct marketers often use historical data and similar variables to build their customer selection models and to forecast future sales. Researchers have used these variables many times and have substantial knowledge of these variables. The effects of these variables do not change significantly over time. Thus, informative priors can provide real benefits for variable selection in direct marketing if such priors can be elicited from previous data.

4.2. Distribution of the informative priors

The informative prior distribution is based on the idea of the existence of a previous study that measures the same response variable and covariates as the current study. Using the covariance structure of a previous study, one can construct a prior distribution for the parameters of the current study. As proposed by Chen et al.'s (1999) in Eq. (1), the natural choice for X_{01} is the raw covariance matrix from a previous study. In this case, the covariance matrix from a previous study provides the basis for eliciting the informative priors for the current study. The maximum likelihood estimation (MLE) is used to generate the covariance matrix. Then, the diagonal of the inverse Hessian matrix are used for eliciting the priors.

$$\pi\left(\beta^{(m)}\middle|D_0^{(m)},a_0\right)\alpha\exp\left\{a_0\left(y_0'X_0^{(m)}\beta^{(m)}-J_0'Q_0^{(m)}\right)\right\}\pi\left(\beta^{(m)}|C_0^{(m)}\right) \quad (1)$$

For the ease of explanation, let us assume only one previous study, as the extension to multiple previous studies is straightforward. Given a precision matrix, c_0 is a fixed hyper-parameter that controls the impact of $\pi_0(\beta^{(m)}|c_0)$ on the entire prior. Then, a_0 is a scalar prior precision parameter that weighs the historical data relative to the likelihood of the current study. Small values of a_0 give little prior weight to the historical data relative to the likelihood of the current study, whereas values of a_0 close to 1 give roughly equal weight to the prior and the likelihood of the current study. Specifically, the case of $a_0 = 1$ corresponds to the formal Bayesian update of $\pi_0(\beta^{(m)}|c_0)$ using the Bayes theorem. With $a_0 = 1$, the prior and likelihood of the current study are equally weighted. The case of $a_0 = 0$ results in no incorporation of historical information. In this case, the prior reduces to the initial prior. Thus, the parameter a_0 allows the investigator to control the influence of the historical information on the current study. Such control is important in cases where there is heterogeneity between the historical data and the current study, or when the sample sizes of the two studies are different. In actual analysis, researchers have choices that give small and large weights to the historical data to conduct sensitivity analyses to achieve optimum results.

4.3. Posterior probabilities

Using the algorithms proposed by Chen et al. (1999), we can incorporate real prior information and the quantification of prior information from past studies for variable selection. With the proposed informative priors, the computational methods are well suited to handle a moderate number of covariates, such as $k < 20$. To compute the marginal and posterior distribution of the data, we adopt the Monte Carlo approach to estimate all the prior model probabilities by using a single Gibbs sample from the full model and computing the marginal distribution of the data via ratios of normalizing constants. The method produces prior and posterior probabilities for all possible models, which serves as an effective method of exhaustive search for variable selection.

Although Bayesian variable selection has been proposed for classification problems, it has been tested only with conventional statistical methods like logistic regression. Traditional methods like the logit model, which assume a model of conditional distribution, cannot fully take advantage of the benefits provided by Bayesian variable section. Models of joint distribution that strive on interactions and nonlinearity can derive greater benefits from BVS. In this study, we compare the effect of Bayesian variable selection on logistic regression with that on Bayesian networks, a model of joint probability distribution that is especially apt for handling interrelations among predictor variables (Cui et al., 2006).

5. Methodology

The data for this study was provided by a US-based catalog direct marketing company that sells multiple product lines of general merchandise ranging from gifts and fashion to electronics. The company regularly sends mailings to its customers. This particular database from the Direct Marketing Education Foundation stores the records of 106,284 consumer responses to a recent promotion as well as their purchase history and responses to the promotions from the company in the last four years. Each customer record contains 361 variables. This study focuses on the most recent catalog promotion with a 5.4% response rate, representing 5740 buyers.

To examine the effect of variable selection methods on model performance, we compare the conventional statistical methods of variable selection (forward, backward and stepwise) with BVS. For BVS, we use the Gibbs sampler that minimizes the difficulties in prior selection associated with regression models. A feasible subset will include those variables with the highest Bayesian probabilities of appearing in the Gibbs sample. The results are compared with those by forward, backward and stepwise methods of variable selection. For modelling consumer responses, we use the standard logistic regression procedure and Bayesian networks. The main task of Bayesian networks is to decompose a joint probability distribution into a set of local distributions. In practice, a Bayesian network is a graphical representation that depicts conditional independence among variables and encodes the joint probability distribution. The optimisation and selection of models are achieved using evolutionary programming (Cui et al., 2006). Both logistic regression and Bayesian networks generate a probability score for each case or unseen data for the purpose of validation and forecasting.

We examine the effect of variable selection on the performance of the two classification methods by testing the predictive ability of the models based on the selected variables on unseen cases, i.e., the testing (validation) data. Error rate or the percentage of correct classification for a new dataset is often used as a measure of predictive accuracy. However, in direct marketing applications, simple error rate may not be appropriate for assessing the performance of classifiers. Most direct market campaigns have very low response rate (e.g., 5–15%). Moreover, due to budget constraints, direct marketing campaigns often only contact a small pre-set percentage of potential customers, say, the top 10% of customers based on their probability to purchase. Thus, the percentage of responders in the top decile of a customer list identified by a model based on the response probability scores, or the top decile "lift" will be used as the evaluation criterion. To measure the performance of a classifier at different depths of file, the measure of cumulative "lift" is the percentage of true positives (multiplied by 100) identified by the classifier in comparison with that identified by a random model or no model based on the total number of records at a given decile. This criterion can help evaluate the effectiveness of variable selection and the performance of forecasting models using the selected subsets. To validate the results of variable selection on these two modelling methods, we perform the validation experiments on a holdout dataset as well as on the entire dataset.

6. Results

For our variable selection experiments, we include 11 predictor variables that direct marketers usually work with to build forecasting models. They consist of six variables from customer purchase history and five demographic and credit history variables: lifetime orders (Liford, V1), credit card orders (Hcrd, V2), telephone orders (Tele, V3), recency (V4), monetary value of previous purchases

(Money, V5), income (V6), household size (V7), education (V8), home value (V9), credit rating (V10), and frequency (V11). First, we drew a stratified random sample of 10,000 cases from the entire dataset as the estimation sample for variable selection. Then, from the remaining data, we drew another stratified random sample of 10,000 for the purpose of holdout validation. So both datasets have the same response rate as the original dataset (5.47%).

For logistic regression, we use the forward, backward and stepwise selection methods and adopt the default significance level of 0.05 as the criterion for including a variable. Backward and stepwise methods, which are more time consuming, are considered superior to the forward selection method. However, the forward and backward selection methods produce the same set of five variables out of the 11 variables: V1, V2, V4, V9, and V11 (Table 1). Stepwise selection failed to converge after repeated trials, perhaps due to the large size of the dataset. Thus, we only use the forward selection variable set for subsequent comparison and discussion. Then, we compare the performance of logistic regression models in terms of model fitness and predictive accuracy. Logistic regression using the variable set by forward selection results in a simple error rate of 4.95% and an *R*-squared of 0.045. Thus, it performs nearly as well as the logistic regression using the full variable set (4.94% and an *R*-squared of 0.045). The results suggest that forward selection method makes little difference in model fitness, although the forward selection set has slightly lower AIC and BIC values than the full set. In terms of predictive performance on the holdout data (Table 4), the top decile lift of the logistic regression model using the forward selection set is exactly the same as the one using the full variable set (398.8). But on the second file, the logistic regression model using the forward selection set performs only slightly better than the model using the full variable set (283.7 vs. 282.6).

For BVS, we first generate the variance–covariance matrix of all the variables using the data from the previous year. Then, an inverse Hessian matrix was generated, and its diagonal was used to elicit the informative priors for the BVS procedure. An intercept is included in every model. Since the sample size is very large ($N = 10{,}000$), we used only 200 Gibbs iterations to reach convergence (Table 1). Altogether, the BVS procedure compares 2047 models or sets of variables. We also used 1000 and 10,000 iterations but did not produce any better results (Table 2). In Table 2, we present the standard measures of fitness, including Bayesian Information Criterion, Akaike Information Criterion, the log-likelihood, *R*-squared and simple error rate. It appears that given the large dataset, there is little difference in the results of different methods of variable selection. Only the full variable set and the BVS set (c0.001, i1000) have a slightly lower simple error rate (4.94%).

Then, we need to tune the priors controlled by parameter a_0, which assumes a normal distribution with 0.5 as the mean of the distribution and the weight of the prior. For a higher mean of the normal distribution, we apply a greater weight for the prior (e.g., 0.98). C_0 controls the prior model space, which assumes values from 0.01 to 15. When C_0 assumes a value of 10 or higher, it almost includes all the variables. The results of the sensitivity analyses for the BVS procedures are included in Table 3. It appears that BVS (c0.01, a(0.5, 0.008)) offers the highest probability with the following variables: V2, V4, V5, V6, V7, V8, V11, with a probability score of 0.705. The selected variables are different from those selected by the forward selection method using logistic regression. In Fig. 1, we observe a monotonic decrease in the posterior probabilities of alternative models. After the 130th model, the posterior probabilities of models virtually become zero. The forward selection variable set from logistic regression is the 12th best model among those ranked by BVS.

Table 1
Variables selected by different methods.

Model	Variables	Probability
LR full	V1–V11	
LR forward	V1, V2, V4, V9, V11	
BVS(c0.01, i200)	V2, V4, V5, V6, V7, V8, V11	0.705
BVS(c10, i200)	V2, V4, V5, V6, V7, V8, V9, V10, V11	0.708
BVS(c15, i200)	V1, V2, V4, V6, V7, V8, V9, V10, V11	0.629

Note: For all BVS models, $a = (0.5, 0.008)$.

To examine the effect of variable selection on model performance, we have conducted two validation experiments, one on the holdout dataset ($N = 10{,}000$, Table 4) and another on the entire dataset ($N = 106{,}284$, Table 5). We compare the predictive performance of the variable sets selected by forward selection and by BVS. We also compare the performance of logistic regression with that of Bayesian networks. The left side columns of Table 4 contains the results of the validation using logistic regression on the holdout dataset: cumulative lifts across the 10 deciles and different methods of variable selection. The results indicate that the BVS set has the highest top decile lift (409), which is nearly 10 points higher than that of the forward selection set and the full set of variables (both at 398.8). Using Bayesian networks (right side columns of Table 4), the BVS set again has a higher top decile lift than the forward selection set and the full set (410.1 vs. 405.6).

Validation on a larger dataset is a better test of the variable selection methods and the modelling methods. Then, we also validate both the logistic regression and Bayesian networks on the entire dataset ($N = 106{,}248$) Based on the results in Table 5, logistic regression using the BVS set again has a very small advantage over the forward selection set and the full set (352 vs. 351). However, with Bayesian networks, the BVS set achieves a top decile lift of 435, which is significantly higher than that of the forward selection set (427) and the full set (403). On the second decile, both the BVS and the forward selection set have higher cumulative lift than the full set. Overall, the results suggest that BVS provides a better set of variables that can help improve the performance of predictive models than the forward selection method and the full variable set. Bayesian networks, a model of joint distribution that considers the interrelations among variables, benefit more from Bayesian variable selection than logistic regression.

7. Discussion

7.1. Findings and implications

First, given the large sample size, forward and backward selection of variables makes little difference in the variables that are actually selected. Stepwise selection is computationally inefficient with a large dataset. Second, BVS by relying on the informative priors results in a different set of variables. Third, by using the top decile lift as the performance criterion for the forecasting models, BVS improves the performance of logistic regression over the forward selection method based on the holdout validation. But the improvement is negligible when validation is performed on the entire dataset. Furthermore, Bayesian networks, a model of joint distribution, achieve better predictive results based on the BVS set by exploring the interactions among variables, thus benefit more from BVS than logistic regression. In other words, BVS can potentially supply a set of variables with less noise and give a better opportunity to identify the underlying data distribution. Overall, the results suggest that BVS using informative priors from customer purchase history provides a feasible solution for exhaustive search to select variables and build forecasting models.

Having more variables and data is often a mixed blessing and does not guarantee building better forecasting models. Managers

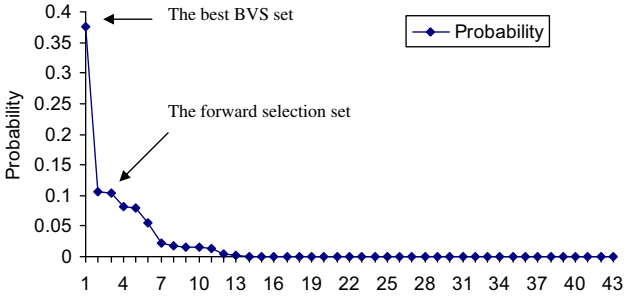
Table 2
Comparison of different variable selection methods.

Model/fitness criteria	AIC	BIC	$-2\log L$	R-Squared	Simple error rate (%)
LR full	3566.7	3653.2	3542.7	0.045	4.94
LR forward	3557.2	3600.4	3545.2	0.045	4.95
BVS(c0.01, i200)	3568.9	3626.6	3552.9	0.044	4.96
BVS(c10, i200)	3571.3	3643.4	3551.3	0.044	4.99
BVS(c15, i200)	3565.0	3637.1	3545.0	0.045	4.96
BVS(c0.01, i1000)	3559.1	3609.6	3545.1	0.045	4.94
BVS(c10, i1000)	3566.9	3617.4	3552.9	0.044	4.96
BVS(c10, i10000)	3560.0	3617.7	3544.0	0.045	4.95

Table 3
Tuning the weight of the priors.

Model	Selected variables	Probability
BVS(c0.01, a(0.5, 0.083))	V1, V2, V4, V5, V6, V7, V8, V9, V10, V11	0.526
BVS(c0.01, a(0.5, 0.023))	V2, V4, V5, V6, V7, V8, V9, V10, V11	0.648
BVS(c0.01, a(0.5, 0.008))	V2, V4, V5, V6, V7, V8, V11	0.705
BVS(c0.01, a(0.98, 3.7×10^{-4}))	V2, V4, V5, V6, V7, V8, V9, V10, V11	0.581

Note: "a" is actually controlled by a normal distribution: "μ" $\to$ 1, a large prior weight desired; and, reasonable $\mu/100 \leqslant \sigma 2 \leqslant \mu/10$, Models (for $i = 200$).



Note: X axle represents very 10th set before 13 and every 100th set thereafter.

Fig. 1. Posterior probabilities of variable sets by BVS.

Table 5
Validation of the results on the entire dataset.

Method/decile	Logistic regression			Bayesian networks		
	Forward	BVS	Full	Forward	BVS	Full
1	351.5	352.3	351.2	427.1	434.6	403.4
2	244.9	244.4	246.4	287.6	287.6	284.7
3	207.4	207.2	208.7	228.6	224.8	220.8
4	183.3	182.3	183.9	189.0	189.5	186.4
5	160.9	161.1	160.8	163.5	164.9	161.8
6	143.0	143.5	143.3	144.9	144.8	143.8
7	129.2	129.5	129.4	130.0	129.9	129.9
8	118.1	118.2	117.9	118.4	118.7	118.1
9	108.4	108.4	108.2	108.5	108.4	108.3
10	100.0	100.0	100.0	100.0	100.0	100.0

Note: $N = 106{,}284$.

Table 4
Validation results on the holdout data.

Method/decile	Logistic regression			Bayesian networks		
	Forward	BVS	Full	Forward	BVS	Full
1	398.8	409.0	398.8	405.6	410.1	405.6
2	283.7	280.6	282.6	288.1	287.4	288.1
3	223.4	223.4	217.2	223.0	224.5	223.0
4	183.4	184.5	181.9	186.0	187.4	186.0
5	159.1	157.4	155.8	161.3	161.9	161.3
6	141.5	140.1	139.1	143.1	143.3	143.1
7	128.0	126.6	127.7	128.9	129.0	128.9
8	117.9	116.6	117.2	118.2	117.9	118.2
9	108.3	108.0	107.1	108.8	108.6	108.8
10	100.0	100.0	100.0	100.0	100.0	100.0

Note: BVS model with (c10, i1000), and $N = 100{,}000$.

face an ever-growing need to reduce the number of variables effectively. Although researchers can rely on prior experience and exercise their judgment in trial-and-error selection processes, the increasing variety and number of variables would make an automated variable selection solution more desirable. BVS provides an efficient and exhaustive method to select variables for subsequent model building. It allows researchers to use the insight from previous studies to build more accurate forecasting models and can potentially improve the performance of direct marketing operations. An advanced method of variable selection also requires more sophisticated modelling methods that consider the interrelations among variables. Given the demand for useful information on a timely basis, methodological and technological advances should be undertaken to greatly reduce the marketing research cycle time (Malhotra, Peterson, & Kleiser, 1999). As the amount and variety of data collected by marketers continue to grow, the method advanced here provides an efficient tool for marketing managers to extract and update knowledge from the continuous data inflow in a timely fashion and to select better subsets for building forecasting models and assisting management decisions.

7.2. Limitations and suggestions

Due to space and time limitation, we compare the BVS method only with the forward and backward selection in logistic regression. Other approaches to variable selection can offer more interesting comparisons among the competing methods. The results of the study are based on direct marketing dataset. The proposed method and its generalizability need to be tested on other types of data and problems. Although the results are very encouraging for applications of BVS in business forecasting, the validation was done only on one holdout dataset and the entire dataset of the same period. Validating the model and BVS set on future data would provide stronger evidence on the merit of the proposed method and its applicability and potential benefits under the real-life business scenarios. This approach is also very useful when the researchers have new variables to examine while incorporating the effect of pre-existing variables from a previous study. It is

possible to give historical data different weights. Moreover, variable selection can also be seen as a natural by-product of Bayesian networks learning (Chen, Hao, & Ibrahim, 2000).

Furthermore, the BVS approach can also be applied to other marketing research problems that are based on large customer databases, such as predicting brand switching, churn behaviour, loan default and other issues related to forecasting sales or losses and managing customer relationships. These problems are similar to direct marketing forecasting in many ways, including large datasets, a small class of the target customers, and perhaps budget constraints that require accurate forecasting models and targeted marketing actions. These large noisy datasets and the great number and variety of variables make an automated or semi-automated process of variable section an attractive alternative. Given its efficiency in computing conditional and marginal probabilities, Bayesian variable selection has the potential to provide an efficient solution for fully automatic variable selection that can help to improve forecasting accuracy and the performance of marketing and business operations.

References

Austin, P. C., & Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology, 57*(11), 1138–1146.

Baesens, B., Viaene, S., van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research, 138*(1), 191–211.

Berger, P., & Magliozzi, T. (1992). The effect of sample size and proportion of buyers in the sample on the performance of list segmentation equations generated by regression analysis. *Journal of Direct Marketing, 6*(1), 13–22.

Blattberg, R. C., & Dolan, R. J. (1981). An assessment of the contribution of log linear models to marketing research. *Journal of Marketing, 45*(2), 89–97.

Chen, M.-H., Ibrahim, J. G., & Yiannoutsos, C. (1999). Prior elicitation, variable selection, and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society, Series B, 61*, 223–242.

Chen, M.-H., Hao, Q. M., & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. Berlin: Springer.

Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.-A., Morishita, S., Page, D., et al. (2002). KDD cup 2001 report. *SIGKDD Explorations, 3*(2), 1–18.

Cui, W., & George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference, 138*(4), 888–900.

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science, 52*(4), 597–612.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association, 95*(452), 1304–1308.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistics Association, 88*, 881–889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica, 7*, 339–373.

George, E. I., McCulloch, R. E., & Tsay, R. (1995). Two approaches to Bayesian model selection with applications. In D. Berry, K. Chaloner, & J. Geweke (Eds.), *Bayesian statistics and econometrics: Essays in honor of Arnold Zellner* (pp. 339–348). New York: Wiley.

Malhotra, N. K., Peterson, M., & Kleiser, S. B. (1999). Marketing research: A state-of-the-art review and directions for the twenty-first century. *Journal of the Academy of Marketing Science, 27*(2), 160–183.

Miller, A. J. (1990). *Subset selection in regression*. London: Chapman & Hall.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2), 134–148.

Rogue Wave Software (2009). Business analysis module user's guide. <http://www2.roguewave.com/support/docs/sourcepro/edition9-update1/html/analyticsug/4-2.html> Accessed 14.03.09.

Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science, 22*(3), 304–328.

Shtatland, E.S., Moore, S., Dashevsky, I., Miroshnik, I., Cain, E. & Barton, M.B. (2000). How to be a Bayesian in SAS: Model selection uncertainty. In *Proc Logistic and Proc Genmod, Proceedings of the Northeast SAS(r) Users Group, Inc. Annual Conference* (pp. 724–732), Philadelphia.

Zahavi, J., & Levin, N. (1997). Applying neural computing to target marketing. *Journal of Direct Marketing, 11*(4), 76–93.