Exploration of Multivariate Adaptive Regression Trees and its Evaluation

Masashi Goto
Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University.
Machikaneyama-cho 1-3.
Toyonaka, Osaka 560-8531, Japan.
gotoo@sigmath.es.osaka-u.ac.jp

Tomoyuki Sugimoto
Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University.
Machikaneyama-cho 1-3.
Toyonaka, Osaka 560-8531, Japan.
sugimoto@sigmath.es.osaka-u.ac.jp

Yasunobu Furukawa
KYOWA HAKKO KOGYO Co., LTD
Otemachi 1-6-1, Chiyoda, Tokyo
Tokyo, Japan

1. Introduction

In regression approach, we explore how a single or many variables influence a target variable. In particular, we may focus on three primary aspects: selection of variables, deviation from linearlity, and an introduction of nonlinear terms and additive terms. Then, as an approach for achieving these requirements, we can appeal to Multivariate Adaptive Regression Splines(MARS: Friedman, 1991). In MARS, the result applied to the model can be expressed as a hierarchical tree, and therefore the complicated nonlinear and interaction effects can be visualized. However, when the wrong effects are detected at the roots of the tree, unfortunately it is found that our interpretations are quite different from the phenomenon.

In this paper, we focus on to what degree MARS can reproduct the underlying model for the reification, and evaluate this reproductive performance through some case studies and small-scale simulation.

2. Simulation

We consider the underlying models as follows:

$$\begin{aligned} (1)\ Y &= 10\sin(x_1 x_2 \pi) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5 + \epsilon, \\ (2)\ Y &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon, \end{aligned}$$

where Y is a target variable, x_1, x_2, x_3, x_4,and x_5 are predictor variables, and β_1, β_2, and β_3 are regression coefficients. ϵ is assumed to be normally distributed with mean zero and variance σ^2. On the basis of above models, let the correlation coefficient for each Y, ρ be 0.1, 0.5, and 0.9, the variance of ϵ, σ^2 be 1, 1/4, and 1/16, and the sample size n be 100, 200, and 500. For all combinations of these factors, $n \times p$ matrix is generated from a uniform distribution U [0,1] such that each component of X is

$$x_{ij} = \begin{cases} z_{ij} & : \quad i = 1, 2, ..., n, \quad j = 1, \\ (1 - \rho_{1j}^2)^{1/2} z_{ij} + \rho_{1j} z_{ij} & : \quad i = 1, 2, ..., n, \quad j = 2, 3, 4, 5. \end{cases}$$

Under this design, we evaluated the influences of these factors on GCV by analysis of variance (ANOVA).

3. Result

Table 1 shows the result of ANOVA on GCV for model (1). As a result, the main effect of ρ and σ^2 were significant. The effect of ρ was the largest in terms of PVE (Proportion of Variation Explained by each source). It was noted that all interaction effects were very small. Similarly, the result for model (2) was that the main effect of ρ and σ^2 were highly significant.

Table 1. ANOVA on GCV : Result of model (1)

Source	d.f.	F-ratio	p-value	PVE(%)
n	2	0.825	0.4616	4.017
ρ	3	4.062	0.0331	29.652
σ^2	2	5.048	0.0256	24.565
$n\ \rho$	6	0.106	0.9941	1.547
$n\ \sigma^2$	4	0.142	0.9631	1.382
$\rho\ \sigma^2$	6	0.493	0.8018	7.205

REFERENCES

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**, 1-141.

Goto, M. & Matsubara, Y. (1979). Evaluation of ordinary ridge regression. *Bull. Math. Statist.*, **19**, 1-35.

RESUME

Nous ayons simule deux sortes de modeles en vue de estimer reproduction du laten modele par MARS.