대한산업공학회/한국경영과학회 2002 춘계공동학술대회
한국과학기술원(KAIST) 2002 년 5월 3일-4일

군집분석 기법과 단계별 회귀모델을 결합한 예측 방법

정일교, 전치혁
포항공과대학교 산업공학과

A Prediction Method Combining Clustering Method and Stepwise Regression

Il-gyo Chong, Chi-Hyuck Jun
Division of Mechanical and Industrial Engineering, POSTECH
E-mail: chig@postech.ac.kr

Abstract

A regression model is used in predicting the response variable given predictor variables. However, in case of large number of predictor variables, a regression model has some problems such as multicollinearity, interpretation of the functional relationship between the response and predictors and prediction accuracy. A clustering method and stepwise regression could be used to reduce the amount of data by grouping predictors having similar properties and by selecting the subset of predictors, respectively. This paper proposes a prediction method combining clustering method and stepwise regression. The proposed method fits a global model and local models and predicts responses given new observations by using both models. This paper also compares the performance of proposed method with stepwise regression via a real data example obtained in a steel process.

1. Introduction

Regression model provides an adequate and interpretable description of how the predictors affect the response. However with a large number of predictors, regression model often suffers from its poor prediction accuracy and interpretation. To overcome these drawbacks, many statisticians propose various methods: variable subset selection (best subset regression, forward stepwise selection, backward stepwise regression, stepwise regression and etc), coefficient shrinkage (ridge regression, the Lasso and etc), and methods using derived input directions (PCR, PLS and etc). These methodologies are described in detail [1]. Frank et al also discuss the pros and cons of these methodologies [2]. Variable subset selection has good points of producing an easy interpretable model by retaining a subset of the predictors and discarding the rest while the others has difficulty in interpretation because they predict responses with linear combinations of original predictors.

Clustering method is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes (dimensions). Objects within each cluster are more closely related to one another than objects assigned to different clusters. Clustering methods have been studied extensively in statistics. Sharma describes the details of clustering method [3]. In case of large number of predictors, clustering method can be used to reduce the dimensions of predictors by grouping predictors having similar properties.

In this paper, we assume a large number of predictors and one response. The proposed method predicts responses given new observations via a global model and local models. They are grouped and fitted by clustering method and stepwise regression. The way to divide into and fit a global model and local models will be explained in detail in the section 3. In the section 2, we will explore some basic concepts concerning clustering method and ten-fold cross validation to understand this paper. In the section 4, we compare the performance of the proposed method with stepwise regression through a real data example obtained in a steel process. Finally, section 5 concludes with some remarks.

2. Literature review

2.1. Clustering method

The objective of cluster analysis is to group objects into clusters such that each cluster is as homogeneous as possible with respect to the clustering variables. Current clustering techniques can be broadly classified into two categories: nonhierarchical and hierarchical. Given a set of objects and clustering criterion, nonhierarchical clustering such as K-means and K-medoids obtains a partition of the objects into clusters such that the objects in a cluster are more similar to each other than to objects in different clusters. A hierarchical clustering is a nested sequence of partitions. An agglomerative hierarchical clustering start by placing each object in its own cluster and then merge these atomic clusters into larger clusters until all objects are in a single cluster [3].

Various clustering methods are identified according to definition and evaluation of similarity. Euclidean, Minikowski, Mahalanobis distance are generally used similarity definitions. Centroid, nearest neighbor (single linkage), farthest neighbor

(complete linkage), average linkage and Ward's method are also generally used techniques to evaluate similarity between an object and a cluster. In this paper, we employ Euclidean distance and Ward's method to obtain local groups. The definition of Euclidean distance between objects i and j with p dimensions is as follows

$$D_{ij} = (\sum_{k=1}^{p} (X_{ik} - X_{jk})^2)^{1/2} \quad \ldots(1)$$

The Ward's method tries to group objects minimizing the total within group or within cluster sums of squares of defined similarities (Euclidean distance here) [3].

2.2. Ten-fold cross validation

The simplest and most widely used method for estimating prediction error is cross validation. Ten-fold cross validation works by dividing the training data randomly into ten equal parts. The learning method is fit to nine-tenths of the data, and the prediction error is computed on the remaining one-tenth. This is done in turn for each one-tenth of the data, and the ten prediction error estimates are averaged. Here are more details. Assume data of N observations and divide the data into 10 partitions having equal size. Then we have 10 partitions consisting of $(N/10)\times 9$ train observations and $N/10$ test observations. Denote by $\hat{f}^{(-i)}(x)$, $(i = 1,\ldots,10)$ the fitted function, computed with the i-th partition removed. Now, prediction error is defined as follows. In the case of training,

$$CV_i = 1/(9N/10) \sum_{k \in train\ set} (y_k - \hat{f}^{(-i)}(x_k))^2 \quad \ldots(2)$$

In the case of testing,

$$CV_i = 1/(N/10) \sum_{k \in test\ set} (y_k - \hat{f}^{(-i)}(x_k))^2 \quad \ldots(3)$$

And

$$CV_{avg} = \sum_{i=1}^{10} CV_i / 10 \quad \ldots(4)$$

$$CV_{std} = \sum_{i=1}^{10} (CV_i - CV_{avg})^2 / 9 \quad \ldots(5)$$

3. Proposed method

Suppose that there are *one* response and n predictors having m observations each. Let y_i, $(i = 1,\ldots,m)$ be i-th observation of a response and Y be $m\times 1$ response matrix composed of y_i. Let x_{ij}, $(i = 1,\ldots,m,\ j = 2,\ldots,n+1)$ be i-th observation of $(j-1)$-th predictor and X be $m\times(n+1)$ predictor matrix composed of x_{ij} and $x_{i1} = 1$. Summary of the proposed method are as follows.

3.1 Modeling algorithm

Step 1: Standardize predictor variables.

$$(x_{ij} - \bar{x}_j)/s_j$$
$$where\ \bar{x}_j = \sum_{i=1}^{m} x_{ij} / m$$
$$and\ s_j = \sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2 / (m-1), \quad \ldots(6)$$
$$(j = 2,\ldots,n+1)$$

Step 2: Fit a stepwise regression model of Y on X with *add1* and *remove1*, which are levels of significance for adding a predictor into a model and removing a predictor from a model, respectively. We define the $(n+1)\times 1$ coefficient matrix of this stepwise model as that of a global model denoted by $B^{g}_{(add1,remove1)}$. Then the hat matrix of a response by global model is as follows

$$\hat{Y} = XB^{g}_{(add1,remove1)} \quad \ldots(7)$$

And transform original response as follows

$$Z = Y - XB^{g}_{(add1,remove1)} \quad \ldots(8)$$

Step 3: Divide n predictors into n_g global predictors and n_l local predictors ($n = n_g + n_l$). If the global model includes a predictor, we consider this predictor as one of global predictors and local predictors, otherwise.

Step 4: Consider global predictors determined by step3 as cluster variables and divide observations of predictors into K groups by clustering method (In this paper, we adopt Euclidean distance and Ward's method in cluster method). Now let m_k be the number of observations assigned in k-th group, X_k be $m_k\times(n+1)$ matrix composed of x_{ij}s assigned in k-th group and Z_k be $m_k\times 1$ matrix composed of the observations of a transformed response which correspond with those of X_k, $(k = 1,\ldots,K)$. Note that cluster method in this step uses only global variables and the observations of n_g global predictors in each K groups become more similar than those of n_l local predictors.

Step 5: Fit a stepwise regression model of Z_k on X_k with two levels of *add2* and *remove2* and with global predictors excluded from a model in each K groups (i.e., coefficients of global predictors are all zeros). We define the $(n+1)\times 1$ coefficient matrix of this stepwise model as that of a local model in each K groups and denote this by $B^{l}_{k,(add2,remove2)}$. Then the hat matrix of a Z_k is as follows

$$\hat{Z}_k = X_k B^{l}_{k,(add2,remove2)},\ (k = 1,..,K) \quad \ldots(9)$$

3.2 Prediction algorithm without smoothing

Suppose new observation $P^T = (p_1,\ldots,p_n)$.

Step 1: Transform new observation as follows

$$(p_j - \bar{x}_j)/s_j,\ (j = 1,2,\ldots,n) \quad \ldots(10)$$

Step 2: Identify the group to which new observation is assigned by clustering method. Let the index of this

group be *c*.
Step 3: Predict a response given new observation as follows

$$Y_{new} \mid P = P^T B^g_{(add1,remove1)} + P^T B^l_{c,(add2,remove2)} \quad \cdots(11)$$

3.3 Prediction algorithm with smoothing
Step1~Step2: The same as described in 3.2.
Step3: Predict a response given new observation as follows

$$Y_{new} \mid P = P^T B^g_{(add1,remove1)} + \sum_{k=1}^{K} \alpha_k P^T B^l_{k,(add2,remove2)}$$
$$where \ \sum_{k=1}^{K} \alpha_k = 1 \ and \ \alpha_k \geq 0 \quad \cdots(12)$$

4. Example

4.1 Data description

In order to investigate the performance of the proposed method, we apply the proposed method to real data obtained in sequential steel process: steel-making → hot rolling mill →cold rolling mill.

Table 1. The notations and brief descriptions of predictors

Production Process	Process Variables	A Brief Description
Steel Making (SM)	Cu	The amount of copper
	Mn	The amount of manganese
	Nb	The amount of niobium
	Si	The amount of silicon
	Sol_Al	The amount of aluminum
	Ni	The amount of nitrogen
	V	The amount of vanadium
	Mo	The amount of molybdenum
	B	The amount of boron
	Cr	The amount of chromium
	N	The amount of nitrogen
	Ti	The amount of titanium
	C	The amount of carbon
Hot Rolling Mill (HM)	H_Line_Spd	Line speed in hot rolling mill process
	M2_Min	Time spent in the reheating furnace
	CT_Top	Temperature in the top part of coil while winding coil
	CT_Tail	Temperature in the tail part of coil while winding coil
	FT0_Top	Temperature before finishing mill
	FDT_Tail	Temperature after finishing mill
	M1_Min	Time spent in the heating furnace
Cold Rolling Mill (CM)	TCM_Rate	Pressure reduction ratio
	TM_Rate	Pressure reduction ratio in roughing mill
	HR_Thickness	The thickness of hot coil
	C_Line_Spd	Line speed in annealing process
	SS_Temp	Annealing temperature

The purpose of this study is to investigate the relationship between the hardness of steel (i.e., one response) and 25 process variables (i.e., 25 predictors). 25 predictors are selected through opinions of the experts and some statistical tests. The brief descriptions of predictors are shown in Table 1.

4.2 Results of stepwise regression

Before adopting the proposed method, we consider stepwise regression and estimate prediction error via ten-fold cross validation. In the result shown in Table 2 we obtain estimated cross validation error of 0.584.

Table 2. Ten fold cross validation error over train set and test set with *add1*=0.1 and *remove1*=0.1

	Train set	Test set
CV_1	0.462	0.663
CV_2	0.455	0.739
CV_3	0.492	0.424
CV_4	0.454	0.790
CV_5	0.443	0.876
CV_6	0.485	0.439
CV_7	0.512	0.300
CV_8	0.470	0.642
CV_9	0.496	0.448
CV_{10}	0.498	0.521
CV_{avg}	0.477	0.584
CV_{std}	0.023	0.186

4.3 Results of proposed method

4.3.1 How to determine the number of local models

To determine the complexity parameter *k* (the number of local models) of the proposed method, we consider the estimated prediction error curve shown in Figure 1 and obtain this over test and train set by ten-fold cross validation. In this example, *k*=5 seems appropriate due to the lowest cross validation error over test set.

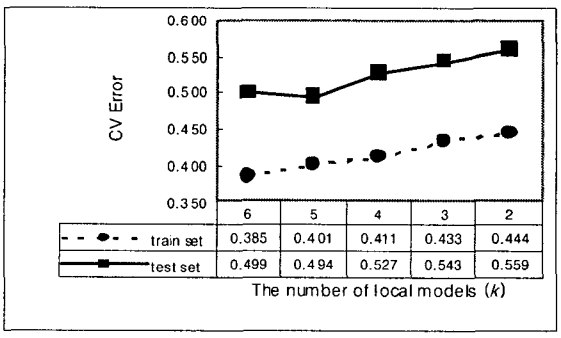

Figure 1. An estimated prediction error curve that shows how to determine the number of local models with *add1*=0.1, *remove1*=0.1, *add2*=0.15 and *remove2*=0.15

4.3.2 Relationship between global and local model

Consider the relationship between a response and 25 predictors. Table 3 shows the index of the tenth model in ten-fold cross validation with *add1*=0.1, *remove1*=0.1, *add2*=0.15, *remove2*=0.15 and k=5. In the Table 3, index '1' denotes a predictor included in a model. For example, the global behavior of steel hardness is explained well by Cu, V, M2_Min, FT0_Top and TCM_Rate. The local behavior could be divided into 5 groups each having identified properties. This embedded information is helpful for more insight to control steel hardness.

Table 3. The tenth model that shows which predictor is included in global model and local model with *add1*=0.1, *remove1*=0.1, *add2*=0.15, *remove2*=0.15 and k=5

Description		Global model	Local model				
			I	II	III	IV	V
SM	Cu	1	0	0	0	0	0
	Mn	0	0	0	0	0	0
	Nb	0	0	0	0	0	1
	Si	0	0	0	0	1	0
	Sol_Al	0	1	0	0	0	0
	Ni	0	0	0	0	1	0
	V	1	0	0	0	0	0
	Mo	0	0	0	0	0	0
	B	0	1	0	0	0	0
	Cr	0	0	0	0	0	0
	N	0	0	0	0	0	0
	Ti	0	0	1	0	0	1
	C	0	1	0	0	1	0
HR	M2_Min	1	0	0	0	0	0
	CT_Top	0	1	1	1	0	0
	CT_Tail	0	1	1	1	0	0
	H_Line_Spd	0	1	1	0	0	0
	FDT_tail	0	0	0	0	0	1
	FT0_Top	1	0	0	0	0	0
	M1_Min	0	0	0	0	1	0
CR	TCM_Rate	1	0	0	0	0	0
	HR_Thickne	0	0	0	0	0	0
	TM_Rate	0	0	0	0	0	1
	C_Line_Spd	0	0	1	1	0	0
	SS_Temp	0	1	1	0	0	0

4.3.3 Smoothing of local models

Let *Dist(A,B)* be the Euclidean distance between A and B. Let C_k be the average sums of squares of Euclidean distances between observations within k th local model (In short, C_k means the Centroid of k th local model). Now, we adopt smooth parameter α_k as follows

$$\alpha_k = (Dist(P, C_k) \Big/ \sum_{k=1}^{K} Dist(P, C_k))^q \quad \ldots(13)$$

Above q is shape parameter of smoothing weight. In Table 4, q was set to 7 and this results in about 0.01 decrease of estimated prediction error compared with without smoothing.

Table 4. The effect of adopting smoothing weight with *add1*=0.1, *remove1*=0.1, *add2*=0.15, *remove2*=0.15, k=5, and q=7

	Without smoothing		With smoothing	
	Train set	Test set	Train set	Test set
CV_1	0.399	0.561	0.389	0.576
CV_2	0.388	0.735	0.390	0.780
CV_3	0.422	0.151	0.414	0.150
CV_4	0.424	0.829	0.424	0.738
CV_5	0.349	0.757	0.344	0.742
CV_6	0.394	0.382	0.393	0.371
CV_7	0.386	0.203	0.386	0.233
CV_8	0.454	0.434	0.433	0.382
CV_9	0.396	0.506	0.395	0.506
CV_{10}	0.396	0.383	0.402	0.416
CV_{avg}	0.401	0.494	0.397	0.489
CV_{std}	0.028	0.230	0.024	0.218

5. Conclusion

We propose a prediction method combining clustering method and stepwise regression. This method divides observations of predictors into some local groups having homogeneous observations of some predictors and heterogeneous observations of the others. Those are considered as global predictors explaining global behavior of a response and these as local predictors explaining local behavior. Both are independently used to fit global and local model by stepwise regression. Eventually, a response given new observation of predictors will be predicted by both global and local model. A real example showed that the proposed method results in 16.27% decrease of estimated prediction error compared with typical stepwise regression.

The proposed method can be improved by adopting alternative smoothing parameter and clustering algorithm. We leave these future works.

6. Reference

[1] Hastie, Trevor et al. (2001), *The Elements of Statistical Learning: Data Mining Inference and Prediction*, New York: Springer

[2] Frank, Ildiko E. and Friedman, Jerome H. (1993), "A Statistical View of Some Chemometrics Regression Tools", *Technometrics, VOL. 35, No.2, 109-135*

[3] Sharma, Subhash (1996), *Applied Multivariate Techniques*, John Wiley & Sons, Inc.