한국지능시스템학회 논문지 2008, Vol. 18, No. 6, pp. 876-880

Support Vector Regression을 이용한 이상치 데이터분석

An Outlier Data Analysis using Support Vector Regression

전성해

Sung-Hae Jun

청주대학교 바이오정보통계학과

요 약

주어진 데이터에서 대부분의 다른 관측치들에 비해 지나치게 크거나 작은 관측치를 이상치라고 한다. 이상치는 몇 가지 원인에 의해 발생한다. 이상치를 포함한 데이터의 분석결과는 이 값을 포함하지 않은 경우와 크게 달라질 수 있다. 일반적으로 이상치는 탐지를 통하여 찾아내어 제거한 후에 데이터분석을 수행한다. 하지만 사기탐지, 네트워크 침입 등의 데이터 마이닝 분야에서는 이상치가 중요한 정보를 포함하고 있기 때문에 반드시 포함하여 데이터분석을 수행하여야 한다. 본 논문에서 다루는 회귀모형에서는 기존의 단순, 다중 회귀분석은 이상치에 대하여 안정된 모형을 구축하기 어렵기 때문에 표준화 잔차 또는 스튜던트화된 잔차를 이용하여 이상치를 찾아내고 제거한 후의 데이터분석 수행을 추천한다. 본 논문에서는 회귀모형에서 이상치를 포함하여 효과적으로 데이터분석을 수행할 수 있는 한 방법으로 Vapnik이 제안한 통계적 학습이론에 기반한 Support Vector Regression(SVR)을 이용하였다. 인공 데이터를 생성한 모의실험 결과 기존의 회귀모형에 비해 SVR의 향상된 결과를 확인할 수 있었다.

Abstract

Outliers are the observations which are very larger or smaller than most observations in the given data set. These are shown by some sources. The result of the analysis with outliers may be depended on them. In general, we do data analysis after removing outliers. But, in data mining applications such as fraud detection and intrusion detection, outliers are included in training data because they have crucial information. In regression models, simple and multiple regression models need to eliminate outliers from given training data by standadized and studentized residuals to construct good model. In this paper, we use support vector regression(SVR) based on statistical learning theory to analyze data with outliers in regression. We verify the improved performance of our work by the experiment using synthetic data sets.

Key Words : 이상치, Support Vector Regression, 이상치 분석, 단순회귀모형, 다중회귀모형

1. 서 론

사기탐지(fraud detection), 침입탐지(intrusion detection) 등의 데이터 마이닝 분야에서 이상치(outlier)는 제거되어야 할 대상이 아니라 전체 데이터에 대한 중요한 정보를 포함하고 있는 개체(object)로 인식된다[1]. 주어진 학습데이터에서 대부분의 다른 관측치들에 비해 지나치게 크거나 작은 개체를 이상치라고 부른다[1]. 이상치는 몇 가지 원인에 의해 발생한다. 이상치를 포함한 데이터의 분석결과는 이 값을 포함하지 않은 경우와 크게 달라질 수 있다. 때문에 일반적으로 이상치를 찾아내어 제거한 후에 모형을 통한 데이터분석을 수행한다. 하지만 경우에 따라서는 이상치가 전체 데이터에 대한 중요한 정보를 담고 있기 때문에 데이터분석 수행과정에서 반드시 포함해야 할 경우가 발생한다. 대표적인 지도학습 방법인 회귀모형(regression model)에서 전통적인 단순(simple), 다중(multiple) 회귀분석은 이상치에 대하여 안정된 모형을 제공하지 못하기 때문에 표준화 잔차(standardized residual) 또는 스튜던트화된 잔차(studentized residual)를 이용하여 이상치를 찾아내고, 제거한 후에 데이터분석의 수행을 추천한다[2,3]. 본 논문에서는 이와 같은 회귀모형에서 이상치를 포함한 데이터를 효과적으로 분석할 수 있는 하나의 방법으로 Vapnik이 제안한 통계적 학습이론(statistical learning theory)에 기반한 Support Vector Regression(SVR)을 제안한다[4,5,6]. SVR은 주어진 데이터공간(data space)의 개체들을 커널함수(kernel function)를 이용하여 고차원(high dimension)의 형상공간(feature space)로 사상(mapping)시킨 후에 이 형상공간에서 데이터분석을 수행한다[5]. 데이터공간에서 고차원의 형상공간으로 사상되는 과정에서 전체 데이터에 대한 이상치의 영향은 줄어들게 된다. 본 논문의 제안 방법에 대한 성능평가를 위하여 인공 데이터(synthetic data)를 생성하여 모의실험을 수행한 결과 이상치를 포함한 데이터의 경우에 기존의 회귀모형에 비해 SVR에 의한 데이터분석이 더 우수한 결과를 제공함이 확인되었다.

접수일자 : 2008년 5월 16일
완료일자 : 2008년 7월 30일

2. 이상치 데이터

이상치는 주어진 데이터 중에서 다른 관측치에 비하여 지나치게 크거나 작은 것을 의미한다[7]. 네트워크 침입탐지와 같은 데이터 마이닝 분야에서 이상치는 그 자체로 중요

한 정보를 포함할 수 있기 때문에 이상치를 포함한 데이터 분석이 가능한 방법에 대한 연구가 필요하게 된다. 일단 발생한 이상치의 탐지방법은 크게 모델기반(model based), 근접성기반(distance based), 그리고 밀도기반(density based)의 3가지가 있다[1]. 모델기반 접근방식은 주로 통계적 측도를 사용하고 근접성기반과 밀도기반 접근방식들은 각각 K-nearest neighbor와 밀도함수에 의존한다. 일반적으로 이상치가 발생하는 원인은 여러 가지가 있다. 상이한 데이터 그룹으로부터의 관측, 데이터 측정이나 기타 수집 및 입력 오류 등에 의한 이상치인 경우에는 당연히 찾아내어 제거한 후에 분석하면 되겠지만 이상치가 정상적으로 발생되었고 전체 데이터에 대하여 중요한 정보를 가지고 있는 관측치라고 판단되면 학습데이터에 포함하여 분석해야 한다. 이 때 기존의 회귀모형은 다음 그림과 같은 문제점이 발생한다.

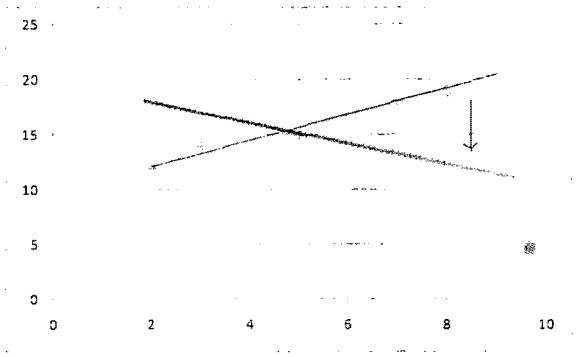

그림 1. 이상치에 의한 회귀방정식의 변형
Fig. 1. Changing regression by outlier

원형으로 표시된 관측치가 없다면 나머지 다이아몬드 형태의 관측치들만의 회귀방정식을 구할 수 있다. 하지만 원형으로 표시된 이상치가 전체 데이터에 포함되면 회귀직선이 화살표에서 표시한 직선으로 이동하게 된다. 이렇게 되면 모형의 성능이 떨어지게 된다. 즉 모든 데이터에 대한 실제값과 예측값의 차이가 크게 발생하게 된다. 이러한 문제점을 해결하기 위하여 본 논문에서는 이상치가 포함된 데이터공간에서 고차원 형상공간으로 모든 데이터를 사상시켜서 형상공간에서 데이터분석이 이루어지는 SVR 모형을 사용한다.

3. SVR을 이용한 이상치 포함 데이터의 분석

일반적으로 이상치의 탐지에 대한 많은 기준들이 제안되고 있다. 모든 기준들이 같은 결론을 제시하지는 않지만 나름대로 효과적인 방안을 제시하고 있다. 본 논문에서는 마하라노비스 거리(Mahalanobis' distance)와 상자그림(box plot)을 이용하여 이상치에 대한 탐지를 수행하였다. 마하라노비스 거리는 데이터 내의 변수 간의 상관관계를 포함한 유클리드 거리(Euclidean distance)에 대한 일반적인 확장이다. 다음은 두 개의 변수 X와 Y의 마하라노비스 거리에 대한 정의이다[1,7].

$$MD(X,Y)=(X-Y)\Sigma^{-1}(X-Y)' \tag{1}$$

Σ^{-1}은 변수들 간의 분산-공분산행렬(variance-covariance matrix)에 대한 역행렬을 나타낸다. 기존의 이상치 탐지와 마찬가지로 본 논문에서도 $\pm 3\sigma$를 벗어나는 영역에 속한 데이터 개체를 이상치로 판정한다. 일반적으로 표준정규분포(standard normal distribution)를 사용하게 되면 σ의 값은 1이 된다.

상자그림은 통계학 분야에서 연속형 데이터의 요약에서 기본적으로 구하는 5숫자 요약(five number summary)에 해당되는 최소값, 제1사분위수(Q1), 중앙값(median), 제3사분위수(Q3), 최대값을 이용하여 그려진다. 중앙값을 기준으로 Q1과 Q3를 이용하여 상자를 그린 후에 Q3 값에서 Q1 값을 뺀 IQR(inter-quartile range) 값을 이용하여 다음과 같이 상한(upper bound; UB)과 하한값(lower bound; LB)을 계산한다[8].

$$UB=Q_3+1.5IQR \tag{2}$$

$$LB=Q_1-1.5IQR \tag{3}$$

상자그림을 이용한 이상치 탐지는 위의 상한과 하한을 벗어나는 관측치 개체로 정의된다. 본 논문에서는 마하라노비스 거리측도와 상자그림을 통하여 이상치의 탐지가 확인되면 다음에서 설명되는 SVR을 이용하여 이상치를 포함하는 데이터공간의 모든 개체들을 커널함수를 통하여 고차원의 형상공간으로 사상시킨 후, 이 형상공간에서 데이터의 분석모형을 구축한다. SVR의 이론적 기반이 되는 Vapnik이 제안한 통계적 학습이론은 분류(classification), 예측(prediction), 그리고 군집(clustering)를 위한 구체적 분석모형인 SVM(support vector machine), SVR, 그리고 SVC(support vector clustering)의 3가지 학습모형을 갖추고 있다[5]. SVR과 SVC는 모두 통계적 학습이론 중 가장 먼저 제안되었던 SVM에서 확장되었다. SVM에서 목표변수(target variable) Y와 입력벡터(input vector) X로 구성된 데이터집합 D는 다음과 같은 표현된다[5,6].

$$(x_i,y_i)_{i=1}^{l},\quad x_i\in R^N,\quad y_i\in\{-1,1\} \tag{4}$$

일반적으로 주어진 데이터공간에서 서로 다른 클래스를 정확히 분류하는 초평면(hyperplane)을 찾는 것은 매우 제한적이다. 이러한 문제를 해결하기 위하여 SVM에서는 입력공간을 더 높은 차원의 형상공간으로 사상시키고, 형상공간에서 최적의 초평면을 찾는다. $z=\psi(x)$를 데이터공간 R^N에서 형상공간 Z로의 사상 ψ를 갖는 형상공간벡터로 표현하면, (w,b)의 쌍으로 이루어진 다음의 초평면을 구해야 한다.

$$w\cdot z+b=0 \tag{5}$$

위의 초평면식을 구하면 다음 함수에 의해 각 x_i를 분류한다.

$$\begin{cases}(w\cdot z_i+b)\geq 1, & \text{if } y_i=1\\ (w\cdot z_i+b)\leq -1, & \text{if } y_i=-1\end{cases}\quad i=1,2,\cdots,l \tag{6}$$

선형 분류가능(linearly separable) 집합 D는 이진 클래스를 갖는 학습 데이터의 사영(projection)들 사이의 마진(margin)을 최대화 하는 유일한 초평면을 구한다. S가 선형 분류가능이 아니면 음이 아닌 여유변수(slack variable) ξ_i를 도입하여 다음과 같이 식 (6)을 일반화한다.

$$y_i(w \cdot Z_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, l \tag{7}$$

식 (7)에서 ξ_i는 식 (6)을 만족하지 않는 x_i이다. $\sum_{i=1}^{l} \xi_i$는 오분류(misclassification)의 양을 나타내는 측도로서 고려된다. 따라서 최적 초평면을 구하는 문제는 아래의 문제에 대한 해(solution)가 된다.

$$\begin{aligned} &\text{minimize} \ \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i \\ &\text{subject to} \ y_i(w \cdot z_i + b) \geq 1 - \xi_i \end{aligned} \tag{8}$$

정규화상수(regularization constant) C는 조정모수(control parameter)이다. 이 상수값의 조정을 통하여 모형의 정확성(accuracy)과 복잡성(complexity) 사이의 균형을 맞출 수 있다. 식 (8)에서 최적 초평면을 찾는 것은 다음의 라그랑지 변환(Lagrangian transformation)을 통하여 풀 수 있다.

$$\begin{aligned} &\text{maximize} \ W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j z_i \cdot z_j \\ &\text{subject to} \ \sum_{i=1}^{l} y_i \alpha_i = 0 \qquad 0 \leq \alpha_i \leq C, \ i = 1, 2, \cdots, l \end{aligned} \tag{9}$$

여기서 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)$는 식 (7)의 제약조건과 관련된 음이 아닌 라그랑지 승수(multiplier) 벡터이다. 정규화상수 C와 함께 SVM에서 중요하게 고려되는 것은 데이터공간에서 형상공간으로의 사상을 담당하는 커널함수(kernel function)이다. 일반적으로 SVM에서는 다음 표와 같이 다항학습기계(polynomial learning machine(PLM)), 방사기저함수(radial basis function(RBF)), 그리고 다층 퍼셉트론(multi layer perceptron(MLP))을 사용한다[9].

표 1. 커널함수의 종류
Table 1. Types of kernel function

커널종류	함수식	커널모수
PLM	$(x^t x_i + 1)^p$	p
RBF	$e^{\left(-\frac{1}{2\sigma^2}\|x - x\|^2\right)}$	σ^2
MLP	$\tanh(\beta_1 x^t x_i + \beta_0)$	β_0, β_1

위 표는 SVM의 커널함수와 함수식 그리고 각 커널함수에서 사용되는 커널모수를 나타내고 있다. PLM에서는 다항함수의 차수 p가 커널모수이고 RBF에서는 분산 σ^2가 커널모수가 된다. MLP에서는 편이(bias) β_0와 기울기 β_1이 커널모수이다.

회귀(regression) 문제를 해결하기 위하여 사용되는 SVM을 SVR이라 한다. SVR은 SVM과 같은 이론구조를 가지며 추가적으로 다음과 같은 ϵ-insensitive 손실함수(loss function)를 사용한다[5,6].

$$L(d,y) = \begin{cases} |d-y| - \epsilon, & \text{for } |d-y| \geq \epsilon \\ 0 & , o.w. \end{cases} \tag{10}$$

SVR도 SVM과 마찬가지로 커널모수와 정규화상수를 주관적으로 결정해야 한다. 커널모수와 정규화상수는 SVR 성능에 직접적으로 영향을 준다. 본 논문에서는 격자탐색(grid searching)을 통하여 커널함수와 정규화상수의 최적값을 구한다.

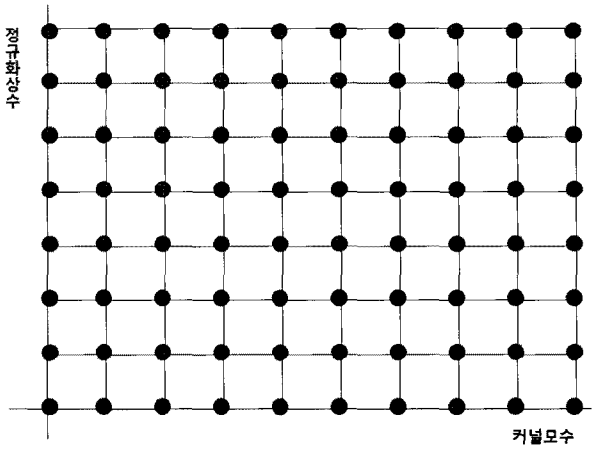

그림 2. 격자탐색을 통한 SVR의 모수선택
Fig. 2. Parameter selection using grid searching

위의 그림에서 격자의 간격을 좀 더 조밀하게 하면 더 상세한 해공간의 탐색이 가능하다. 하지만 해공간의 크기에 비례하여 컴퓨팅 비용이 추가적으로 필요하게 된다. 최근에는 격자탐색에 비해 빠르고 보다 정확한 커널모수와 정규화상수의 탐색을 위하여 진화 컴퓨팅(evolutionary computing)을 이용한 SVM, SVR, 그리고 SVC 모형이 제안되고 있다[9,10,11]. 하지만 이 방법들은 추가적으로 진화연산 알고리즘(evolving algorithm)에 대한 고려가 이루어져야 하기 때문에 본 논문에서는 격자탐색을 이용하였다.

4. 실험 및 결과

제안방법에 대한 성능평가를 위하여 본 논문에서는 인공데이터를 생성하여 모의실험을 수행하였다. 실험을 위한 데이터분석은 R-project를 이용하였다[12]. 인공데이터 생성은 다음과 같은 확률모형에 의한 난수를 생성하였다[13,14].

$$Y = f(X) + \epsilon \tag{11}$$

위 난수발생모형에서 Y는 반응변수이고 X는 설명변수벡터이다. ϵ는 오차항이다. 본 논문에서는 반응변수는 1개로 고정하고 설명변수를 1개인 경우(단순회귀)와 2개 이상인 경우(다중회귀)로 구분하여 모의실험 데이터를 생성하였다. 확률모형에서 사용되는 확률분포는 일반적으로 대부분의 데이터에서 가정하는 정규분포를 사용하였다. 첫 번째 모의실험데이터는 다음과 같은 단순회귀모형의 경우로서 모두 1000개의 개체를 생성하였다.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{12}$$

위식에서 오차항은 모든 관측치가 동일한 표준정규분포에서 생성된 것으로 가정하고 이와 같은 가정 하에서 오차항 데이터를 생성하였다. 다음 그림은 마하라노비스 거리측도와 상자그림을 통하여 인공적으로 생성된 데이터 내에 포함된 이상치의 분포를 보여주고 있다.

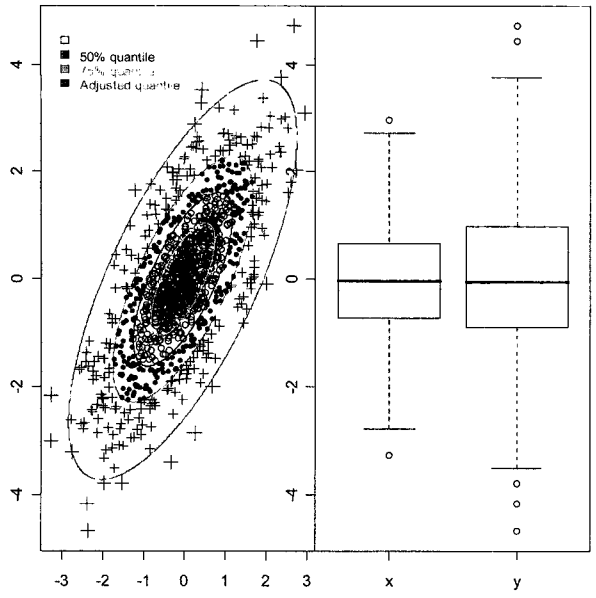

그림 3. 단순회귀 데이터의 이상치 탐지
Fig. 3. Outlier detection of simple regression

위 그림의 왼쪽은 마하라노비스의 거리를 통하여 ±3 범위내의 데이터 개체와 이 범위를 벗어나는 개체를 보여주고 있다. 3개의 타원 중 ±3 범위의 경계를 나타내는 가장 바깥의 타원을 벗어난 개체들을 볼 수 있다. 즉 마하라노비스의 거리측도를 통하여 실험을 위하여 인공적으로 생성된 데이터 집합에 이상치가 포함되어 있음을 알 수 있다. 또한 위 오른쪽의 상자그림을 통하여 주어진 데이터의 상한과 하한을 벗어난 데이터 개체인 이상치를 확인할 수 있다. 두 번째 모의실험 데이터는 다변량 정규분포(multivariate normal distribution)로부터 생성된 인공 데이터를 이용한다. 평균벡터(mean vector)는 다음과 같다.

$$\mu = (1.12, 12.21, 1.98) \tag{13}$$

마찬가지로 분산-공분산 행렬은 다음과 같다.

$$\Sigma = \begin{pmatrix} 0.9 & 2 & 0 \\ 2 & 5 & 0.55 \\ 0 & 0.55 & 2.98 \end{pmatrix} \tag{14}$$

위의 평균벡터와 분산-공분산 행렬을 이용하여 한 개의 반응변수 Y와 2개의 설명변수 X1과 X2를 생성하였다. 본 논문에서는 1000개의 난수 중에서 한 개를 임의의 이상치로 대체하여 실험하였다. 즉 Y, X1, 그리고 X2의 1000번째 데이터 개체를 각각 10, 25, -5로 대체하였다. 다음 그림은 Y와 X1 그리고 Y와 X2와의 마하라노비스 거리와 상자그림을 통한 이상치의 탐지과정이다.

앞의 단순선형회귀의 경우와 마찬가지로 위의 그림들의 마하라노비스 거리와 상자그림을 통하여 이상치의 존재를 확인할 수 있다. 다음 표는 이상치를 포함한 2개의 모의실험 데이터를 이용하여 실험한 결과이다. 실험결과에서 비교되는 모형들 간의 성능평가를 위한 측도로서 본 논문에서는 대부분의 예측모형에서 사용되는 MSE(mean square error) 특도를 사용하였다. 다음은 MSE에 대한 정의이다[7].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

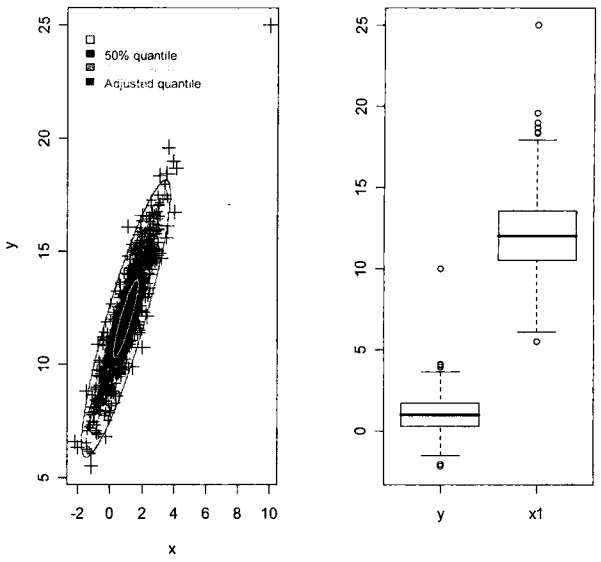

그림 4. 다중회귀 데이터의 이상치 탐지 1 (Y-X1)
Fig. 4. Outlier detection of multiple regression 1 (Y-X1)

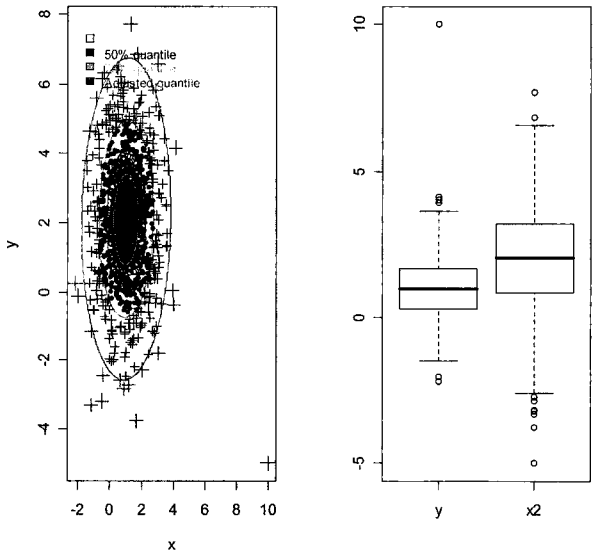

그림 5. 다중회귀 데이터의 이상치 탐지 2 (Y-X2)
Fig. 5. Outlier detection of multiple regression 1 (Y-X2)

위 식에서 y_i는 실제값이고 $\hat{y}_i$는 예측값이다. 즉 MSE는 실제값과 예측값의 차이의 제곱합을 전체 데이터의 개수(n)로 나눈 값이다. 이 값이 작을수록 우수한 모형이 된다.

표 2. 인공데이터 실험결과: MSE
Table 2. Result of synthetic data: MSE

Data	Regression	SVR
단순회귀모형	simple	c=8, G=0.5
	1.48	0.90
다중회귀모형	multiple	c=3, G=0.5
	0.26	0.18

위 결과를 통하여 단순회귀모형과 다중회귀모형의 모든

데이터의 실험에서 SVR에 의한 결과의 성능이 우수하게 나타남을 알 수 있다. SVR의 정규화상수(c)와 커널모수(G)의 값은 각각 격자 탐색을 통하여 결정된 최적모수값들이다. 결과를 통하여 SVR에 의한 이상치 포함 회귀 데이터의 분석이 과대적합(overfitting) 문제에도 적절한 해결책을 보여줄 것으로 보인다.

5. 결론 및 향후 연구과제

본 논문에서는 일반적으로 데이터에 포함된 이상치가 분석에 반드시 포함되어야 하는 경우에 이상치를 포함한 데이터분석의 효과적 수행을 위하여 Vapnik이 제안한 통계적 학습이론에 기반의 SVR을 제안하였다. 이상치를 포함한 데이터공간의 모든 개체들을 커널함수를 이용하여 고차원의 형상공간으로 사상시킨 후에 이 형상공간에서 데이터분석을 수행하여 이상치의 영향을 감소시켰다. 2개의 모의실험 데이터를 이용한 실험을 통하여 일반적으로 사용되는 회귀모형에 비해서 SVR 모형이 향상된 성능을 확인하였다. 본 논문에서는 연속형 데이터의 이상치만을 고려 대상으로 하였다. 범주형 데이터에서의 이상치 데이터분석은 별도의 주관적 고려가 이루어져야 할 뿐만 아니라 연속형 데이터에서의 이상치 탐지 기준과는 다른 측도가 필요하기 때문에 본 논문에서는 다루지 않았다. 향후 연구과제로서 이 부분에 대한 연구가 진행될 것이다.

참 고 문 헌

[1] 용환승, 나연묵, 박종수, 송현우, 이민수, 이상준, 최린 역, *데이터 마이닝*, 인피니티북스, 2007.

[2] 박성현, *회귀분석* 제3판, 민영사, 2007.

[3] R. H. Myers, *Classical and Modern Regression with Applications*, Duxbury, 1989.

[4] C. J. Burges, "A Tutorial on Support Vector Machine for Pattern Recognition", *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[5] S. Haykin, *Neural Networks A Comprehensive Foundation*, Prentice Hall, 1999.

[6] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[7] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.

[8] 류기열, 박일렬, 최승두 역, *앤더슨의 통계학*, 한올출판사, 2007.

[9] Sung-Hae Jun, "A Co-Evolutionary Computing for Statistical Learning Theory", *International Journal of Fuzzy Logic and Intelligent Systems*, Vol. 5 No. 4, pp. 281~285, December 2005.

[10] 전성해, "차분진화 기반의 Support Vector Clustering", *한국퍼지 및 지능 시스템학회 논문집*, 제17권 제5호, pp.679-683, 2007.

[11] Sung-Hae Jun, Kyung-Whan Oh, "A Competitive Co-Evolving Support Vector Clustering", *Lecture Note in Computer Science* (LNCS, ICONIP'2006), vol. 4232, pp. 864-873, Springer-Verlag, 2006.

[12] R-Project www.r-project.org

[13] W. L. Martinez, A. R. Martinez, *Computational Statistics Handbook with MATRAB*, Chapman & Hall, 2002.

[14] S. M. Ross, *Simulation*, Academic Press, 1997.

저 자 소 개

전성해(Sung-Hae Jun)

1993년 : 인하대 통계학과(학사)

1996년 : 인하대 통계학과(이학석사)

2001년 : 인하대 통계학과(이학박사)

2007년 : 서강대학교 컴퓨터공학과 (공학박사)

2003년~현재 : 청주대학교 바이오정보통계학과 조교수

관심분야 : 진화연산, 통계적학습이론, 신경망

Phone : 043-229-8205

Fax : 043-229-8432

E-mail : shjun@cju.ac.kr