Using simulated annealing to optimize feature selection problem in marketing applications

Article *in* European Journal of Operational Research · June 2006

DOI: 10.1016/j.ejor.2004.09.010

2 authors, including:

Jacob Zahavi
Tel Aviv University
50 PUBLICATIONS **891** CITATIONS

Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 171 (2006) 842–858
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Using simulated annealing to optimize the feature selection problem in marketing applications

Ronen Meiri [a,*], Jacob Zahavi [b,1]

[a] Tel Aviv University, P.O. Box 39040, Tel Aviv 69978, Israel
[b] The Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19102, USA

Available online 28 October 2004

Abstract

The feature selection (also, specification) problem is concerned with finding the most influential subset of predictors in predictive modeling from a much larger set of potential predictors that can contain hundreds of predictors. The problem belongs to the realm of combinatorial optimization where the objective is to find the subset of variables that optimize the value of some goodness of fit function. Due to the dimensionality of the problem, the feature selection problem belongs to the group of *NP*-hard problems. Most of the available predictors are noisy or redundant and add very little, if any, to the prediction power of the model. Using all the predictors in the model often results in strong over-fitting and very poor predictions. Constructing a prediction model by checking out all possible subsets is impractical due to computational volume. Looking on the contribution of each predictor separately is not accurate because it ignores the intercorrelations between predictors. As a result, no analytic solution is available for the feature selection problem, requiring that one resorts to heuristics. In this paper we employ the simulated annealing (SA) approach, which is one of the leading stochastic search methods, for specifying a large-scale linear regression model. The SA results are compared to the results of the more common stepwise regression (SWR) approach for model specification. The models are applied on realistic data sets in database marketing. We also use simulated data sets to investigate what data characteristics make the SWR approach equivalent to the supposedly more superior SA approach.

Keywords: Feature selection; Stepwise regression; Simulated annealing; Database marketing

1. Introduction

Feature selection (also, specification) is a process of selecting the most influential subset of features (or predictors) from a usually much larger set of original variables to optimize a predefined

* Corresponding author.
E-mail addresses: meirir@post.tau.ac.il (R. Meiri), zahavi@wharton.upenn.edu (J. Zahavi).
[1] On leave from Tel Aviv University.

0377-2217/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2004.09.010

"goodness of fit" measure. Blum and Langley (1997) and Dash and Liu (1997) defined the objective of feature selection as "the improvement of the prediction accuracy or decreasing the size of the structure without significantly decreasing prediction accuracy". When the number of predictors entering the model is large, the probability that insignificant predictors will enter the model increases (Type-I error). Reducing the significant level will reduce Type-I error but it may prevent significant predictors from entering the model, thus causing a Type-II error. The overall quality of the model is determined by two components: the variance of the prediction error calculated based on the training data and the bias estimated on a set of independent validation data (Geman et al., 1992).

The feature selection problem belongs to the class of *NP*-hard problems known as *induction of minimal structures* (Hancock, 1989; Blum and Rivest, 1992; John et al., 1994). When the number of potential predictors, k, is large, the selection process cannot be solved exactly with an acceptable amount of computation time. Searching the exact solution is equivalent to checking all $O(k)^2$ states and comparing them to find the best ones. Even with medium size k this task is practically impossible. Consequently, heuristic optimization algorithms have evolved, including *iterative improvement algorithms* and *stochastic search methods*, to solve large scale combinatorial problems. In the feature selection problem, the objective is typically to find subset of features that minimize the prediction errors (or some function of them). However, we note that since the distribution of the data is not known, the best we can do is to optimize the errors relative to some known data observations rather than relative to the "true", but unknown, distribution of the population. Thus the goal of the feature selection process is to get as close as possible to the "true" optimal solution with reasonable amount of computation time.

Classic optimization methods such as *iterative improvement algorithms* starts with an initial solution and seek the optimal solution in an iterative process where each step moves in the direction of the best improvement until no further improvement is possible (van Laarhoven and Aarts, 1987). These algorithms are greedy (or myopic), in the sense that they look only one step ahead, and thus can get "stuck" at the first local optima. Furthermore, these algorithms are very sensitive to the initial conditions and can yield different solutions for different starting points. Stepwise regression (SWR) (Miller, 2002) is a typical example of an iterative improvement algorithm. Stochastic search methods, on the other hand, such as simulated annealing (SA), have the advantage that they can escape local optimum and converge go the global optimum under certain conditions (Anily and Federgruen, 1987; Cruz and Dorea, 1998).

One can view the feature selection problem as a search process for the optimal subset configuration of predictors in a 2^n space. Subset selection algorithms can be divided into three categories: Exhaustive/complete search methods, Heuristics methods, and Non-deterministic search methods. (Motoda and Liu, 2002).

1.1. Exhaustive/complete

These methods interrogate all possible subsets, under some criterion, to look for the best possible configuration. For instance, if we limit the number of features to M, the number of possible configurations is $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{M}$. Actually, in some cases (e.g., the subset has some monotonic properties) it is not necessary to go over all possible configurations (Schilmmer, 1993). However, when n is large, even this shortcut may not help since the remaining search space is still large (Liu and Motoda, 1998).

1.2. Heuristics

These methods, further divided into filters and wrappers methods, are used to find a satisfactory solution within an acceptable amount of time.

1.2.1. Filter methods

The filter approach attempts to remove irrelevant features from the feature set before it is used by the learning algorithm (Liu and Motoda, 1998). For example, features can be selected according to their correlation with the dependent variable. The

disadvantage of the correlation test is that it does not depend on the type of the prediction model involved, whether linear regression, logistic regression, decision trees, or other. Several filtering methods have been developed to cope with this issue. FOCUS uses a sequential forward search method based on consistency criteria (Almuallim and Dietterich, 1991, 1994); The Relief algorithm assign a "relevance" weight for each feature according to the target (Kira and Rendell, 1992; Kononenko, 1994); IC3 is based on decision trees (Cardie, 1993); Finally, we mention the filter approach of Koller and Sahami (1996) which uses the entropy criteria. The filtering approach is not computationally expensive, but it may return a large feature subset (Liu and Motoda, 1998).

1.2.2. Wrapper approach

The wrapper approach uses a heuristic search that evaluates the quality of the feature subset by using the prediction accuracy of the induction algorithm (Kohavi and Sommerfield, 1995). Wrapper methods include SWR, forward and backward feature selection algorithms (Devijver and Kitler, 1982), Greedy algorithms such as hill climbers (Caruana and Freitag, 1994) and the best first search that looks for the best promising node of a tree (Kohavi and John, 1998). Others are Bayesian search methods like Bayesian variable selection (Mitchell and Beauchamp, 1998), Gibbs sampling (George and McCulloch, 1993) that uses an iterative algorithm to evaluate and select the best subset of predictors and the Bayes factor method (Kasse and Raftery, 1995) which is a general method for multiple comparisons between a set of competing hypothesis (subsets).

1.2.3. Non-deterministic search

Finally, the non-deterministic methods search the subset space using stochastic search algorithms. Genetic algorithms (GA) and simulated annealing (SA) belong to this class of models (Eiben et al., 1991). In particular, GA selection process that keeps the one or more of the best solutions, also known as *elitism*, converges to the global optimum (Rudolph, 1994). Therefore, GA and SA are more or less equivalent with respect to the quality of the solution and hence the conclusions derived with respect to the SA algorithm also extend to the GA algorithm. GA and SA prove to be promising methods for feature selection in image perception (Siedlecki and Sklansky, 1988; Jain and Zongker, 1997). Similar results can be found in other domains using GA (Melab et al., 2002).

In this paper we focus on the SA algorithm (Kirkpatrik et al., 1983; Metropolis et al., 1953) and compare it with SWR to address the feature selection problem for building a large-scale linear regression model. The objective is to investigate to what extent the supposedly superior SA can give a better predictive model by escaping the local optimum that SWR is likely to fall into. We use realistic data from the area of database marketing for the comparative analysis of these two methods. We also use simulated data sets to investigate what data characteristics make the SWR approach equivalent, if not better, than the SA approach.

2. The goodness of fit function

As mentioned above, the objective of feature selection is to find the subset of predictors which attains the "best" fit, as measured by a pre-specified goodness-of-fit (GOF) criterion. The GOF criterion can be any function that is capable of assessing the prediction power of the model. In general, we seek a GOF that is monotonically increasing or decreasing as the prediction power of the model improves. Many functions may fit this definition, each having its own advantages and disadvantages. The choice of the "right" GOF may depend on the problem. Table 1 presents a list of candidate GOFs from two "families", one which is based on the R-squared criterion, the other, the Akiaki Information Criterion (AIC), which is based on the likelihood function.

Two related criteria are the Bayesian information criteria (BIC) and the risk information criteria (RIC) penalty function. For the BIC, $f(n) = \log(n)/2$ (Schwarz, 1978) and for RIC, $f(n) = \log(p)$, where p is the number of available predictors (Foster and George, 1994). Other modified criteria have been suggested by Tibshirani and Knight (1999), Ye (1998), Hurvich and Tsai (1989, 1998), Wei

Table 1
Candidate GOFs

Name	Equation	Remarks	Refference
R^2	1-RSS/RSS0	Never decrease when new predictors are added	Miller (2002)
Adj-R^2	$1-(1-R_p^2)\frac{n-1}{n-q}$	Increases only if change in R^2 is significant	Miller (2002)
AIC	$-2(L_q-q)$	Decreases with significant changes in the likelihood function. AIC can be used in non-linear prediction models such as logistic-regression	Akaike (1974)
$AIC_{log-log}$	$-2(L_q-q\log(\log(n)))$	Improves the AIC for large samples	Hannan and Quinn (1979)
$AIC_{(1/2)log}$	$-2(L_q-q/2\log(n))$	Improves the AIC for large samples	Schwarz (1978)
AIC_{log}	$-2(L_q-q\log(n))$	Improves the AIC for large samples	Proposed in this work

(1992), Shao (1993), Zheng and Loh (1997), Benjamini and Hochberg (1995), Clyde and George (1999, 2000), Foster and George (2000) and Johnstone and Silverman (1998).

In our study we will explore several GOF functions from Table 1 including: Adj-R^2, *AIC*, and the modified *AIC* with penalty functions $f(n) = \log(\log(n)), \log(n)/2$ and $\log n$. We note that we are concerned here more with the feasibility of using combinatorial optimization methods for feature selection problems rather than with the efficiency of these methods.

3. Setup of study

Recapping, our objective is to compare the performance of SWR and SA for addressing the feature selection problem using real marketing data. The evaluation criteria are based on lift values, Gini coefficients and gains charts. These metrics are further discussed in Appendix A.

Each of the SWR and the SA algorithm is controlled by a set of parameters. In setting up the study, we experimented with several parameters of these algorithms. For one thing, in order to find the "best" model configuration; for another, in order to examine the stability of the algorithm.

The SA method can be controlled by means of parameters which influence the cooling process: the *cooling method*, the *cooling rate* and the *termination condition*. In particular, we use the homogenous annealing process in our study, which implies that when cooled slowly enough the system reaches a thermal equilibrium at the prevailing temperature. We use this concept to change the temperature in the SA approach only when the system has reached equilibrium at the current temperature, as indicated by the rate of change in the objective function value. The latter is given by the ratio j/m, where m is the number perturbations (iterations) required to attain j changes in the objective function value. We define the set of m iterations as a "round". A perturbation is a neighboring solution obtained by randomly adding or removing a single predictor from the current set of predictors. The system attains equilibrium at a given temperature level if the rate of change between two successive "rounds" of iterations remains constant within a pre-defined confidence interval. Practically, we calculate the confidence interval for the difference in the rate of change for the given confidence level based on the binomial distribution. Then, if the resulting rate of change falls within the confidence interval, we conclude that the system has reached thermal equilibrium and reduce the temperature; otherwise, we keep perturbing the solutions by creating neighboring solutions, until reaching equilibrium. In this research we set $j = 10$. Different *confidence intervals* were tested in the range $\alpha = 0.4$ to $\alpha = 0.8$ with increments of $\Delta\alpha = 0.2$. Once the system has reached equilibrium, the temperature parameter is reduced by a *cooling rate* r, where r varies in the range $[0.85, 0.97]$ with increments. $\Delta r = 0.03$. So all in all, we tested 3 different *confidence intervals* with 4 different *cooling rates* and 5 different GOF functions for a total of 60 configurations. For each configuration we have stopped the process after 60,000 iterations.

The SWR algorithm is controlled by two significance values—the level of significance to enter a

Table 2
Significance values for SWR algorithm

Configuration	*P*-to-enter (%)	*F*-to-delete (%)
1	10.0	15.0
2	5.0	10.0
3	2.5	5.0
4	1.0	2.0
5	0.5	1.0

variable into a model, *F-to-enter* (F_e), and the level of significance to remove a variable from the model, *F-to-delete* (F_d). Several values of (F_e) and (F_d) were tested in this study, as exhibited in Table 2.

4. The data

Four realistic data sets, all drawn from the marketing domain, were used in this study. These data sets were made available for the research community by the DMEF (Direct Marketing Educational Foundation). Based on the application, we refer to these data sets as: ''Non-Profit'', ''Catalog'', ''Specialty'' and ''Gift'', respectively. The observations in these data sets correspond to individual customers with attributes which represent purchase and promotion history and demographics. Table 3 provides more details about the types and the magnitude of each of the data files. Several transformations were applied to all data sets to prepare them for the modeling purposes:

(a) Dates were converted to number of days from a reference date (e.g., the date of entry to the database).
(b) Categorical variables were expressed by means of dummy variables, one variable per each value of the categorical variable (less one, which is used as a reference point).
(c) Zip codes were converted to categorical variables according the first two digits.
(d) Variables with more than 5% of missing values were removed from the model. Otherwise, we have substituted the missing value in the observation record with a zero value and added a dummy variable to represent the missing value.
(e) Variable values which were more than six standard deviations away from the variable's mean (outliers), were truncated by setting their value to $6 * \sigma$, where σ is the standard deviation. This transformation was not applied on binary variables.

5. Results

The SWR results are exhibited in Table 4, the SA results in Table 5. We use several metrics to assess the goodness of fit of the models: the number of predictors entering the model (# Pred.), R^2 in the training set (R^2_{tr}), R^2 in the validation set (R^2_{vl}), R^2-ratio between the validation set and the training set, the Gini coefficient and the maximum lift (ML). These metrics are presented for each data file and for each of the parameter configuration tested.

Table 6 compares the SWR and SA modeling results for the metrics above. The SWR results corre-

Table 3
The datasets

Name	# Observ.	# Initial var.	# Total pred.	Response variable	Remarks
Non-profit	99,200	77	307	Binary	Non-profit organization. Objective is to predict the expected donation amount for each customer
Catalog	96,551	165	226	Binary	Catalog mailing. Objective is to predict the probability that a customer responds to a catalog mailing
Specialty	106,284	287	350	Continues	Mailing offers. Objective is to predict the response rate to a mailing solicitation
Gift	101,532	99	104	Counter	Catalog mailing. Objective is to predict the expected number of orders for each customer

Table 4
SWR results

File	*F*-to-enter	*F*-to-delete	# Pred.	R^2_{tr}	R^2_{vl}	R^2-ratio (vl versus tr)	Gini	ML
Non-profit	10.0	15.0	61	0.1279	0.1202	0.9395	0.314	0.228
	5.0	10.0	51	0.1274	0.1204	0.9453	0.315	0.227
	2.5	5.0	39	0.1264	0.1203	0.9523	0.315	0.227
	1.0	2.0	31	0.1255	0.1200	0.9560	0.315	0.228
	0.5	1.0	28	0.1251	0.1200	0.9592	0.315	0.228
Catalog	10.0	15.0	38	0.0146	0.0050	0.3454	0.238	0.174
	5.0	10.0	19	0.0129	0.0070	0.5444	0.261	0.189
	2.5	5.0	9	0.0119	0.0081	0.6783	0.273	0.199
	1.0	2.0	9	0.0119	0.0081	0.6783	0.273	0.199
	0.5	1.0	8	0.0117	0.0082	0.6955	0.273	0.202
Specialty	10.0	15.0	42	0.0758	0.0661	0.8719	0.528	0.384
	5.0	10.0	36	0.0744	0.0662	0.8890	0.527	0.387
	2.5	5.0	25	0.0740	0.0658	0.8893	0.522	0.383
	1.0	2.0	13	0.0726	0.0650	0.8962	0.528	0.384
	0.5	1.0	11	0.0722	0.0649	0.9000	0.527	0.384
Gift	10.0	15.0	49	0.3437	0.3385	0.9849	0.610	0.477
	5.0	10.0	37	0.3413	0.3388	0.9924	0.608	0.475
	2.5	5.0	37	0.3413	0.3388	0.9924	0.608	0.475
	1.0	2.0	33	0.3410	0.3388	0.9934	0.609	0.479
	0.5	1.0	31	0.3408	0.3388	0.9940	0.609	0.480

Table 5
SA results

File	GOF	# Pred.	R^2_{tr}	R^2_{vl}	R^2-ratio (vl versus tr)	Gini	ML
Non-profit	Adj-R^2	118.3	0.1296	0.1200	0.926	0.314	0.228
	AIC	49.1	0.1272	0.1200	0.943	0.314	0.228
	AIC $f(n) = \log(\log(n))$	32.3	0.1257	0.1197	0.952	0.314	0.229
	AIC $f(n) = \log(n)$	18.2	0.1222	0.1185	0.969	0.315	0.237
	AIC $f(n) = \log(n)/2$	26.2	0.1248	0.1197	0.959	0.315	0.237
Catalog	Adj-R^2	93.3	0.0169	0.0026	0.152	0.224	0.173
	AIC	32.5	0.0145	0.0050	0.344	0.240	0.182
	AIC $f(n) = \log(\log(n))$	14.3	0.0127	0.0076	0.598	0.265	0.192
	AIC $f(n) = \log(n)$	3.1	0.0106	0.0076	0.721	0.229	0.213
	AIC $f(n) = \log(n)/2$	4.1	0.0111	0.0081	0.733	0.268	0.198
Specialty	Adj-R^2	137.9	0.0785	0.0652	0.831	0.507	0.389
	AIC	46.7	0.0748	0.0659	0.881	0.506	0.386
	AIC $f(n) = \log(\log(n))$	17.1	0.0730	0.0657	0.900	0.505	0.387
	AIC $f(n) = \log(n)$	3.6	0.0676	0.0613	0.907	0.489	0.365
	AIC $f(n) = \log(n)/2$	8.3	0.0712	0.0651	0.914	0.500	0.381
Gift	Adj-R^2	64.8	0.3441	0.3381	0.983	0.608	0.476
	AIC	45.6	0.3427	0.3375	0.985.	0.610	0.478
	AIC $f(n) = \log(\log(n))$	37.5	0.3430	0.3381	0.986	0.611	0.479
	AIC $f(n) = \log(n)$	29.2	0.3415	0.3378	0.989	0.610	0.482
	AIC $f(n) = \log(n)/2$	33.5	0.3425	0.3382	0.987	0.611	0.480

sponds to the model with $F_e = 0.5\%$ and $F_d = 1.0\%$. The GOF function in SA method in Table 6 is AIC with $f(n) = \log(n)/2$, the cooling rate is $r = 0.94$ and the confidence interval is $\alpha = 0.4$.

Table 6
SWR versus SA

	SA				SWR			
	Non-profit	Catalog	Specialty	Gift	Non-profit	Catalog	Specialty	Gift
Number of predictors	25	4	28	34	28	8	35	31
R^2-training	0.1245	0.0111	0.0710	0.3426	0.1251	0.0117	0.0722	0.3408
R^2-validation	0.1196	0.0081	0.0650	0.3383	0.1200	0.0082	0.0649	0.3388
R^2-ratio	0.9600	0.7345	0.9157	0.9874	0.9592	0.6955	0.9000	0.0649
Gini	0.3155	0.2678	0.5002	0.6114	0.3151	0.2733	0.5018	0.6093
ML	0.2362	0.1987	0.3816	0.4797	0.2282	0.2020	0.3837	0.4797

While the SA algorithm seems to be slightly better than SWR, there is no substantial difference between the two methods. This result may looks at first surprising because of the prior assumption that SA is a "superior" optimization scheme due to its capacity to skip local optimal points. While this might be true in general, we see, by carefully investigating the results in Table 6 that this does not hold true for our four marketing data sets. In fact, the final set of predictors entering the model were almost the same in the two methods with only slight variations (see Appendix B for the list of predictors that made it to the final model in each case). Moreover, in both cases, the optimal set of predictors set contained relatively very few predictors that carry most of the prediction power. The fact that we obtained almost the same solution from two algorithms which are every bit as different from one another suggests that not only this solution is dominant, but that it is also most likely to be a global one. The interesting thing to note here is that SWR algorithm found this global solution, skipping whatever local optimal solutions that it may have encountered along the way. For example, in the "non-profit", "catalog" and "Specialty" data files, the set of predictors found in the SA algorithm were almost identical to those found in the SWR model. The "Gift" data file showed similar results. The majority of the predictors (29) were common to both models, and only 5 predictors appeared in the SA model but not in the SWR model and 2 appeared in the SWR model but not in the SA model.

We note, however, that even though the SA and SWR produce similar models, each method has its own advantages. SA was found to be almost insensitive to the optimization parameters, *cooling ratio* and *significance level*, with the modified AIC criteria with penalty function of $f(n) = \log(n)/2$ yielding the best performance for all the four files. Thus SA has the advantage that it is more stable. SWR, on the other hand, is faster and easier to compute. However SWR is extremely sensitive to the threshold significance levels F_e and F_d. For example, for the data files non-profit and Gift and SWR parameters $F_e = 0.5\%$ and $F_d = 1\%$, SWR and SA models were almost identical. However, the same parameters values, when applied on the Catalog and Specialty data files, resulted in the SWR model containing more predictors than the SA model. Therefore, no single set of the parameters exist for SWR that is suitable for all problems. This requires that one test different parameter sets to find out the best configuration for the specific problem, which could be a very tedious process especially when dealing with large scale problems.

Given that the SWA and SA gave similar results, we now proceed, in the next section, to understand the reason behind this phenomenon, by means of a simulation study.

6. Simulated data experimentation

Theoretically, the main advantage of SA over the "linear" SWR methods for feature selection is that SA can capture complex non-linear relationships that SWR cannot. To recall, this did not happen in our case, as in all data sets that we tested, SWR gave results that are very comparable to SA. We hypothesize that this phenomenon occurs because marketing databases are well-behaved and "homogenous". We test this hypothesis in this section. The idea is to create simulated data-

sets with complex relationships between predictors, which will cause SWR to fail, and then explore whether SA is capable of identifying these complex relationships and introducing them to the model.

We recall that SWR is a "greedy" algorithm, whereas SA is not. That is, the algorithm looks only on the next variable in line (the "current" variable) to determine whether to introduce or eliminate it from the model. In the simulated files we artificially create two sets of independent variables—a set of variables that as a group are strongly correlated with the dependent variable by means of a linear transformation, and a much larger set of predictors which are not related to the dependent variable. Clearly, in a "good" model, one would expect that the set containing the independent variables that are correlated with the dependent variable will make it into the model, whereas the set containing the uncorrelated independent variables will not. Our hypothesis is that because of the way SWR works, most, if not all, of those linearly-dependent predictors will be missed by the greedy SWR algorithm but not by the non-greedy SA algorithm.

6.1. The simulation process

The simulation is aimed at creating a set of p predictors, $p \ll k$, that are linearly related to the dependent variable via a linear combination (the "true" model), such that when tested each on its own, each of these p predictors will turn to be insignificant and thus eliminated from the model. However, when tested as group, most of them will turn significant and worthy of entering the model. Two simulated datasets are created with different number of variables, k, and different number of predictors in the "true" model, p.

The idea is to randomly select the value of $p-1$ variables (out of a larger set of $k-1$ predictors) and the value of the dependent variable, y, and then find the value of the pth predictor as a linear combination of the $p-1$ predictors and y. This creates a dependency between the p variables and y that justifies introducing these variables into the model. The simulation process consists of several steps:

- Generate the dependent variable, y, from the standard normal distribution.
- Generate $k-1$ normally distributed independent variables denoted $x_2,\ldots,x_k$.
- Create the "true" model by expressing the first predictor, x_1, as a linear combination of $p-1$ independent variables, $x_2,\ldots,x_p$, and the dependent variable, y, as follows:

$$x_1 = y - \gamma_1(x_1 + x_2 + \cdots + x_{p-1}) + \gamma_2 \cdot \mathrm{rand}(\mathrm{normal}), \quad (1)$$

where γ_1, and γ_2 are two parameters that control the correlation between x_1 and y as well as the overall significance of the model. The function *rand*(*normal*) generates random numbers from the standard normal distribution. Its purpose is to introduce noise into the process to make the data more "realistic".

Clearly, because the $p-1$ variables, $x_2,\ldots,x_p$, are random variables, drawn independently of the dependent variable y, none of these variables is expected to be significant enough on its own to warrant introducing it into the model. However, the variable x_1, being related to y via Eq. (1), is expected to be significant and worthy of entering the model. But we note that if γ_1 is large enough, x_1 may also turn out to be insignificant because it may be dominated by the random numbers x_2 to x_p. We therefore seek the value of γ_1 and γ_2 that guarantees that when tested on its own, x_1 will still not make it to the model, yet when tested along with the $p-1$ predictors, it will be introduced into the model. In fact, a "good" algorithm should introduce all these p variables into the model.

Two files were created for each combination of γ_1 and γ_2, each containing 100,000 records equally split into training and validation datasets, the first with $k=100$ variables and $p=5$ predictors in the "true" model and the second with $k=200$ variables and $p=8$ predictors in the "true" model. Different values of the parameters γ_1 and γ_2 were tested. We then tested the hypothesis about x_1, seeking the parameter values in Eq. (1) that renders the variable x_1 insignificant.

Table 7 exhibits the results for $k=100$ and $p=5$ for $\gamma_2=2$ and varying γ_1. We have selected

Table 7
Testing the significance of the simulated file with $k = 100$ $p = 5$

γ_1	Significance
1	~0.0
2	~0.0
3	1.29600E−161
4	3.63793E−91
6	2.06779E−41
8	7.91054E−24
10	2.39592E−15
12	1.21881E−10
14	5.11250E−08
16	2.59085E−06
20	0.000289
24	0.002629
28	0.013963
32	0.036592
40	0.112621
48	**0.222950**
56	0.334542
64	0.430413
128	0.905463

Table 8
Testing the significance of the simulated file with $k = 200$ $p = 8$

γ_1	Significance
1	~0.0
2	4.0169E−237
3	4.5731E−109
4	2.09604E−60
6	8.77455E−28
8	2.16776E−16
10	4.41293E−10
12	3.87728E−07
14	2.62896E−05
16	0.0003340
20	0.0045026
24	0.0306005
28	0.0806689
32	0.1524315
40	**0.3303783**
48	0.4855706
56	0.6314329
64	0.7551025
128	0.8126419

the file with $\gamma_1 = 48$ because it yields a large significance value that will guarantee that when considered on its own, x_1 hardly makes it into the model.

Table 8 exhibits the results for $k = 200$ and $p = 8$ for $\gamma_2 = 2$ and varying γ_1. Here we selected the file with $\gamma_1 = 40$ to make sure that x_1 hardly makes it into the model.

6.2. SWR results

The SWR results are presented in Table 9. File names in Table 9 (and also in Table 10 below) consists of the prefix 'Sim' (for simulation), followed by the number of predictors in the model, and then the number of predictors in the "true" model. As suspected, SWR has not been able to locate the optimal configuration, as none of the p variables were found significant to introduce into the model ("# true" in model is zero in both cases). The simulation has used $P_e = 1\%$ and $P_d = 2\%$.

Fig. 1 displays the gains chart of the SWR model with $k = 100$ and $p = 5$; Fig. 2 for $k = 100$ and $p = 8$. The data is sorted in descending order

Table 9
SWR results

File	*P*-to-enter	*P*-to-delete	# Pred in model	# True pred.	R^2_{tr}	R^2_{vl}	R^2-ratio (vl versus tr)	Gini	ML
Sim_100_5	1.0	2.0	2	0	0.00030	−0.00056	–	−0.2830	0.0234
Sim_200_8	1.0	2.0	1	0	0.00026	−0.00063	–	−0.3146	0.5070

Table 10
SA results

File	No. runs	# Pred.	# True	R^2_{tr}	R^2_{vl}	R^2-ratio (vl versus tr)	Gini	ML
Sim_100_5	15	5	5	0.20042	0.20289	1.0123	21.3646	15.1399
Sim_200_8	3	0	0	–	–	–	–	–
Sim_200_8	11	8	8	0.19866	0.20150	0.2015	70.9147	50.7274
Sim_200_8	1	10	8	0.19882	0.20127	0.2013	70.8800	50.7469

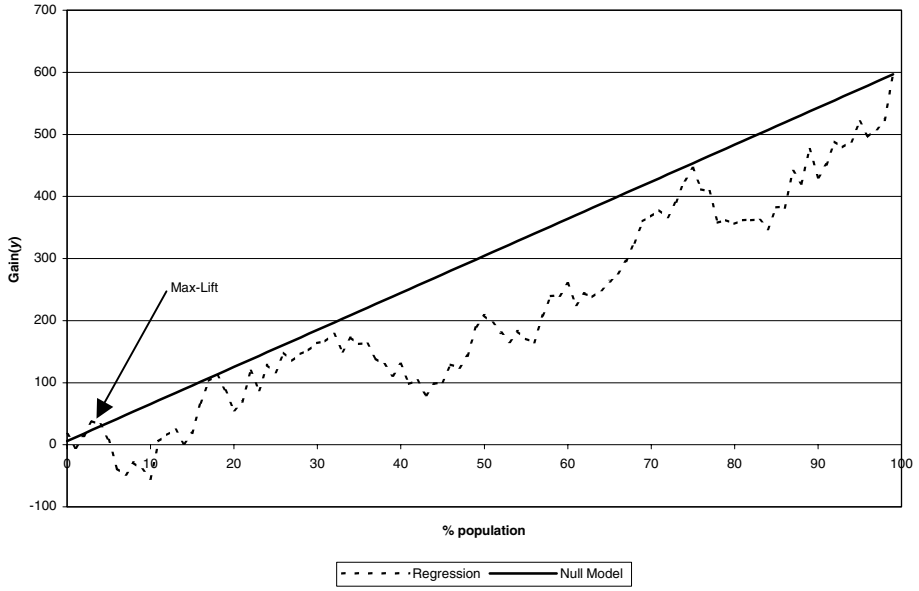

Fig. 1. Gains chart for SWR with $k = 100$ and $p = 5$. The chart shows a random behaviour of the population.

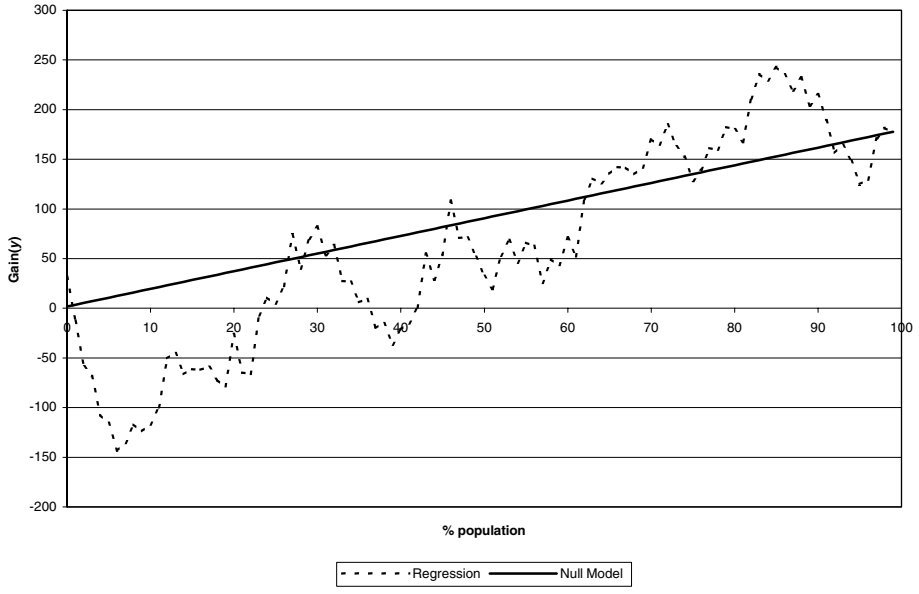

Fig. 2. Gains chart for SWR with $k = 200$ and $p = 8$. The chart shows a random behaviour of the population.

according to the model predictions. The X-axis represents the percentage of the population and the Y-axis the corresponding cumulative gains. We note that the sum of the gains over all observations in the simulated file is 580. The null model assumes that all customers are the same, resulting in an equal gain (=58) per each decile. This model is represented by the straight line in the gains chart

with a slope of 58 per decile. A "good" model should place the best customers in the first decile and the worst customers in the last decile. This should lead to a convex curve above the null model. Both gains charts in Figs. 1 and 2 indicate that the SWR model results looks nothing more than a random noise, meaning that SWR is not capable of separating between the "good" and the "bad" customers.

6.3. Simulated annealing results

Unlike SWR, SA is a stochastic algorithm that can converge, in the short term, to different optimal points. To test the convergence and the stability of the algorithm, we have repeated the process several times with different cooling parameters to test both the convergence and the stability of the algorithm. We used the following parameters for the SA runs:

- GOF function is the AIC criteria with a penalty function of the form $f(n) = \log(n)/2$.
- *Cooling ratio* is between 0.85 and 0.97, with intervals $\Delta I = 0.03$.
- *Confidence Interval* between 0.4 and 0.8 with increments $\Delta = 0.2$.

A total of 15 iterations, with different seed values, have been run for each of the simulated files. The results are presented in Table 10. From the table we see that all of the simulations in the first file ($k = 100$, $p = 5$) have converged to the global optimum with exactly all the 5 "true" predictors making it into the model and no other extra predictors. The second file with $k = 200$ and $p = 8$ is more complicated. Here, in 12 runs all the eight predictors in the "true" model were introduced to the model by the SA algorithm, in 10 of which these were the only predictors, in 2 others these predictors were accompanied by two other insignificant predictors. Only in 3 runs, out of the 15 different runs, neither one of the eight "true" predictors made it to the model.

Fig. 3 presents the gain-chart of the model with $k = 100$ and $p = 5$, Fig. 4 the gain-chart of the model with $k = 200$ and $p = 8$. As we can see, the SA model was able to place the observations with the positive y-values ("revenues") at the top of the deciles list and those with negative y-values

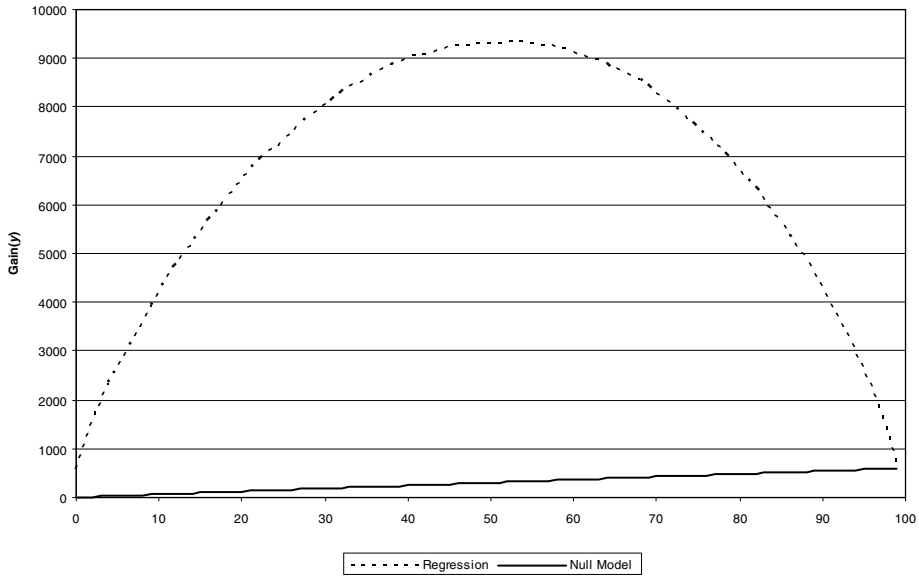

Fig. 3. Gain chart for SA with $k = 100$ and $p = 5$. The model clearly separates between positive y-values and negative y-values. This results in large *Gini-coefficient* and *maximum lift*.

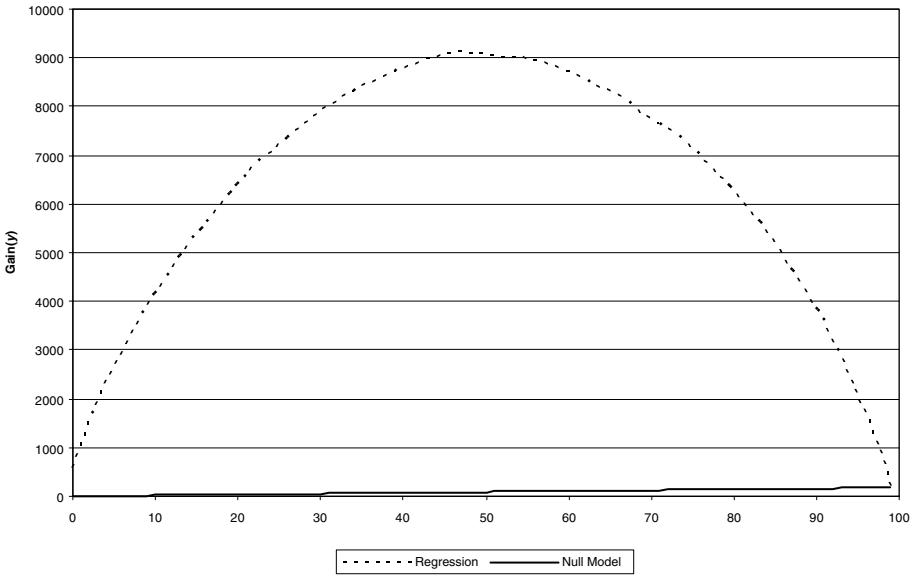

Fig. 4. Gain chart for SA with $k = 200$ and $p = 8$. The model clearly separates between positive y-values and negative y-values. This results in large *Gini-coefficient* and *maximum lift*.

(''losses'') at the bottom deciles. Consequently, the gains (defined here as the cumulative sum of the y-values), first increases until reaching a maximum level and then decreases until it meets the edge of the null model. This gains chart explains why we received such large values for the *Gini coefficient* and *maximum lift*.

We note that the gains charts in Figs. 3 and 4 are very much different than ''conventional'' gains charts that we are used to see in database marketing applications, the reason being that the dependent variable in these applications is always non-negative, and as a result the cumulative gains on the Y-axis is always increasing, yielding a monotonic gains chart curves. In our case, however, some of the y-values (the gains) are positive and some are negative, and as a result the gains chart exhibits a non-monotonic, almost a quadratic shape curve.

7. Conclusions

This work has compared the SWR and the SA algorithms for solving the feature selection problem. We have shown that while the SA algorithm is slightly better than the SWR algorithms, there is basically no substantial difference between the output results as both algorithms yield almost the same solution. The major difference between these models lies with the stability of the algorithm. SA was found to be more stable and almost insensitive to variations in the optimization parameters. SWR, on the other hand, was more affected by fluctuations in the optimization parameters, requiring that one experiment with several parameter configurations to find the ''best'' model. But the SWR algorithm is easier to implement and software availability is much wider.

The fact that SWR yields comparable results to SA may look at first surprising, since it contradicts the common assumption that an *iterative improvement algorithm* often gets ''trapped'' in local optima and may not converge to the global configuration. But it was found that when applied on marketing data, SWR was able to overcome any local optima along the way and converge to the global optimum. As evident from the results, the optimal solution is dominated by a relatively small set of predictors which carry most of the predictive power. Therefore, although several local

optimums may exist, their impact on the optimal solution is minor.

We hypothesized that this phenomenon occurred because marketing databases are well-behaved and "homogenous", for which SA do not have clear advantage. We tested this hypothesis by creating a simulated dataset with complex relationships between predictors, which causes the SWR to fail, and then exploring whether SA is capable of capturing these complex relationships and introduce them to the model. Indeed the simulation study showed that when complex structure exists in the data, SA outperforms the SWR algorithm, and is capable of converging to the global optimum, overcoming complex combination between predicting variables that SWR cannot.

The bottom line is that if complex structures exist in the data, SA is preferable to SWR because it is most likely to identify these structures. If no complex structure exist in the data, then SWR is likely to behave almost the same as the more complex SA algorithms. Because of the relative simplicity of the SWR algorithm and software availability, SWR may be preferable to other models in these cases.

The fact that SA and SWR yield similar results when run on real marketing databases, support our hypothesis that marketing data are well-behaved and homogenous with no irregularities and complex structure between the data elements. In these cases, the SWR algorithm is suitable enough for addressing the feature selection problem, ruling out the need to use more sophisticated algorithms such as SA, Genetic Algorithms, or others.

But we emphasize the fact that the results of this research are limited to the domain of database marketing and may not extend for other domains, such as insurance, finance, medicine, engineering, etc. In all likelihood, there are domains where the above phenomenon may still hold, others where they may not. Further research is required to generalize the results of this research to other domains.

Appendix A. Evaluation metrics

Several metrics are used in the database marketing field to evaluate models. In this work, we discuss a couple measures which are based on the so-called gains charts.

A.1. The gains chart

Gains chart displays the added gains (e.g., profitability) in using a predictive model versus a null model that assumes that all customers are the same. The X-axis represents the cumulative proportion of the population $X_i = 100 * i/n$. The Y-axis represents the cumulative percentage of the predicted quantity (e.g., purchase probability), or $Y_i = 100 * \sum_{j=1}^{i} y_j / \sum_{j=1}^{n} y_j$, where the observations are ordered according to the predicted values such that $\hat{y}_i \leqslant \hat{y}_{i+1}$. A typical gains chart is exhibited in Fig. A.1. We note that gains charts are similar to Lorenz curve in economics,

Two metrics, based on the gains chart, are typically used to assess how the predicted variable y differs from the null model:

- Maximum *lift* (ML), more commonly known as the Kolmogorov Smirnov (K–S) criterion (Lambert, 1993, Chapter 2), which is the maximum distance between the model curve and the null model. The K–S statistics has a distribution known as the D distribution (DeGroot, 1991, Chapter 9). A large ML indicates, in most problems, that the distribution of predicted variable y is different from the null model (see Fig. A.1). The D distribution can be approximated when the number of observation n is large. For large n, the NULL hypothesis, that the two distributions are the same, is rejected with a confidence interval if $\mathrm{ML} > D_{95} \approx \frac{1.36}{\sqrt{n}}$ (Churchill, 1999; Hodges, 1957).
- The *Gini coefficient* (Lambert, 1993, Chapter 2) which is calculated as the area between the model curve and the null model (the gray area in Fig. 1) divided by the area below the null model. A large Gini coefficient indicates, in most problems, that the distribution of predicted variable y is different from the null model.

We note that the metrics reflecting the true prediction quality is the gains chart. The lift and the Gini coefficients are summary measures that may

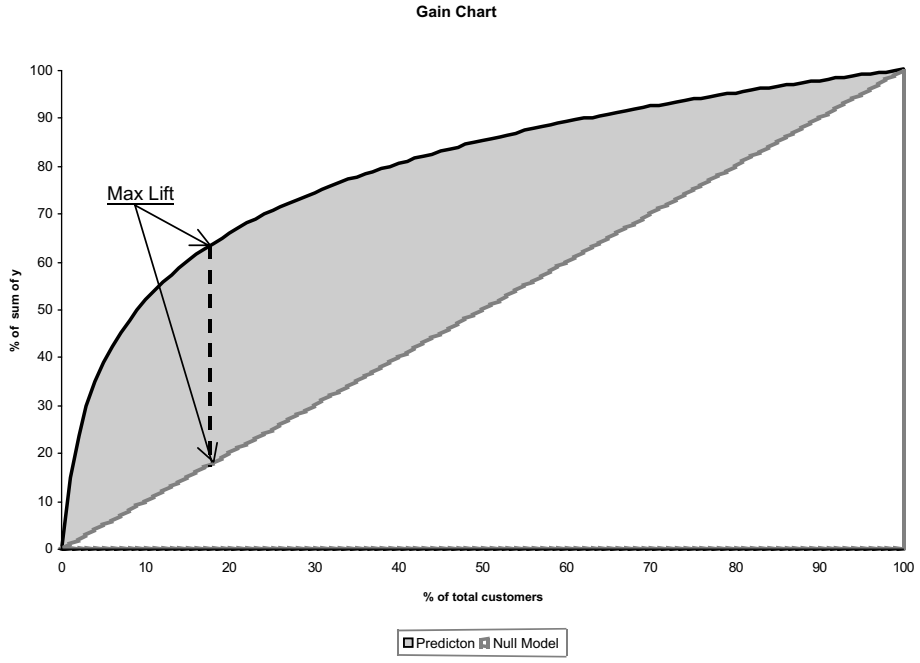

Fig. A.1. Gains chart: The null model is represented by the dashed line (45°) emanating from the origin. The model curve is given by the solid convex curve above it. The highest distance between the model and the null curve is the maximum lift (K–S statistics). The Gini coefficient is the ratio between the model gains (gray area) and the area under the null model.

not be consistent with one other. For example, it is possible to find two distributions where in one, the Gini coefficient is larger, and in the other the ML is larger.

Appendix B. Final SA and SWR predictors

The tables below present the predictors that made it to the final model in each application. Each predictor is identified by its name and followed by a short description. Predictors selected by both SA and SWR are marked with "=" sign, predictors selected only by SA are marked with "A" whereas predictors selected only by SWR are marked with "R".

Predictors in the "non-profit" data file

Predictor	In	Descript
CNDOL1	=	Latest contribution
CNDOL10	=	10th latest contribution
CNDOL2	=	2nd latest contribution

Predictors in the "non-profit" data file

Predictor	In	Descript
CNTIME1	=	Time since latest contribution
CNTMLIF	=	Time contribution lifetime
CONLARG	=	Largest contribution
CONTRFST	=	First contribution
CT1_A	=	Latest contribution type "A"
CT10_	=	10th latest contribution is " " ("blank")
CT4_M	=	4th latest contribution is "M"
CT9_A	R	9th latest contribution is "A"
CT9_C	R	9th latest contribution is "C"
MONTHFST	=	Time since first contribution
MONTHLRG	=	Time since large contribution

(*continued on next page*)

Appendix B (*continued*)

Predictors in the "non-profit" data file

Predictor	In	Descript
NONPRCOD_	=	No premium contact
REINCODE_R	=	Reinstate
RENTCODE_R	=	Rental exclusion
SEX_B	=	Gender = 'Both"
SEX_M	R	Gender = "Male"
SLTMLIF	=	Time solicited lifetime
SLTYPE_M	=	# Solicitation type "M"
ST1_O	=	Latest solicitation = "O"
ST10_	=	10th latest solicitation = " " ("blank")
ST10_O	=	10th latest solicitation = "O"
ST2_O	=	2nd latest solicitation = "O"
ST3_B	=	3rd latest solicitation = "B"
TYPE_A	=	# type "A" contributions
TYPE_O	=	# type "O" contributions

Predictors in the "specialty" data file

Predictor	In	Descript
CAT25	=	Purchased form cat. 25
CAT33	=	Purchased form cat. 33
CNVCAT25	=	First purchase from cat. 25
CNVCAT26	=	First purchase from cat. 26
CNVCAT33	R	First purchase from cat. 33
CRCPR17	=	Lifetime # of prom. 17
CRCPR35	=	Lifetime # of prom. 35
CRCPR50	=	Lifetime # of prom. 50
CRCPR65	=	Lifetime # of prom. 65
CRCPR85	=	Lifetime # of prom. 85
CRCPR93	=	Lifetime # of prom. 93
CVS5	=	First order >$300
FSTCLS2	R	1st order prod. from class 2
LSTYCLS3	=	Last order prod. from class 2
ORD165	=	Purchase in prom. 63 last year
ORD172	R	Purchase in prom. 72 last year
ORD185	=	Purchase in prom. 85 last year
ORD250	=	Purchase in prom. 50 2 years ago
ORD285	=	Purchase in prom. 85 2 years ago

Predictors in the "specialty" data file

Predictor	In	Descript
ORD301	R	Purchase in prom. 01 2 years ago
ORD350	=	Purchase in prom. 50 3 years ago
ORD385	=	Purchase in prom. 85 3 years ago
ORD400	R	Purchase in prom. 00 4 years ago
ORD485	=	Purchase in prom. 85 4 years ago
ORDCLS1	R	Total 5 years orders in cat. 1
ORDCLS3	=	Total 5 years orders in cat. 3
PRORD72	A	Lifetime orders in prom. 72
PRORD80	=	Lifetime orders in prom. 80
REC5	=	Last order placed 25–30 m′ ago
REC6	=	Last order placed 31–36 m′ ago
REC7	=	Last order more than 36 m′ ago
SALCLS3	=	Total 5 years orders in cat. 3
SSODCLS5	R	Total 12 months sales in cat. 5
TOTSAL1	=	Total sales last year
TOTSAL5	=	Total sales 5 years ago
ZIP2DIG_48	R	Zip: first 2 digit = 48

Predictors in the "gift" data file

Predictor	In	Descript
AMEXSLS	A	LTD AmerExp card orders
CHARGORD	A	LTD Credit card orders
CHARGSLS	A	LTD Credit card Dollars
DATEFP	=	Months since first purchase
DATELP	=	Months since last purchase
FALORD	=	LTD Fall orders
FORDORD	=	First season orders
GRP3S10	=	LTD purchases in ProdGrp 10
GRP3S11	=	LTD purchases in ProdGrp 11
GRP3S2	=	LTD purchases in ProdGrp 2
LORDORD	=	Latest season orders

Predictors in the "gift" data file

Predictor	In	Descript
LPURSEAS_1	=	Latest purchase season = 1
LPURSEAS_2	R	Latest purchase season = 2
LPURYEAR_0	=	Latest purchase Year = 0
LPURYEAR_1	=	Latest purchase Year = 1
LPURYEAR_2	=	Latest purchase Year = 2
LPURYEAR_4	=	Latest purchase Year = 4
LPURYEAR_5	=	Latest purchase Year = 5
LPURYEAR_6	=	Latest purchase Year = 6
LPURYEAR_7	=	Latest purchase Year = 7
LPURYEAR_8	=	Latest purchase Year = 8
LPURYEAR_9	=	Latest purchase Year = 9
MCVISSLS	A	LTD MS & VISA card Dollars
ORD2AGO	=	Orders 2 years ago
ORD3AGO	=	Orders 3 years ago
ORDHIST	=	LTD orders
ORDLYR	=	Orders last year
ORDTYR	=	Orders this year
PGRP11TY	=	This year purchased ProdGrp 11
PHONORD	R	LTD phone orders
PURSEAS	=	Seasons with purchase
PURYEAR	=	Years with purchase
SEASAD_1	=	Season added to file = 1
SEASFORD_2	=	Season of first order = 2
SPRORD	=	LTD Spring orders

Predictors in the "catalog" data file

Predictor	In	Descript
AGE	R	Age in years
DOL12MT	R	Div. T total $12 months
DRFMD	=	Div. D RFM pints
NUMCRED	R	# of credit cards used
ORDLTDD	=	Total LTD orders div. D
TOTAMT12	R	Total 12 month amount
TOTHSCRD	=	House credit (Yes = 1)
TOTORD24	=	Total 24 month orders

References

Akaike, H., 1974. A new look at the statistical identification model. IEEE Transactions on Automatic Control 19, 716–723.

Almuallim, H., Dietterich, T.G., 1991. Learning with many irrelevant features. In: Ninth National Conference on Artificial Intelligence. MIT Press, pp. 547–552.

Almuallim, H., Dietterich, T.G., 1994. Learning Boolean concepts in the presence of many irrelevant features. Artificial Intelligence 69, 279–306.

Anily, S., Federgruen, A., 1987. Simulated annealing methods with general acceptance probabilities. Journal of Applied Probability 24, 657–667.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 57, 289–300.

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271.

Blum, A.L., Rivest, R.L., 1992. Training a 3-node neural networks is NP-complete. Neural Networks 5, 117–127.

Cardie, C., 1993. Using decision trees to improve case-based learning. In: Proceedings of the Tenth International Conference on Machine Learning. Morgan Kaufmann Publishers, Inc., pp. 25–32.

Caruana, R., Freitag, D., 1994. Greedy attribute selection. In: Proceedings of the 11th International Conference on Machine Learning.

Churchill Jr., G.A., 1999. Marketing Research, seventh ed. The Dryden Press.

Clyde, M., George, E.I., 1999. Empirical Bayes estimation in wavelet nonparametric regression. In: Muller, P., Vidakovic, B. (Eds.), Bayesian Inference in Wavelet Based Models. Springer-Verlag, pp. 309–322.

Clyde, M., George, E.I., 2000. Flexible empirical Bayes estimation for wavelets. Journal of the Royal Statistics Society, Series B, 62, 681–698.

Cruz, J.R., Dorea, C.C.Y., 1998. Simple conditions for the convergence of simulated annealing type algorithms. Journal of Applied Probability 35, 885–892.

Dash, M., Liu, H., 1997. Feature selection methods for classifications. Intelligent Data Analysis: An International Journal 1 (3).

DeGroot, M.H., 1991. Probability and Statistics, third ed. Addison-Wesley, Reading, MA.

Devijver, P.A., Kitler, J., 1982. Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs, NJ.

Eiben, A.E, Aarts, E.H.L., Van Hee, K.M., 1991. Global convergence of genetic algorithms. In: Schwefel, H.-P., Männer, R. (Eds.), Parallel Problem Solving from Nature. Springer, Berlin and Heideberg, pp. 4–12.

Foster, D.P., George, E.I., 1994. The risk inflation criterion for multiple regression. Annals of Statistics 22, 1947–1975.

Foster, D.P., George, E.I., 2000. Calibration and empirical Bayes variable selection. Biometrica 87 (4), 731–747.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and bias/variance dilemma. Neural Computation 4 (1), 1–58.

George, E., McCulloch, R., 1993. Variable selection in Gibbs sampling. Journal of the American Statistical Association 88, 881–889.

Hancock, T.R., 1989. On the difficulty of finding small consistent decision trees. Unpublished Manuscript, Harvard University.

Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. Royal Statistical Society B (2), 190–195.

Hodges Jr., J.L., 1957. The significance probability of the Smirnov two-sample test. Arkiv for Matematik 3, 469–486.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Hurvich, C.M., Tsai, C.L., 1998. A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. Biometrika 85 (3).

Jain, A., Zongker, D., 1997. Feature selection evaluation, application and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2), 153–158.

John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 121–129.

Johnstone, I.M., Silverman, B.W., 1998. Empirical Bayes approaches to mixture problems and wavelet regression. Tech Report, University of Bristol.

Kasse, R., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–785.

Kira, K., Rendell, L.A., 1992. A practical approach to feature selection. In: Proceeding of the Ninth International Conference on Machine Learning. Morgan Kaufmann.

Kirkpatrik, S., Gelatt Jr., C., Vecchi, M., 1983. Optimization by simulated annealing. Science 220, 671–680.

Kohavi, R., John, G.H., 1998. Wrappers for feature subset selection. Artificial Intelligence 97 (1–2), 273–324.

Kohavi, R., Sommerfield, D., 1995. Wrappers for feature subset selection. Artificial Intelligence 97 (1–2), 273–324.

Koller, D., Sahami, M., 1996. Toward optimal feature selection. In: Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, July 1996, pp. 284–292.

Kononenko, I., 1994. Estimating attributes: Analysis and extensions of relief. In: Bergandano, F., Raedt, L.D. (Eds.), Proceedings of the European Conference on Machine Learning.

Lambert, P.J., 1993. The Distribution and Redistribution of Income. Manchester University Press.

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers.

Melab, N., Cahon, S., Talbi, E.-G., Duponchel, L., 2002. Parallel GA-based wrapper feature selection for spectroscopic data mining. International Parallel and Distributed Processing Symposium: IPDPS 2002 Workshops, April 2002.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Simulated Annealing. Journal of Chemical Physics 21, 1087.

Miller, A.J., 2002. Subset Selection in Regression, (2nd Edition). Chapman and Hall.

Mitchell, T.J., Beauchamp, J.J., 1998. Bayesian variable selection in linear regression (with discussion). Journal of the American Statistical Association 83, 1023–1036.

Motoda, H., Liu, H., 2002. Feature selection. Handbook of Data Mining and Knowledge Discovery. Oxford University Press, pp . 208–213.

Rudolph, G., 1994. Convergence analysis of canonical genetic algorithms. IEEE Transactions on Neural Networks, Special Issue on Evolutionary Computation 5 (1).

Schilmmer, L.C., 1993. Efficiently inducing determinations: A complete and systematic search algorithm that uses optimal pruning. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 284–290.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Shao, J., 1993. Linear model selection by cross-validation. Journal of the American Statistical Association 88, 486–494.

Siedlecki, W., Sklansky, J., 1988. On automatic feature selection. International Journal of Pattern Recognition and Artificial Intelligence 2, 197–220.

Tibshirani, R., Knight, K., 1999. The covariance inflation criterion for model selection. Journal of the Royal Statistical Society, Series B, Biometrika 85, 701–710.

van Laarhoven, P.J.J., Aarts, E.H.L., 1987. Simulated Annealing: Theory and Application. Kluwer, Boston.

Wei, C.Z., 1992. On predictive least squares principles. Annals of Statistics 29, 1–42.

Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association 93, 120–131.

Zheng, X., Loh, WY., 1997. A consistent variable selection criterion for linear models with high-dimensional covariates. Statistica Sinica 7, 311–325.