한국통계학회논문집 제4권 2호
1997년 8월 pp. 327 - 332

회귀분석을 위한 로버스트 신경망

황 창 하[1], 김 상 민[2], 박 희 주[3]

요 약

다층 신경망은 비모수 회귀함수 추정의 한 방법이다. 다층 신경망을 학습시키기 위해 역전파 알고리즘이 널리 사용되고 있다. 그러나 이 알고리즘은 이상치에 매우 민감하여 이상치를 포함하고 있는 자료에 대하여 원하지 않는 회귀함수를 추정한다. 본 논문에서는 통계물리에서 자주 사용하는 방법을 이용하여 로버스트 역전파 알고리즘을 제안하고 수학적으로 신경망과 매우 유사한 PPR(projection pursuit regression) 방법, 일반적인 역전파 알고리즘과 모의실험을 통해 비교 분석한다.

1. 로버스트 역전파 알고리즘의 제안

독립변수가 두개 이상 일 때 비선형 회귀함수 추정을 위해 주로 사용되는 비모수적 방법으로 PPR과 다층 신경망이 있다. 두 방법은 차원문제(curse of dimensionality)를 극복하기 위해 독립변수 벡터의 투영을 생각한다. 원래 다층 신경망은 패턴인식을 위한 도구로서 제안되었다. 패턴인식의 기본이론은 회귀분석의 특별한 경우인 판별분석인데 공학에서는 여러가지 이유로 PPR 보다 다층 신경망을 선호한다. 3개의 층으로 구성된 신경망을 이용한 회귀분석 방법은 PPR과 수학적으로 매우 비슷하다. Friedman & Stuetzle(1981)은 PPR 방법을 처음으로 제안하였고 Diaconis & Shahshahani(1984)는 어떤 연속함수라도 PPR 방법으로 근사시킬 수 있음을 보였다. 한편 Hwang 등(1994)은 Hermite 함수를 이용한 PPR을 제안하여 신경망과 비교 분석하였고 Roosen & Hastie(1994)는 평할 스플라인 함수를 이용한 PPR을 제안하여 기존의 PPR과 비교 분석하였는데, 일반적으로 PPR 방법이 신경망 보다 우수하다고 결론 내렸다.

다층 신경망을 학습시키기 위한 학습 알고리즘으로 역전파 알고리즘이 널리 이용되어져 왔다. 이 알고리즘은 최급강하법을 이용하여 오차제곱의 합이 최소가 되도록 신경망의 모수, 즉 가중치를 반복적으로 조정하는 알고리즘이다. 역전파 알고리즘으로 학습되는 신경망은 비선형 함수를 추정하게 된다. 그런데 실제상황에서는 많은 경우에 자료에 과대오차(gross error)와 이상치가 포함되게 된다. 그리고, 일반적으로 회귀분석을 위한 자료는 오차를 수반하고 많은 경우에 이상치 또는 이상치로 의심되는 관측치가 포함된다. 따라서 과대오차에 민감하지 않고, 이상치의 영향을 최소화 시키는 로버스트 역전파 알고리즘의 필요성이 대두되었다. 본 논문에서는 통계물리에서 많이

1) (712-702) 경북 경산시 하양읍 금락리 대구효성가톨릭대학교 정보통계학과 조교수
2) (740-200) 경북 김천시 삼락동 김천전문대학 전산정보처리과 조교수
3) (712-702) 경북 경산시 하양읍 부호리 경일대학교 전자계산학과 교수

사용되는 방법을 이용하여 로버스트 역전파 알고리즘을 제안하고 모의실험을 통해 기존의 역전파 알고리즘 및 PPR 방법과 비교 분석한다.

훈련표본 집합 $T=\{(\boldsymbol{x}_p, \boldsymbol{y}_p); p=1, \cdots, P\}$를 사용해서 미지의 회귀함수를 추정하는 문제를 생각해 보자. 이때 입력벡터 $\boldsymbol{x}_p=(x_{p1}, \cdots, x_{pn})^t$와 출력벡터 $\boldsymbol{y}_p=(y_{p1}, \cdots, y_{pm})^t$는 미지의 함수 $\boldsymbol{f}$에 대해 $\boldsymbol{y}_p=\boldsymbol{f}(\boldsymbol{x}_p)+\boldsymbol{e}_p$이라 가정한다. 여기서 $\boldsymbol{e}_p$는 오차벡터이다. 일반적인 역전파 알고리즘은 다음과 같은 오차제곱합

$$E_{LS}(\boldsymbol{W}, T)=\frac{1}{2}\sum_{p=1}^{P}\sum_{j=1}^{m}(y_{pj}-\hat{y}_{pj})^2$$

을 최소화하는 $\boldsymbol{f}$의 추정치 $\hat{\boldsymbol{f}}$을 구하기 위해 신경망의 가중치를 반복적으로 조정해 나가는 것이다. 이때 $\boldsymbol{W}$는 가중치집합을, T는 훈련표본 집합을, y_{pj}는 p번째 출력벡터의 j번째 원소를 나타내고 $\hat{y}_{pj}$는 신경망에 의해 추정된 p번째 출력벡터의 j번째 원소를 나타낸다.

로버스트 역전파 알고리즘을 유도하기 위해 통계물리에서 자주 사용되는 다음과 같은 일반화된 에너지함수를 사용한다.

$$E(\boldsymbol{V}, \boldsymbol{W})=\sum_{p=1}^{P}\sum_{j=1}^{m}V_p z(y_{pj}, \hat{y}_{pj})+E_{prior}(\boldsymbol{V})$$

여기서 $z(y_{pj}, \hat{y}_{pj})=\frac{1}{2}(y_{pj}-\hat{y}_{pj})^2$이고 V_p는 0 또는 1의 값을 가지는 확률변수이며 $\boldsymbol{V}=\{V_p, p=1, \cdots, P\}$이다. 즉, V_p는 입력자료가 이상치이면 0, 아니면 1의 값을 갖는다. 한편 $E_{prior}(\boldsymbol{V})$는 $\{V_p\}$의 사전분포에 의해 공헌되어진 에너지의 양을 나타내며, 이것의 일반적인 선택은 다음과 같다.

$$E_{prior}(\boldsymbol{V})=\eta\sum_{p=1}^{P}(1-V_p).$$

그 의미를 설명하면 다음과 같다. 만일 $z(y_{pj}, \hat{y}_{pj})<\eta$이면, $V_p=1$이 되어 주어진 관측치가 표본으로 간주되고, 그렇치 않으면 $V_p=0$이 되어 이상치로 간주된다.

한편 η는 지정된 반복횟수에 도달할 때 마다 계산되는 우측경계값으로 정의한다. 즉, 지정된 반복횟수에 도달 할때 마다 관측치들에 대응되는 $\sum_{j=1}^{m}z(y_{pj}, \hat{y}_{pj})$값들을 구한 후에 정렬하여 삼사분위수($Q_3$)와 일사분위수($Q_1$)를 계산한다. 그리고 삼사분위수에서 일사분위수를 공제한 값인 사분위범위수(IQR)를 계산하여 우측경계값, $Q_3+1.5\times IQR$을 η값으로 취한다.

우리들의 목표는 V_p가 이진 값을 가진다는 제약조건 하에서 $\{V_p\}$와 $\boldsymbol{W}$에 관해 $E(\boldsymbol{V}, \boldsymbol{W})$를 최소화 시키는 것이다. 그런데 이 문제는 연속형변수와 이산형변수가 혼합된 경우의 최적화 문제이기 때문에 해석적인 해를 구할 수 없을 뿐 아니라, 최급강하법을 사용하여 해를 구하는 것도 쉽지는 않다. 따라서 이런 문제점을 해결하기 위해 Gibbs 분포를 사용하며, 그 분포는 $P[\boldsymbol{V}, \boldsymbol{W}]=1/Ze^{-\beta E[\boldsymbol{V}, \boldsymbol{W}]}$로 정의된다. 이때, Z는 관계식 $\sum_{\boldsymbol{V}}\int_{\boldsymbol{W}}P[\boldsymbol{V}, \boldsymbol{W}]=1$을 만족한다. 따라서 $E(\boldsymbol{V}, \boldsymbol{W})$를 최소화하는 문제는 $P[\boldsymbol{V}, \boldsymbol{W}]$를 최대화하는 문제로 귀착된다. 그러나 이것 또한 연속형변수와 이산형변수가 혼합된 경우의 최적화 문제이기 때문에 어려움이 따른다. 따라서 이런 문제에 대한 하나의 해결책으로는 $\boldsymbol{W}$의 주변분포 $P_{margin}[\boldsymbol{W}]$를 구하여 이것을 최대화 시키는 것이다. 이때 $\boldsymbol{W}$의 주변분포는 다음과 같이 계산된다.

$$P_{margin}(\boldsymbol{W}) = \frac{1}{Z}\sum_{V}\exp\left(-\beta\sum_{p=1}^{P}\{\sum_{j=1}^{m}V_p\frac{1}{2}(y_{pj}-\hat{y}_{pj})^2+\eta(1-V_p)\}\right)$$
$$= \frac{1}{Z}\prod_{p=1}^{P}\sum_{V_p=\{0,1\}}\exp\left(-\beta\{\sum_{j=1}^{m}V_p\frac{1}{2}(y_{pj}-\hat{y}_{pj})^2+\eta(1-V_p)\}\right)$$
$$= \frac{\exp(-P\beta\eta)}{Z}\prod_{p=1}^{P}\left\{1+\exp\left(-\beta\{\sum_{j=1}^{m}\frac{1}{2}(y_{pj}-\hat{y}_{pj})^2-\eta\}\right)\right\}.$$

이제, $Z_m = Ze^{P\beta\eta}$와 $E_{eff}(\boldsymbol{W}) = -1/\beta\sum_{p=1}^{P}\log\left\{1+\exp\left(-\beta\{\sum_{j=1}^{m}\frac{1}{2}(y_{pj}-\hat{y}_{pj})^2-\eta\}\right)\right\}$로 두면, 주변 분포는 $P_{margin}(\boldsymbol{W}) = 1/Z_m\exp(-\beta E_{eff}(\boldsymbol{W}))$가 된다. 따라서, $P_{margin}(\boldsymbol{W})$를 최대화하는 문제는 $E_{eff}(\boldsymbol{W})$를 최소화하는 문제로 귀착된다. 한편, $z(\boldsymbol{y},\hat{\boldsymbol{y}}) = \sum_{j=1}^{m}z(y_{pj},\hat{y}_{pj})$의 값이 작다면 $E_{eff}(\boldsymbol{W})$의 합기호 내의 각 항의 값은 $z(\boldsymbol{y},\hat{\boldsymbol{y}})$가 되고, 반면 $z(\boldsymbol{y},\hat{\boldsymbol{y}})\to\infty$이면 $E_{eff}(\boldsymbol{W})$의 합기호 내의 각 항의 값은 일정한 상수가 된다. 따라서 $z(\boldsymbol{y},\hat{\boldsymbol{y}})$값을 크게 하는 관측치를 이상치로 간주하며 표본과 다르게 취급할 수 있다. 이 방법을 로버스트 M-추정법의 일반화로 간주할 수 있다.

이제 로버스트 에너지함수 $E_{eff}(\boldsymbol{W})$를 사용하는 로버스트 역전파 알고리즘의 유도과정을 은닉층이 하나인 신경망에 대해서 설명한다. 비선형 활성함수를 사용하여 다음 층으로 변화량을 전달하는 방법은 다음과 같다. w_{ji}는 입력층에서 은닉층으로의 가중치를, θ_j는 은닉층 노드의 오프셋을, v_{kj}는 은닉층에서 출력층으로의 가중치를, θ_k는 출력층 노드의 오프셋을, x_{pi}는 입력층의 i번째 노드의 p번째 입력값을, h_{pj}는 은닉층의 j번째 노드에서 p번째 입력벡터의 출력을, $\hat{y}_{pk}$는 출력층의 k번째 노드에서 p번째 입력벡터의 출력을, a_{pj}는 입력층에서 은닉층으로의 가중치와 입력층의 출력값과의 곱의 합을, b_{pk}는 은닉층에서 출력층으로의 가중치와 은닉층의 출력값과의 곱의 합을, g_j는 은닉층의 시그모이드 비선형 활성 함수를, g_k는 항등함수로서 출력층의 활성함수를 나타낸다고 하자. 그러면, $a_{pj} = \sum_i w_{ji}x_{pi}$, $h_{pj} = g_j(a_{pj})$, $b_{pk} = \sum_j v_{kj}h_{pj}$, $\hat{y}_{pk} = g_k(b_{pk})$이 된다.

다층 전방향 신경망에서 가중치의 변화량은 오차제곱합이 가장 많이 감소하는 방향으로 변화한다. 즉, 수식으로 표현하면 $\Delta w_{ji}\propto -\frac{\partial E_{eff}}{\partial w_{ji}}$이다. 합성함수 미분공식과 역전파 알고리즘을 유도하는 방법을 사용하여 로버스트 역전파 알고리즘을 다음과 같이 두 가지 경우에 대하여 구한다.

(1) 출력층 노드의 경우:

$$\Delta v_{kj} = \alpha\sum_{p=1}^{P}\frac{1}{1+\exp(\beta\{\sum_{j=1}^{m}\frac{1}{2}(y_{pj}-\hat{y}_{pj})^2-\eta\})}\times\{y_{pk}-\hat{y}_{pk}\}\times h_{pj}$$
$$\equiv \alpha\delta_{pk}h_{pj},$$
$$\Delta\theta_k = \beta\delta_{pk}$$

(2) 중간층 노드의 경우:

$$\Delta w_{ji} = \alpha\sum_{k}\delta_{pk}v_{kj}\hat{y}_{pi}$$
$$\equiv \alpha\delta_{pj}\hat{y}_{pi},$$
$$\Delta\theta_j = \beta\delta_{pj}$$

여기서, α, β는 학습률이다.

2. 모의실험

본 논문의 모의실험의 첫번째 목적은 회귀분석 문제에 대해 제안된 로버스트 역전파 알고리즘과 일반적인 역전파 알고리즘을 비교 분석하는 것이고 그 다음 목적은 역전파 알고리즘 보다 우수하다고 알려진 PPR과 로버스트 역전파 알고리즘을 비교 분석하는 것이다. 이상치가 포함된 자료에 대한 사전 모의실험에서 로버스트 역전파 알고리즘이 일반적인 역전파 알고리즘 보다 더 좋은 결과를 보여 주었는데 예상되는 결과이고 지면 관계상 생략하였다.

세가지 방법의 성능을 분석하기 위해 Hwang 등(1994)이 사용한 5개의 이변량 비선형함수를 추정하는 문제를 생각한다. 5개의 함수는 다음과 같다.

• Simple Interaction Function

$$g^{(1)}(x_1, x_2) = 10.391\{(x_1 - 0.4)\cdot(x_2 - 0.6) + 0.36\}$$

• Radial Function

$$g^{(2)}(x_1, x_2) = 24.234\{r^2(0.75 - r^2)\},\quad r^2 = (x_1 - 0.5)^2 + (x_2 - 0.5)^2$$

• Harmonic Function

$$g^{(3)}(x_1, x_2) = 42.659\{(2 + x_1)/20 + Re(z^5)\},$$

여기서 $z = x_1 + ix_2 - 0.5(1 + i)$. 혹은

$$g^{(3)}(x_1, x_2) = 42.659\{0.1 + \tilde{x}_1(0.05 + \tilde{x}_1^4 - 10\tilde{x}_1^2\tilde{x}_2^2 + 5\tilde{x}_2^4)\},$$

여기서 $\tilde{x}_1 = x_1 - 0.5$, $\tilde{x}_2 = x_2 - 0.5$.

• Additive Function

$$g^{(4)}(x_1, x_2) = 1.3356[1.5(1 - x_1) + e^{2x_1 - 1}\sin\{3\pi(x_1 - 0.6)^2\} + e^{3(x_2 - 0.5)}\sin\{4\pi(x_2 - 0.9)^2\}]$$

• Complicated Interaction Function

$$g^{(5)}(x_1, x_2) = 1.9[1.35 + e^{x_1}\sin\{13(x_1 - 0.6)^2\}e^{-x_2}\sin(7x_2)]$$

모의실험을 위한 225개의 훈련자료(training data)의 가로축 좌표값 $\{(x_{l1}, x_{l2})\}$의 x_{l1}, x_{l2}는 서로 독립이며, 균등분포 $U[0, 1]$로 부터 생성되었다. 5개의 함수에 대한 모의실험에서 항상 같은 225개의 쌍 $\{(x_{l1}, x_{l2})\}$이 사용된다. 그리고 이들 가로축 좌표값에 대해 오차가 없는 훈련자료 $y_l^{(j)} = g^{(j)}(x_{l1}, x_{l2})$, $l = 1, 2, \cdots, 225$, $j = 1, \cdots, 5$를 만들었다. 아울러 iid 정규오차를 더하여 훈련자료 $y_l^{(j)} = g^{(j)}(x_{l1}, x_{l2}) + 0.25\varepsilon_l$, $l = 1, 2, \cdots, 225$, $j = 1, \cdots, 5$를 만들었다. 여기서, $\varepsilon_l \sim N(0, 1)$이다. 그리고 적합된 모형의 성능평가를 위해서 검정자료(test data)를 사용하는데 이 검정자료는 $[0, 1]^2$ 상의 10,000개의 등간격 격자점 $x_{li} = (2l - 1)/200$, $l = 1, \cdots, 100$, $i = 1, 2$과 이들 격자점에 대한 10,000개의 함수값 $\{g^{(j)}(x_{l1}, x_{l2})\}$로 구성된다.

역전파 알고리즘, 로버스트 역전파 알고리즘 및 PPR의 비교, 분석을 위한 측도로 우리는 Hwang 등(1994)과 Roosen & Hastie(1994)등이 사용한 FVU(fraction of variance unexplained)를 사용한다. 실험조건 또한 그들의 것과 동일하다. 한편 FVU는 다음과 같이 정의된다.

$$FVU = \frac{\sum_{l=1}^{N} (\hat{g}(\boldsymbol{x}_l) - g(\boldsymbol{x}_l))^2}{\sum_{l=1}^{N} (g(\boldsymbol{x}_l) - \bar{g}(\boldsymbol{x}_l))^2}$$

여기서, $g = g^{(j)}$, $j = 1, \cdots, 5$ 이다.

표1은 모의실험의 결과를 설명한다. 여기서 BP는 일반적인 역전파 알고리즘을, RBP는로버스트 역전파 알고리즘을 의미한다. 표1에 의하면 대부분의 함수들의 경우에 오차를 포함하지 않는 훈련자료 및 검정자료에 대해서는 근소한 차이지만 PPR이 다른 두 방법 보다 다소 좋은 결과를 보여주며, 검정자료에 대해서 몇몇 경우에는 RBP가 오히려 약간 더 좋은 결과를 보여주기도 한다. 한편 BP와 RBP는 거의 비슷한 결과를 보여준다. 오차를 포함하는 자료에 대해서 생각해 볼 때 훈련자료에 대해서는 BP, RBP와 PPR은 거의 비슷한 결과를 보여주며, 검정자료에 대해서는 대체로 RBP가 다른 두 방법 보다 좋은 결과를 보여준다. 회귀분석 문제에서 종속변수의 관측치들은 일반적으로 오차를 포함하고 있고, 그리고 회귀분석에서 적합도 중요하지만 아울러 예측(prediction) 또는 일반화(generalization)도 매우 중요하므로 BP와 PPR 대신에 RBP를 사용하는 것이 좋을 것으로 생각된다. 한편 사전 모의실험을 통해서 이상치가 포함된 자료에 대해서는 RBP가 BP 보다

표 1. FVU에 의해 결정되는 정확도

		오차가 없는자료			오차가 있는 자료		
함수	방법	노드수	훈련표본	검정표본	노드수	훈련표본	검정표본
g(1)	BP	5	0.000534	0.001471	5	0.064836	0.070192
		10	0.000227	0.001338	10	0.064651	0.072336
	RBP	5	0.000534	0.001266	5	0.064836	0.007300
		10	0.000227	0.001285	10	0.064651	0.007984
	PPR	3	0.000075	0.001077	3	0.053925	0.080629
		5	0.000048	0.001095	5	0.050652	0.080352
g(2)	BP	5	0.005934	0.007648	5	0.068849	0.083526
		10	0.003537	0.005400	10	0.058644	0.073024
	RBP	5	0.005949	0.006711	5	0.068866	0.023277
		10	0.003537	0.005127	10	0.059375	0.014247
	PPR	3	0.025906	0.034620	3	0.057967	0.082920
		5	0.001163	0.006069	5	0.058581	0.084407
g(3)	BP	5	0.524021	0.424022	5	0.399383	0.566565
		10	0.142020	0.151487	10	0.146998	0.231935
	RBP	5	0.567275	0.495448	5	0.401656	0.472818
		10	0.142935	0.132279	10	0.147073	0.169004
	PPR	3	0.360373	0.577145	3	0.185445	0.321773
		5	0.112422	0.242643	5	0.138678	0.341155
g(4)	BP	5	0.045945	0.019688	5	0.073419	0.086255
		10	0.004273	0.006913	10	0.059601	0.075581
	RBP	5	0.015946	0.019215	5	0.073427	0.025834
		10	0.004283	0.005262	10	0.061691	0.015751
	PPR	3	0.000389	0.000692	3	0.041929	0.091219
		5	0.000427	0.001915	5	0.038907	0.089882
g(5)	BP	5	0.214445	0.236293	5	0.249337	0.286426
		10	0.025470	0.070035	10	0.096857	0.138735
	RBP	5	0.214724	0.234898	5	0.262230	0.260099
		10	0.025487	0.065402	10	0.099910	0.086531
	PPR	3	0.140191	0.227764	3	0.294997	0.543458
		5	0.016889	0.038313	5	0.060511	0.192520

훨씬 더 좋은 결과를 보여주었다. 그리고 특히 공학분야에서 회귀분석과 관련된 문제에서 여러 이유로 PPR 보다 BP를 더 많이 사용한다. 따라서 RBP의 사용은 나름대로 의미가 있을 것이다.

참 고 문 헌

[1] Diaconis, P. and Shahshahani, M. (1984). *On Linear Functions of Linear Combinations*, SIAM J. SCI. STAT. COMPUT., **5**, 175-191.

[2] Friedman, J. H. and Stuetzle, W. (1981). *Projection Pursuit Regression*, J. Amer. Statis. Assoc., **76**, 817-823.

[3] Hwang, J-N, Lay, S-R, Maechler, M., Martin, D. and Schimert, J. (1994). *Regression Modeling in Back-Propagation and Projection Pursuit Learning*, IEEE Transactions on Neural Networks, **5**, 342-353.

[4] Roosen, C. B. and Hastie, T. J. (1994). *Automatic Smoothing Spline Projection Pursuit*, Journal of Computational and Graphical Statistics, **3**, 235-248.