한국통계학회논문집
2011, 18권, 2호, 171–177

DOI: 10.5351/CKSS.2011.18.2.171

서포트 벡터 기계를 이용한 이상치 진단

서한손[a], 윤 민[1,b]

[a]건국대학교 응용통계학과, [b]부경대학교 통계학과

요 약

실생활에서 얻어지는 자료에서 근사함수를 구성하기 위하여 모델링을 하기 전에 측정된 원자료로부터 이상치를 제거하는 것이 필요하다. 기존의 이상치 진단의 방법들은 시각화나 최대 잔차들을 이용해왔다. 그러나 종종 다차원의 입력자료를 가지는 비선형 함수에 대한 이상치 진단은 좋지 않은 결과를 얻었다. 다차원 입력자료를 갖는 비선형 함수에 대한 전형적인 서포트 벡터 회귀에 기초한 이상치 진단방법들은 좋은 수행능력을 얻어지지만, 계산비용이나 모수들의 보정 등의 실질적인 문제점들을 가지고 있다. 본 논문에서 계산비용을 감소하고 이상치의 문턱을 적절히 정의하는 서포트 벡터 회귀를 이용한 이상치 진단의 실질적인 방법을 제안한다. 제안한 방법을 실제 자료들에 적용하여 타당성을 보일 것이다.

주요용어: 이상치 진단, 서포트 벡터 회귀, 실질적인 접근법.

1. 서론

실생활의 자료의 함수근사는 통계학과 기계학습 등의 분야에서 가장 기본적인 문제이다. 공산품 설계에 대한 반응표면분석 방법과 측정 장비들에 대한 캐리브레이션 곡선 등과 같은 다양한 산업분야에서 응용되어 왔다. 출력 시스템들이 종종 다차원 입력자료들을 갖는 비선형의 특징을 가지고 있기 때문에 선형회귀분석과 같은 전통적인 통계분석 방법들은 요구되는 수행능력에 도달하기가 어려웠다. 게다가 실생활에서 얻어지는 자료들은 노이즈와 이상치들을 포함한다. 그러한 오염된 실제 자료들은 결과 모형의 수행능력을 감소시키게 된다. 그러므로 근사모형에 대한 정확한 추정을 위하여 노이즈와 이상치들의 효과를 감소시킬 필요가 있다.

노이즈는 전원 공급, 압력 그리고 전류의 불안정 등에 의해 일어난다. 노이즈 수준의 범위는 측정 시스템에 의해 영향을 받는다. 이와 같은 노이즈를 감소시키기 위한 방법들로는 평활법, 과적합을 방지하기 위한 로버스트 회귀 등이 전형적으로 사용되어 왔다. 이상치들은 장치의 시동을 걸 때, 전원의 차단, 운전상의오류나 전송오류 등에 의하여 일어난다. 이상치들은 비정상적인 현상들이기 때문에 모형을 구축하기 전에 자료의 나머지 부분들로부터 이상치들을 제거해야 한다. 이상치들을 진단하기 위한 기존의 방법들은 시각화와 최대 오차 에러(the maximum residual error)를 최대화하는 방법들이 사용되어 왔다. 비록 단일한 입력을 가지는 경우에는 시스템에 대하여는 자료들의 구조의 시각화가 유용하지만 다차원의 입력자료를 가지는 시스템의 경우에는 적절하지 않다. 또한, 최대 오차 에러 접근법은 추정함수 자체가 이상치에 영향을 받기 때문에 다른 자료들로부터 멀리 떨어져 있는 자료점들을 진단하기 어렵다. 한편, 이상치에 비교적 덜 민감한 로버스트 부분 최소제곱회귀(robust partial least squares regression) 방법은 분석학(chemometrics)의 응용에서 Pell (2000)이 사용되었으나 단지 선형의 관련성을 갖는 경우에만 적용하였다.

이 논문은 2010년도 건국대학교 학술진흥연구비 지원에 의한 논문임.

[1] 교신저자: (608-737) 부산시 남구 대연3동 599-1, 부경대학교 통계학과, 교수. E-mail: myoon@pknu.ac.kr

최근에 서포트 벡터 회귀(support vector regression; SVR, Hastie 등, 2009; Lahiri와 Ghanta, 2009)는 기존의 방법들보다 특히, 모수를 쉽게 보정하고 둔감함수(insensitivity function)를 갖는 강건성 그리고 커널함수들이 갖는 사전 지식의 이용 등과 같은 유의한 장점들로 인해 공업의 응용에서 사용되기 시작했다. 서포트 벡터 회귀는 또한 다차원 입력을 가지는 비선형 함수에서 이상치 진단의 문제들에 적용할 수 있다 (Jordaan과 Smits, 2004; Dufrenois 등, 2009). 비록 이러한 접근방법들은 서포트 벡터 회귀의 장점들을 이용하지만, 모수들의 보정과 마찬가지로 높은 계산 비용에서의 어려움 때문에 실제적으로 사용하는 데는 어려움이 있다. 이와 같은 사실은 측정된 자료들이 빠르고 명백해야하는 비전문 사용자들이 있는 공업의 응용분야에서 심각한 단점들이다.

이러한 단점들을 해결하기 위하여 기존에 사용하는 표준적인 서포트 벡터 회귀방법보다 개선된 μ-ε-서포트 벡터 회귀를 사용하여 이상치 진단을 위한 실질적인 방법을 제안한다 (Nakayama 등, 2009).

2. 표준적인 서포트 벡터 회귀를 이용한 이상치 진단

이 절에서, 표준적인 서포트 벡터 회귀 (Vapnik, 1999)를 이용하여 기존의 이상치 진단방법에 대하여 설명한다. 우선 훈련자료 집합으로 ℓ개의 입력 벡터들과 대응하는 스칼라 출력값 y를 고려하자. 서포트 벡터 회귀의 목적은 입력벡터 $\boldsymbol{x}$와 대응하는 출력값 y사이의 함수관계를 추정하는 것이다

$$y = \boldsymbol{w}^t \phi(\boldsymbol{x}) + b.$$

여기서 함수 $\phi(\boldsymbol{x})$는 임의의 특징공간(feature space)상에서의 비선형 함수이고, $\boldsymbol{w}$와 b는 각각 특징공간상에서의 가중치 벡터와 바이어스(bias)이다. 주어진 자료집합 $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, \ell$에 대한 서포트 벡터 회귀의 공식은

$$\begin{aligned} \underset{\boldsymbol{w},\, b,\, \xi_i,\, \xi_i'}{\text{minimize}} \quad & \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + \frac{C}{\ell}\sum_{i=1}^{\ell}(\xi_i + \xi_i') \\ \text{subject to} \quad & \boldsymbol{w}^t\phi(\boldsymbol{x}_i) + b - y_i \le \varepsilon + \xi_i, \\ & y_i - \boldsymbol{w}^t\phi(\boldsymbol{x}_i) - b \le \varepsilon + \xi_i', \quad \xi_i,\, \xi_i' \ge 0, \;\; i = 1, \ldots, \ell \end{aligned} \tag{2.1}$$

이다. 여기서 C는 모형의 복잡성과 손실들 사이의 트래드-오프(trade-off)를 보정하는 정규화 모수이고, ε은 ε-둔감 손실함수의 모수이고 ξ_i, ξ_i'들은 slack 변수들이다. 1차 공식 (2.1)에서 목적함수는 손실함수 ε에서 surplus 오차인 평균 slack 변수들을 최소화한다.

식 (2.1)의 쌍대공식은 아래의 볼록 2차문제와 같이 표현 된다

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{\ell}(\alpha_i' - \alpha_i)y_i - \varepsilon\sum_{i=1}^{\ell}(\alpha_i' + \alpha_i) \\ \text{subject to} \quad & \sum_{i=1}^{\ell}(\alpha_i' - \alpha_i) = 0, \quad 0 \le \alpha_i' \le \frac{C}{\ell}, \;\; 0 \le \alpha_i' \le \frac{C}{\ell}, \;\; i = 1, \ldots, \ell. \end{aligned} \tag{2.2}$$

여기서 $K(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^t\phi(\boldsymbol{x}')$는 커널함수이다. 각각의 라그랑쥬 승수 α_i, α_i'들은 상계(upper bound) 제약식 C/ℓ를 가짐에 주의하자.

이상치 진단방법에 기반한 표준적인 서포트 벡터 회귀 (Jordaan과 Smits, 2004)는 식 (2.2)를 풀어서 얻어지는 라그랑쥬 승수들의 성질을 이용한다. 즉, 식 (2.3)의 Karush-Kuhn-Tucker(KKT) 조건들에서

보조 slackness 조건들은 식 (2.1)로부터 얻어지고, 만일 자료점이 위로 유계인 라그랑쥬 승수를 가지지 않으면 이에 대응하는 slack 변수는 0이 되고 결국 자료점은 이상치로 의심되지 않는다.

$$\left\{-\boldsymbol{w}^t\phi(\boldsymbol{x}_i) - b + y_i + \varepsilon + \xi_i\right\}\alpha_i = 0, \qquad \left\{\boldsymbol{w}^t\phi(\boldsymbol{x}_i) + b - y_i + \varepsilon + \xi_i'\right\}\alpha_i' = 0, \tag{2.3}$$
$$\xi_i\left(\frac{C}{\ell} - \alpha_i\right) = 0, \qquad \xi_i'\left(\frac{C}{\ell} - \alpha_i'\right) = 0.$$

그러므로 자료점들의 상계에서 라그랑쥬 승수들 α_i, α_i'를 가지는 자료점들은 이상치의 후보점들로서 간주할 수 있다. 일반적으로, 다수의 자료점들은 상계를 갖는 라그랑쥬 승수들을 가지기 때문에 여러 개의 후보점들 중에서 실제 이상치를 찾아내야 한다. 실제 이상치는 여러 개의 다른 모수값들 ε에서 (2.2)를 이용하여 여러 번의 최적화 계산을 수행한 후에 의심된 이상치의 후보점들 중에서 가장 높은 빈도를 포함하는 하나로 진단된다. 이 과정은 더 이상 이상치가 진단되지 않을 때까지 반복하거나 여러 번의 최적화 계산을 통하여 얻은 제곱근평균제곱오차(root mean square error) 시행착오(trial and error)에 의하여 사전에 정의된 이상치의 문턱(threshold)보다 적게 되는 경우까지 반복하게 된다.

실제 자료를 가지고 응용을 할 때 이상의 접근방법을 적용할 때 다음과 같은 문제점들이 발생한다. 첫째, 이상치의 진단은 최적화 계산을 수차례의 반복이 요구되므로 여러 개의 이상치들을 가지는 대용량의 자료에 대하여 높은 계산비용이 요구된다. 둘째, 이상치의 문턱값을 어떻게 정의하느냐가 명확하지 않기 때문에 정확한 진단을 위하여 시행착오가 요구된다.

3. μ-ε-서포트 벡터 회귀에 기초한 이상치 진단

예측능력을 개선하기 위하여 Nakayama 등 (2009)은 평균 slack 변수들 대신에 모수 μ를 가지는 최대 slack 변수를 최소화하는 서포트 벡터 회귀 공식을 제안하였다. 본 논문에서 기존의 표준적인 서포트 벡터 회귀 방법의 어려움을 극복하기 위하여 모수 μ를 이용하여 μ-ε-서포트 벡터 회귀를 아래와 같이 얻을 수 있다.

$$\begin{aligned} \underset{\boldsymbol{w},\, b,\, \xi_i,\, \xi_i'}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \mu(\xi_i + \xi_i') \\ \text{subject to} \quad & \boldsymbol{w}^t\phi(\boldsymbol{x}_i) + b - y_i \le \varepsilon + \xi_i, \\ & y_i - \boldsymbol{w}^t\phi(\boldsymbol{x}_i) - b \le \varepsilon + \xi_i', \quad \xi_i,\, \xi_i' \ge 0, \;\; i = 1, \ldots, \ell. \end{aligned} \tag{3.1}$$

μ-ε-서포트 벡터 회귀 1차공식 (3.1)에서 단지 최대오차를 가지는 자료점들의 slack 변수들만 계산하고 식 (2.1)의 표준적인 서포트 벡터 회귀에서 사용된 평균 slack 변수들은 이용하지 않는다.

식 (3.1)로부터 라그랑쥬 함수는

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 + \mu(\xi + \xi') - \sum_{i=1}^{\ell}\alpha_i(-\boldsymbol{w}^t\phi(\boldsymbol{x}_i) - b + y_i + \varepsilon + \xi) - \sum_{i=1}^{\ell}\alpha_i'(\boldsymbol{w}^t\phi(\boldsymbol{x}_i) + b - y_i + \varepsilon + \xi') - (\eta\xi + \eta'\xi') \tag{3.2}$$

이고 여기서 쌍변수들$(\alpha_i, \alpha_i', \eta, \eta')$은 양의 제약식을 만족해야 한다.

식 (3.1)의 쌍대공식은 1차식의 변수들 $\boldsymbol{w}$, b, ξ, ξ' 각각에 대하여 L을 편미분하면 아래와 같이 유도되고

$$\begin{aligned} \underset{\alpha_i,\, \alpha_i'}{\text{maximize}} \quad & -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{\ell}(\alpha_i' - \alpha_i)y_i - \varepsilon\sum_{i=1}^{\ell}(\alpha_i' + \alpha_i) \\ \text{subject to} \quad & \sum_{i=1}^{\ell}(\alpha_i' - \alpha_i) = 0, \quad \sum_{i=1}^{\ell}\alpha_i \le \mu, \quad \sum_{i=1}^{\ell}\alpha_i' \le \mu, \quad 0 \le \alpha_i, \quad 0 \le \alpha_i', \quad i = 1, \ldots, \ell, \end{aligned} \tag{3.3}$$

여기서 둔감도 모수 ε과 정규화 모수 μ는 앞에서 정의되었다.

식 (3.1)에 대한 KKT 조건들은

$$\left\{-\boldsymbol{w}^t\phi(\boldsymbol{x}_i) - b + y_i + \varepsilon + \xi_i\right\}\alpha_i = 0, \qquad \left\{\boldsymbol{w}^t\phi(\boldsymbol{x}_i) + b - y_i + \varepsilon + \xi_i'\right\}\alpha_i' = 0, \tag{3.4}$$

$$\xi_i\left(\mu - \sum_{i=1}^{\ell}\alpha_i\right) = 0, \qquad \xi_i'\left(\mu - \sum_{i=1}^{\ell}\alpha_i'\right) = 0$$

에 의해 주어진다.

라그랑쥬 승수의 합이 상계 μ에 도달할 때, 영 아닌 양수의 라그랑쥬 승수를 가지는 모든 자료점들은 동일한 최대오차를 가진다. 그러므로, 이런 자료점들은 실제 이상치 자료일 것이다. 최적화 이론에서 라그랑쥬 승수들은 목적함수에 대응하는 부등식 제약식의 민감도를 나타낸다고 잘 알려져 있다 (Mangasarian, 1969). 결국, 이상치가 존재할 때, 최대 라그랑쥬 승수를 가지는 자료점은 가장 적절한 이상치로 볼 수 있다. 식 (3.3)의 문제로부터 각각의 라그랑쥬 승수는 상계에 의해 제약되지 않으므로 결코 여러 개의 자료점들은 동일한 위로 유계인 라그랑쥬 승수를 가지지 않는다.

한편 라그랑쥬 승수들의 합이 상계 μ의 아래쪽에 있을 때 모든 자료는 공차 ε보다 작은 에러를 가지거나 이상치가 존재하지 않는다. 따라서 라그랑쥬 승수들의 합은 이상치 진단 절차의 종료에 이용된다.

μ-ε-서포트 벡터 회귀에 기초한 제안한 이상치 진단 알고리즘은 아래와 같다.

Step 1: μ-ε-서포트 벡터 회귀를 계산한다.

Step 2: 가장 큰 α_i, α_i'를 찾는다.

Step 3: 자료집합에서 $\boldsymbol{x}_i$와 y_i를 제거한다.

Step 4: 제거된 자료집합을 이용하여 2 단계와 3 단계를 반복한다.

Step 5: α_i또는 α_i'의 합이 μ와 같지 않으면 종료한다.

이와 같은 접근법을 사용하여 이상치를 진단하는 동안에 각각의 반복횟수는 단지 한 번의 최적화 계산이 요구된다. 그러므로 계산비용은 표준적인 서포트 벡터 회귀를 이용한 방법보다 훨씬 낮게 된다. 또한 이상치들은 다른 잡음자료로부터 분리되는 고정된 공차(tolerance) ε를 이용하여 이상치 문턱을 적절하게 정할 수 있다.

4. 타당성 검정

제안한 이상치 진단 접근법의 효율성을 보이기 위하여 우선 아래의 예제를 고려하자.

$$y_i = (\sin 2\pi x_i)^2 + \tau_i + \theta_i, \tag{4.1}$$

여기서 $\boldsymbol{x}_i$, $i = 1, \ldots, 100$는 $[0, 1]$인 균등분포로부터 생성된 자료이고, τ_i는 $N(0, 0.05)$인 정규분포로부터 생성된 잡음이고, θ_i는 세 개의 이상치를 나타낸다. 즉, $\theta_i = [0.5, -0.5, 0.5, 0, 0, \ldots, 0]$이다.

이 예제에서 사용하는 모수로서 $r = 0.65$, $\mu = 1000$ 그리고 $\varepsilon = 0.15$를 갖는 가우지안 커널을 이용한다. 여기서 ε은 잡음의 분포에서 표준편차의 세배에 기초한 공차로 정의한다.

$$K(\boldsymbol{x}_i, \boldsymbol{x}_i') = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_i'\|^2}{2r^2}\right) \tag{4.2}$$

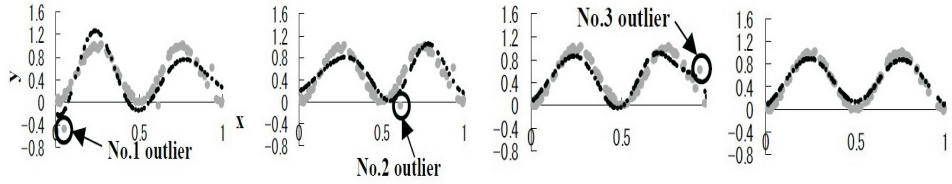

그림 1: *(왼쪽에서 오른쪽 방향): 각 반복에서 실제자료(회색 점들)와 추정된 자료(검은색 점들)*

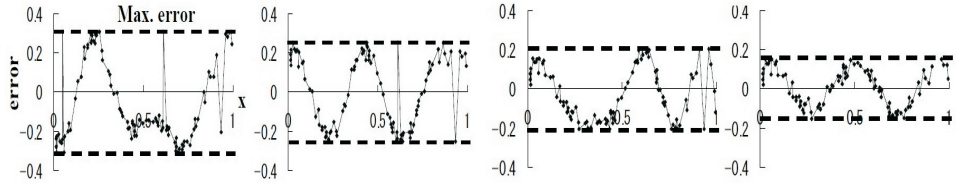

그림 2: *(왼쪽에서 오른쪽 방향): 각 반복에서 실제자료와 추정된 자료 사이의 오차 에러*

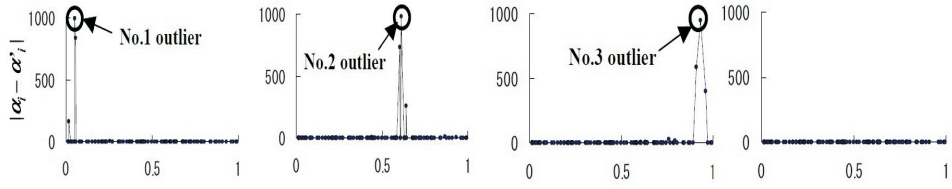

그림 3: *(왼쪽에서 오른쪽 방향): 각 반복에서 라그랑쥬 승수들의 절대값*

표 1: 통풍저항 자료에 대한 기존 방법과 제안한 방법의 결과

제안한 방법	(μ-ε-SVR)	기존의 방법	(표준 SVR)
반복횟수	제거된 자료	반복횟수	제거된 자료
1	No.21	20	No.3, No.4, No.21
1	No.4	20	No.1
1	No.3	20	-
1	No.1		
1	-		
총 반복횟수	5번	총 반복횟수:	60번
총 계산시간	0.9sec.	총 계산시간:	9.1sec.

그림 1은 각 반복단계에서 오차 에러와 대응하는 라그랑쥬 승수의 절대값의 추정된 결과를 나타낸다. 이 그림으로부터 네 번의 반복 후에 세 개의 모든 이상치들이 모두 정확하게 제거됨을 확인할 수 있다. 반복횟수의 증가에 따라 추정된 결과의 정확도가 개선된다. 최종 결과의 최대 오차 에러는 위에서 주어진 바와 같이 공차 0.15와 동일하다. 또 다른 예제는 이전의 많은 연구에서 연구되어진 통풍저항 자료이다 (Brownlee, 1965). 통풍저항 자료는 암모니아가 질산으로 산화에 대한 공장의 가동을 서술하는 자료이다. 이 자료는 하루에 측정된 4차원의 21개의 관측치로 구성되어 있다.

입력변수들은 조업률, 주입구의 냉각수 온도, 그리고 산성농도이고 반응변수는 통풍저항이다. 과거의 연구에서 1, 3, 4 그리고 21번째 자료들은 이상치로 간주되었다 (Rousseeuw와 Baxter, 1987).

제안한 방법을 통풍저항 자료에 적용하였다. 여기서 제안한 방법에서 사용한 모수들은 μ = 1000와

ε = 2.0을 가지고 적용하였다. 이전의 많은 연구들에서 통풍저항 자료에 대하여 이상치 진단에 있어서 선형적인 접근법으로 성공적인 결과를 보였기 때문에 식 (4.3)에서 보여지는 선형 커널을 사용하였다.

$$K(\boldsymbol{x}_i, \boldsymbol{x}_i') = \boldsymbol{x}_i \boldsymbol{x}_i' + 1. \tag{4.3}$$

위의 표에서 보는 바와 같이 제안한 방법을 적용하여 각 반복에서 각각의 이상치를 제거하는데 성공적이었다. 이에 반하여 표준적인 서포트 벡터 회귀를 이용한 기존의 방법은 더 많은 반복횟수와 계산시간이 요구되었다.

두 예제들을 통하여, 이러한 결과들로부터 제안한 방법이 기존의 방법보다 더 작은 반복횟수를 통하여 비선형이고 다차원 자료에서 정확하게 이상치들을 제거하는 것으로 나타났다. 더욱이 이 방법은 이상치 문턱으로 측정 시스템 사용에 관한 정보를 이용하고 있다.

5. 결론

본 논문에서는 μ-ε-서포트 벡터 회귀를 이용하여 이상치를 진단하는 실질적인 방법을 제안하였다. 이 방법은 기존의 접근방법에 비하여 계산비용을 감소시키고 이상치 문턱을 좀 더 정확하게 정의할 수 있기 때문에 장점을 가진다. 제안한 접근법의 효율성은 예제들을 통하여 검토되었다.

추후의 과제로 좀 더 효율적인 이상치 진단으로 다중 이상치들을 동시에 진단하는 문제를 생각할 수 있다. 다중 이상치 진단은 하나의 이상치가 다른 이상치를 숨기는 소위 가면화효과(masking effect) 때문에 어려운 문제로 생각된다. 또한, 제안한 방법과 기존의 대안적인 학습방법들을 개선하여 결합시키는 문제를 생각해 볼 수 있다.

참고 문헌

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering,* 2nd Ed., Wiley.

Dufrenois, F., Colliez, J. and Hamad, D. (2009). Bounded influence support vector regression for robust single-model estimation, *IEEE Transactions on Neural Networks,* **20**, 1689–1705.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, 2nd Ed., Springer, New York.

Jordaan, E. M. and Smits, G. F. (2004). Robust outlier detection using SVM regression, In *Proceedings of International Joint Conference on Neural Networks,* 2017–2022.

Lahiri, S. K. and Ghanta, K. C. (2009). Hybrid support vector regression and genetic algorithm technique–Novel approach in process modeling, *Chemical Product and Process Modeling,* **4**, Article 4.

Mangasarian, O. L. (1969). *Nonlinear Programming*, McGraw-Hill, New York.

Nakayama, H., Yun, Y. B. and Yoon, M. (2009). *Sequential Approximate Multiobjective Optimization Using Computational Intelligence*, Springer-Verlag, Berlin Heidelberg.

Pell, R. J. (2000). Multiple outlier detection for multivariate calibration using robust statistical technique, *Chemometrics and Intelligent Laboratory Systems,* **52**, 87–104.

Rousseeuw, P. J. and Baxter, M. A. (1987). *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*, 2nd Ed., Springer-Verlag, New York.

2011년 2월 접수; 2011년 3월 채택

Outlier Detection Using Support Vector Machines

Han Son Seo[a], Min Yoon[1,b]

[a]Department of Applied Statistics, Konkuk University
[b]Department of Statistics, Pukyong National University

Abstract

In order to construct approximation functions for real data, it is necessary to remove the outliers from the measured raw data before constructing the model. Conventionally, visualization and maximum residual error have been used for outlier detection, but they often fail to detect outliers for nonlinear functions with multidimensional input. Although the standard support vector regression based outlier detection methods for nonlinear function with multidimensional input have achieved good performance, they have practical issues in computational cost and parameter adjustments. In this paper we propose a practical approach to outlier detection using support vector regression that reduces computational time and defines outlier threshold suitably. We apply this approach to real data examples for validity.

Keywords: Outlier detection, support vector regression, practical approach.

This paper was supported by Konkuk University in 2010.

[1] Corresponding author: Assistant Professor, Department of Statistics, Pukyong National University, Daeyeon 3-Dong, Nam-Gu, Busan 608-737, Korea. E-mail: myoon@pknu.ac.kr