

Efficient Computer Experiment Based Optimization through Variable Selection

Dachuan T. Shih, Venkata L. Pilla, Seoung Bum Kim, Jay M. Rosenberger, and
Victoria C. P. Chen

Department of Industrial and Manufacturing Systems Engineering
The University of Texas at Arlington
Arlington, TX 76019 USA

Abstract

Variable selection has been widely used in regression data mining not only to select informative variables, but also to simplify the statistical model. A computer experiment based optimization approach employs design of experiments and statistical modeling to represent a complex objective function that can only be evaluated pointwise by solving an optimization subproblem. In large-scale applications, the number of variables is huge, and direct use of computer experiments would require an exceedingly large experimental design and, consequently, significant computational effort. Typically, a large portion of the variables have little impact on the objective; thus, there is a need to eliminate these before performing the complete set of optimization subproblem computer experiments. Ideally, variable selection would be conducted after a small number of computer experiment runs, likely fewer runs (n) than the number of variables (p). Conventional variable selection techniques cannot be applied in this “large p and small n ” problem. We explore the use of regression trees and a multiple testing procedure based on false discovery rate. Performance of the selected variables is measured using the coefficient of determination (R^2) and relative errors. Two real world applications are studied, an air quality stochastic dynamic program and an airline fleet assignment problem.

Keywords

data mining, variable selection, computer experiment, optimization.

1 Introduction

In recent years, variable selection has received considerable attention in various areas for which datasets with thousands of variables are available. These areas include signal/image processing, bioinformatics, process monitoring, and text mining. The main objective of variable selection is to identify a subset of variables that are most predictive or informative of a given response variable. Further, successful implementation of variable selection simplifies the entire modeling process and, thus, reduces computational and analytical efforts. Variable selection is particularly of interest when the number of candidate explanatory variables is large, and many redundant or irrelevant variables are thought to be present.

In this paper, the purpose of variable selection is to speed up a new class of large-scale optimization methods based on design and analysis of computer experiments (DACE, Chen, Tsui, Barton and Allen 2003). In DACE, a design is used to organize a set of computer experiment runs, so as to enable fitting of a statistical “metamodel” that approximates performance measure output from the computer experiment. The metamodel is then used for additional purposes, often a larger optimization task. In traditional DACE, the computer experiment is a simulation (Kleijnen 2005; Sacks et al. 1989). In DACE-based optimization, the computer experiment solves an optimization problem. This approach has been successfully applied for value function approximation in stochastic dynamic programming and Markov decision processes (Chen 1999; Chen et al. 1999; Chen, Günther and Johnson 2003; Tsai et al. 2004), where the computer experiment is an optimization that provides a point on the value function. More recently, a DACE-based approach has been developed for two-stage stochastic programming (Pilla 2006).

For modeling a performance function in optimization, such as an objective function or a value function, our research has focused on the use of multivariate adaptive regression splines (MARS, Friedman 1991; Tsai and Chen 2005; Shih et al. 2006). In large-scale optimization applications, the number of variables in a DACE-based approach can initially be in the hundreds or thousands. Although one could simply attempt a MARS approximation over these extremely high-dimensional spaces, typically, many variables have little affect on the performance measure. Thus, a data mining step to conduct variable selection is essential. The studies presented in this paper test the use of regression trees and a multiple testing procedure based on false discovery rate for variable selection.

2 Variable Selection

2.1 Variable Importance Scores from CART

Classification and regression trees (CART) developed by Breiman et al. (1984) have become a very popular data mining tool for supervised learning. The CART forward algorithm uses binary recursive partitioning to separate the variable space into rectangular regions based on similarity of the response values. In this paper, we utilized CART software from Salford Systems (www.salfordsystems.com). For variable selection, this software provides “variable importance scores.” The variable that receives a 100 score indicates the most influential variable for prediction, followed by other variables based on their relative importance to the most important one. However, there are some different options for calculating the scores, and selecting the threshold of the scores may be subjective. Further, this method tends to select an overly small number of variables. This motivates the development of an objective and systematic approach for variable selection.

2.2 Multiple Testing Procedure Based on False Discovery Rates

We begin with a brief introduction of the multiple testing procedure. Suppose through some modeling process, we have a collection of hypothesis tests and the corresponding p -values $\{p_i\}_{i=1}^n$, where p_i is the p -value of testing the null hypothesis and n is the number of variables. In the literature, it is standard to choose a p -value threshold τ and declare the variable v_i significant if and only if the corresponding p -value $p_i \leq \tau$. A common approach in multiple testing nowadays is the false discovery rate (FDR) procedure (Benjamini and Hochberg 1995) since it is well-known that we need to adjust the significance level when conducting multiple tests and that the conventional Bonferroni-corrected significance level is too conservative to pick true significant variables. The FDR is defined as the expected proportion of false positives among all the hypotheses rejected (Benjamini and Hochberg 1995). The general FDR-procedure to identify significant variables is as follows: Consider a series of hypotheses, p -values, and ordered p -values, denoted H_i , p_i , and $p_{(i)}$, respectively.

- Choose a fixed α , where $0 \leq \alpha \leq 1$.
- Find $\hat{i} = \max \left[i : p_{(i)} \leq \frac{i}{m} \cdot \frac{\alpha}{\pi_0} \right]$, where $\pi_0 (= \frac{m_0}{m})$ denotes the proportion of true H_i .
- If $\hat{i} \geq 1$, $\Omega = \{\text{All rejected } H_i \text{ with } p_i \leq p_{(\hat{i})}\}$ with $\text{FDR}(\Omega) \leq \alpha$.
If $\hat{i} = 0$, do not reject any hypothesis since $\Omega = \emptyset$.

In general, $\pi_0 = 1$ is the most conservative possible choice. Thus, we use $\pi_0 = 1$ in this paper. For more details of the choice of π_0 , refer to Efron (2004) and Storey and Tibshirani (2003).

2.2.1 FDR-based variable selection from regression trees

Generally, a conventional FDR procedure for variable selection requires a categorical response variable that separates the data into c groups, where c is the number of categories. For each predictor variable, we test for differences in the c samples, using a t -test or F -test. However, because the response variable generated by computer experiments is continuous in most cases, we need to categorize the original response. A mean or median value of the response variable can be used to separate the response variable into two groups, high and low, if the response surface is monotonic. In an air quality application presented later, we show the effectiveness of this simple grouping strategy. However, if the relationship between the response and the predictors is not monotonic, such that the separation by high and low values does not make sense, then alternate grouping strategies are needed. In order to address this problem, we use binary regression trees to partition the response observations into meaningful groups. An algorithm constructing binary regression trees partitions the space into two regions using the predictor variable and splitting-point that achieves the best improvement in fit. This partitioning process is repeated to one or both of these regions until a termination criterion has been reached. Based on the terminal nodes of regression trees, the response value can be separated into a certain number of groups, and an FDR procedure can be applied for variable selection. Note that for three or more groups, an analysis of variance (ANOVA) table is constructed for each predictor variable and its significance is tested using an F -test. This approach simultaneously takes advantage of regression trees and an FDR procedure. The possible drawback of this approach is that we may lose the original characteristics of a continuous response variable by grouping it. Moreover, too many categories in the response may lead to too many significant variables.

2.2.2 Inverse FDR

In order to maintain the continuous characteristic of a response variable in an FDR procedure, we propose a new FDR approach for variable selection, called inverse FDR. The main idea is to create a set of new variables that has the same number of original predictors by grouping the response variable based on the predictors and conducting an FDR procedure on these new variables. This is analogous to the resampling technique because each new variable is re-sampled from the original response based on each predictor variable. Inverse FDR is similar to the original FDR procedure, except that the hypothesis test is conducted on the continuous response grouped by each predictor variable, as opposed to testing the continuous predictor variables grouped by the response values. The setting and procedure for inverse FDR is as follows:

- For each predictor variable, divide the response variable into c groups based on the values of the predictor variable.
- For each predictor variable, conduct a statistical test (e.g., t -test, ANOVA F -test) on its corresponding set of response variable groups, and record the p -value. There will be one p -value for each predictor variable.
- Use the p -values to conduct an FDR procedure that identifies which predictor variables are statistically significant.

If the response surface is known to be convex or concave, a common occurrence in optimization, then inverse FDR with $c = 3$ groups should be sufficient. In general, more complex nonlinear structure can be captured with more groups.

3 Applications

3.1 Air Quality Stochastic Dynamic Program

The increasing concentrations of ground-level ozone in the urban (and often rural) atmosphere continues to be one of the major environmental issues today. Ground-level ozone is not emitted directly into the air, but is created by a complex series of reactions involving nitrogen oxides ($\text{NO}_x = \text{NO} + \text{NO}_2$) and volatile organic compounds (VOCs) in the presence of sunlight (Sillman et al. 1995). The primary sources of NO_x are power plants, automobiles, and industry. VOCs have both anthropogenic sources (cars, industry) and natural sources (vegetation). Therefore, in order to control ground-level ozone it is necessary to control emissions of NO_x and VOCs. Yang (2004) developed an ozone pollution decision-making framework for metropolitan Atlanta using a DACE-based stochastic dynamic programming approach. Atlanta, in particular, is “ NO_x -limited,” which means that targeting VOC emissions is not effective (Chameides et al. 1988). The objective of the decision-making framework was to identify the critical regions and time periods in which to reduce NO_x , so as to attain the EPA ozone limit. The initial set of variables for Atlanta considered NO_x reductions in 25 regions and 102 point sources over 5 time periods, *and* maximum ozone concentrations at 4 monitoring stations over 4 prior time periods. Thus, the total number of variables in the initial set of variables was 651. As part of the framework, the significant point sources were identified, reducing the number of point sources from 102 to 15 and the total number of variables from 651 to 216. Of the 216, 5 were eliminated for reasons related to inactivity of point sources in specific time periods. Hence, the analysis presented here was conducted on a set of 211 predictor variables. A Latin hypercube experimental design with 500 runs was used to collect data on ozone concentrations from the Atlanta Urban Airshed Model (U.S. EPA 1990). These data were used to evaluate the impact of the predictor variables on maximum ozone.

To see the potential for further dimension reduction, the following five cases of variable selection were examined: (i) none, (ii) forward-backward stepwise regression, (iii) FDR using p -values obtained based on a t statistic for each regression coefficient, (iv) FDR using p -values obtained based on two-sample t statistics, (v) FDR applied to a randomly selected subset of the observations (150), using p -values obtained based on two-sample t statistics. Cases (i), (ii), and (iii) use the original continuous response variables, while the response variables used in cases (iv) and (v) were categorized into two groups based on their median values. It is important to note that in case (v), the number of variables (211) is larger than the number of observations (150), i.e., the “large p and small n ” problem.

The numbers of variables selected for each case and the resulting coefficients of determination (R^2) are reported in Table 1. The FDR procedure significantly reduces the number of variables and maintains good prediction accuracy

Table 1: Variable selection performance for the air quality application.

Variable Selection Method	Number of Variables Selected	$R^2(\%)$
None	211	98.9
Forward-backward stepwise regression	49	98.6
FDR with continuous response	9	95.5
FDR with categorized response	13	96.4
FDR with categorized response, $p > n$	5	92.3

compared with the model constructed with all variables. Further, the result from case (v) demonstrates that FDR-based feature selection can efficiently handle the “large p and small n ” problem. In this air quality application, the FDR-based feature selection method with a simple grouping strategy (using median) seems to work well because the response surface, although technically nonlinear, has strong linear components. However, not all applications have strong linear components. In next section, we introduce other variable selection methods that can handle a non-monotonic response surface.

3.2 Robust Airline Fleet Assignment

The robust fleet assignment model addressed in Pilla (2006) uses a two-stage stochastic programming framework along with the Boeing concept of demand driven dispatch to swap crew-compatible aircrafts closer to departure, when most of the demand has been realized. Crew-compatible aircrafts have identical cockpits, allowing an airline to swap aircrafts without swapping crews. The two-stage formulation assigns crew-compatible aircrafts in the first stage, about 90 days prior to departure, so as to enhance the demand capturing potential of swapping in the second stage, about two weeks prior. The stochasticity of the demand is modeled by different demand scenarios in the second stage, and the average over the scenarios estimates the expected profit. The expected profit function is known to be concave.

Traditional two-stage stochastic programming uses a Benders’ approach or L-shaped method (Birge and Louveaux 1997); however, for large-scale problems, this can be slow to converge. Pilla (2006) developed a two-phase DACE approach to reduce the computation involved in conducting the optimization. The DACE Phase uses first-stage constraints in a multi-step process to construct an experimental design within the feasible region, then builds a statistical model that approximates the expected profit function in the first stage of the stochastic program. The Optimization Phase solves the two-stage problem using the DACE expected profit approximation instead of solving many second-stage subproblems in every iteration. This greatly speeds up the optimization, compared to Benders’, because the computation of the subproblems is shifted to the DACE Phase. However, further speedup may be achieved by conducting variable selection prior to the DACE Phase.

For a real airline network with 50 stations and 2358 legs, the DACE Phase reduced the decision space from 6537 to 1264 dimensions, and the multi-step process derived 141 initial extreme points, which were then expanded into 3562 design points in the feasible region. The second-stage subproblem is then solved for each of these design points. Among the 1264 decision variables, there are still many unnecessary ones that could be identified via variable selection, enabling a much smaller set of design points. Using only the subproblem solutions for the 141 initial extreme points, we studied five cases of variable selection: (1) none, (2) CART variable importance scores greater than 0.0, (3) FDR on 3 groups identified by CART, (4) FDR on 4 groups identified by CART, and (5) Inverse FDR. The resulting numbers of variables selected are shown in Table 2. For comparison purposes, a MARS approximation (with automatic stopping rule, Tsai and Chen 2005) for each of the 5 variable sets was fit using the 3562 design points. Relative errors were calculated using a validation data set of 1600 points, and the resulting boxplots are shown in Figure 1. It can be seen that all the variable selection methods produce approximations that are nearly as good as the case without variable selection. As another comparison, 36 variables were randomly selected and tested in the same manner. This resulted in a maximum relative error of 0.0727 and a median relative error of 0.0693, which are clearly larger than those in Figure 1.

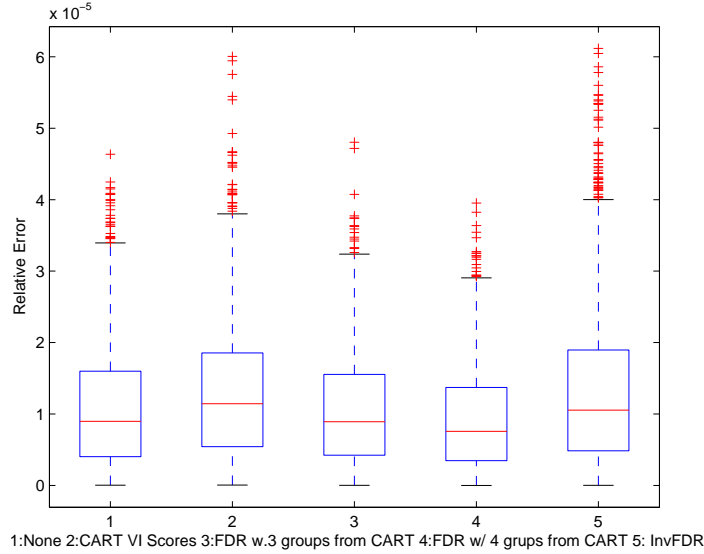


Figure 1: Boxplots of the relative errors from the five variable selection cases for the fleet assignment application.

Table 2: Variable selection results for the fleet assignment application.

Variable Selection Method	Number of Variables Selected
None	1264
CART Variable Importance Scores	36
FDR with 3 groups from CART	565
FDR with 4 groups from CART	820
Inverse FDR	476

4 Concluding Remarks

Several methods for handling the “large p and small n ” problem have been discussed and tested on two large-scale DACE-based optimization applications. The air quality application was seen to be a fairly straightforward application for FDR, given the strong linear components of the response surface. By comparison, the robust fleet assignment application has concave nonlinearity, and the decision variables have many 0 and 1 values. This motivated the inverse FDR approach, although all the methods tested yielded good approximations.

References

- Benjamini, Y. and Hochberg, Y.: 1995, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. B* **57**, 289–300.
- Birge, J. R. and Louveaux, F.: 1997, *Introduction to Stochastic Programming*, Springer, New York, New York.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J.: 1984, *Classification and Regression Trees*, Belmont, Wadsworth, California.

- Chameides, W. L., Lindsay, R. W., Richardson, J. and Kiang, C. S.: 1988, The rise of biogenic hydrocarbons in urban photochemical smog atlanta as a case-study, *Science* **241**, 1473–1475.
- Chen, V. C. P.: 1999, Application of mars and orthogonal arrays to inventory forecasting stochastic dynamic programs, *Computational Statistics and Data* **30**, 317–341.
- Chen, V. C. P., Günther, D. and Johnson, E. L.: 2003, Solving for an optimal airline yield management policy via statistical learning, *Journal of the Royal Statistical Society Series C*(52 Part 1), 1–12.
- Chen, V. C. P., Ruppert, D. and Shoemaker, C. A.: 1999, Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming, *Operations Research* **47**, 38–53.
- Chen, V. C. P., Tsui, K. L., Barton, R. R. and Allen, J. K.: 2003, *A Review of Design and Modeling in Computer Experiments*, Vol. 22 of *In Handbook of Statistics (C. R. Rao and Ravi Khattree, eds.)*, Elsevier Science, NY, pp. 231–261.
- Efron, B.: 2004, Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, *J. Am. Statist. Assoc.* **99**, 99–104.
- Friedman, J. H.: 1991, Multivariate adaptive regression splines (with discussion, *Annals of Statistics* **19**, 1–141.
- Kleijnen, J. P. C.: 2005, An overview of the design and analysis of simulation experiments for sensitivity analysis, *European Journal of Operational Research* **164**, 287–300.
- Pilla, V. L.: 2006, *Robust Airline Fleet Assignment*, PhD thesis, University of Texas at Arlington.
- U.S. EPA: 1990, User's guides for the urban airshed model, EPA450/490007AE.
- Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P.: 1989, Design and analysis of computer experiments (with discussion), *Statistical Science* **4**, 409–423.
- Shih, D. T., Chen, V. C. P. and Kim, S. B.: 2006, Convex version of multivariate adaptive regression splines, *Proceedings of the 2006 Industrial Engineering Research Conference*, Orlando, FL, USA.
- Sillman, S., Al-Wali, K., Marsik, F. J., Nowatski, P., Samson, P. J., Rodgers, M. O., Garland, L. J., Martinez, J. E., Stoneking, C., Imhoff, R. E., Lee, J. H., Weinstein-Lloyd, J. B., Newman, L. and Aneja, V.: 1995, Photochemistry of ozone formation in atlanta, ga: models and measurements, *Atmospheric Environment* **29**, 3055–3066.
- Storey, J. D. and Tibshirani, R.: 2003, Statistical significance for genomewide studies, *Proc. Natl. Acad. Sci.* **100**, 9440–9445.
- Tsai, J. C. C. and Chen, V. C. P.: 2005, Flexible and robust implementations of multivariate adaptive regression splines within a wastewater treatment stochastic dynamic program, *Quality and Reliability Engineering International* **21**, 689–699.
- Tsai, J. C. C., Chen, V. C. P., Beck, M. B. and Chen, J.: 2004, Stochastic dynamic programming formulation for a wastewater treatment decision-making framework, *Annals of Operations Research, Special Issue on Applied Optimization Under Uncertainty* **132**, 207–221.
- Yang, Z.: 2004, *A Decision-Making Framework for Ozone Pollution Control*, PhD thesis, University of Texas at Arlington.