신경망 초기치 탐색방법 비교연구

최 대 우 * 구 자 용 † 박 헌 진 ‡

요 약

데이터 마이닝 분야에서 널리 사용되고 있는 신경망은 최근 많은 통계인들의 관심을 끌고 있다. 그러나 범용 근사성(universal approximator)이라는 성질에도 불구하고 초기치에 따라 적합 결과가 크게 좌우되는 단점이 있다. 본 논문에서는 붓스트랩 표본을 통해 초기치를 발견하는 bumping 기법이 신경망 분야에서 사용되고 있는 무작위 탐색법 보다 더 정확하고 안정적인 초기치를 제공하여 주는가를 살펴 보았다.

주요용어: 신경망, 초기치, bumping

1 서 론

신경망은 인간의 신경전달 과정을 모방한 것으로 McCulloch와 Pitts(1943)가 처음 제안한 이래 전산학의 기계학습(machine learning), 형상인식(pattern recognition) 분야뿐 아니라 경제, 경영학 등 다양한 분야에서 사용되고 있다. 특히 방대한 자료에서 변수 사이의 관계가 매우 복잡한 구조를 가진 경우 이 관계를 근사적으로 추정하여 예측하는데 유용하게 쓰인다.

본 연구에서는 고전적 초기치 탐색법인 무작위 탐색방법 보다 효율적인 bumping 탐색법을 제안하고, 분류예측을 위한 신경망을 세가지 실 자료에 적용하여 새로운 초기치 탐색 알고리즘의 유용성과 안정성을 살펴보았다.

2 신경망의 개요

신경망에는 여러 종류가 있으나 본 논문에서는 한개의 은익층(hidden layer)으로 이루어진 전방 신경망(feed forward neural network)을 고려한다. 은익층이 한

*(449-791) 경기도 용인시 모현면 한국외국어대학교 정보통계학과 조교수
†(200-702) 강원도 춘천시 옥천동 1번지 한림대학교 정보통계학과 교수
‡(402-751) 인천시 남구 용현동 253번지 인하대학교 통계학과 부교수

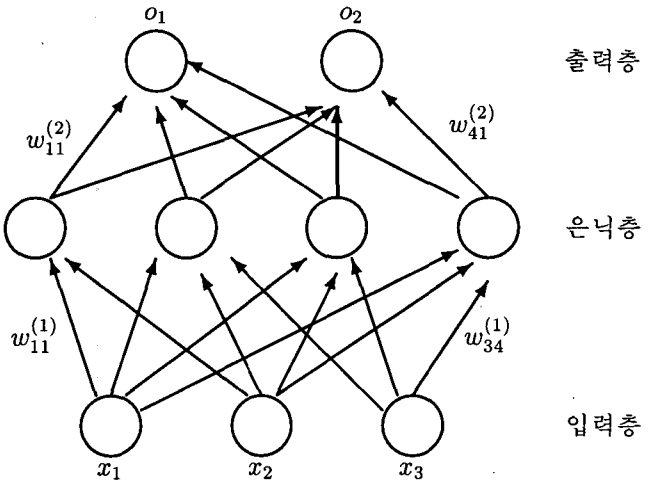

그림 1: 신경망 구조 ($p = 3$; $L = 4$; $q = 2$)

개인 전방 신경망은 그림 1에서 처럼 세개의 층과 각 층에 나열된 노드들로 구성되어 있고, 각 층의 노드는 다음 층의 노드에 연결되어 있다.

신경망에서는 전체의 입력값은 은익층의 각 노드의 입력이 되고, 은익층에서 출력된 값은 출력층의 각 노드의 입력값이 된다. 입력층, 은닉층, 출력층 각각의 노드 개수가 p, L, q인 경우, 신경망은 아래와 같은 함수로 나타낼 수 있다.

$$o_k = f^{(2)}\left(w_{0k}^{(2)} + \sum_{j=1}^{L} w_{jk}^{(2)} f^{(1)}\left(w_{0j}^{(1)} + \sum_{i=1}^{p} w_{ij}^{(1)} x_i\right)\right), \; k = 1, \cdots, q \quad (1)$$

위에서 $(x_1, \cdots, x_p)$는 입력변수의 값, o_k는 신경망에서 출력층의 k번째 노드 출력값을 나타낸다. 식 (1)에서

$$\boldsymbol{w} = \{w_{01}^{(1)}, \cdots, w_{pL}^{(1)}, w_{01}^{(2)}, \cdots, w_{Lq}^{(2)}\} \quad (2)$$

는 알려져 있지 않는 모수로 신경망에서는 가중치라고도 한다. 식 (1)에서 $f^{(1)}$과 $f^{(2)}$는 은익층과 출력층에서 노드의 활성함수를 나타낸다.

식 (1)에서 함수 $f^{(1)}$와 $f^{(2)}$의 형태가 주어진다 하더라도 모수 $\boldsymbol{w}$의 값을 추정하여야 한다. 관측값이 N개 있을 때 입력변수의 n번째 관측값을 $x^{(n)} = (x_1^{(n)}, \ldots, x_p^{(n)})$ 출력변수의 n번째 관측값을 $y^{(n)} = (y_1^{(n)}, \ldots, y_q^{(n)})$이라 하고 입력변수의 n번째 관측값을 신경망에 적용하여 나온 출력노드의 값을 $o^{(n)} = (o_1^{(n)}, \ldots, o_q^{(n)})$이라 하자. 이 때, 출력변수가 다항분포(multinomial distribution)을 따르는 범주형 변수라면 아래의 로그 가능도함수(log-likelihood function)로부터 최대가능도 추정치

를 구할 수 있다.

$$E = \sum_{n=1}^{N} \sum_{k=1}^{q} y_k^{(n)} \log o_k^{(n)} \tag{3}$$

여기서, $\sum_{k=1}^{q} o_k^{(n)} = 1$이다.

식 (3)의 E를 최대화하는 모수 $\boldsymbol{w}$의 추정치는 비선형 최적화(nonlinear optimization)를 통하여 구하게 된다. 이 때 흔히 사용되는 방법이 최대하강법(steepest descent method)으로써 신경망에서는 back-propagation이라고도 한다. 비선형 최적화에서 모수 $\boldsymbol{w}$의 추정치를 구할 때 국소최대(local maxima) 문제가 발생하는데, 특히 신경망에서 입력노드와 은익노드의 개수가 많은 경우 목적함수 E의 표면이 매끄럽지 못하여 여러 개의 국소최대가 존재하게 된다. 이러한 경우 초기값의 선택이 국소최대가 아닌 전체최대(global maxima)를 찾는데 중요한 요인인 것이다.

3 무작위 및 bumping을 이용한 초기치 탐색

신경망에서는 초기치의 선택에 따라 그 적합결과가 크게 변한다. 이러한 초기치 문제를 해결하고자 신경망에서는 고전적인 방법으로 무작위 탐색 방법이 사용되어 왔다. 이 절에서는 고전적인 방법을 간략히 설명하고 개선 방안으로 bumping에 의한 탐색방법을 제안하고자 한다.

3.1 무작위 탐색에 의한 방법

현재 가장 대중적으로 사용되는 초기치 탐색방법으로서 무작위 탐색방법은 아래의 단계로 구성된다.

STEP 1 전방신경망의 각 모수 $\boldsymbol{w}$, 즉 가중치에 대하여 서로 독립인 난수를 발생시키고 가중치 개수만큼 발생시킨 난수들을 가중치의 초기값으로 한다. 난수 발생시 사용되는 분포는 보통 구간 $[-\delta, \delta]$에서의 균일분포이다. 여기서 δ는 보통 0.5에서 1사이의 값으로 한다. 이 난수 값을 초기값으로 하여 비선형 최적화에서 일정한 횟수의 반복을 통하여 가중치 $\boldsymbol{w}$에 대한 추정값 $\hat{\boldsymbol{w}}$를 구하고 추정된 가중치 $\hat{\boldsymbol{w}}$에 대한 오분류 비율의 값 $\hat{r} = r(\hat{\boldsymbol{w}})$을 구한다.

STEP 2 STEP 1을 K번 반복하여 가중치의 추정값 $\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_K$들을 구하고 신경망을 적합하기 위한 과거자료인 훈련자료(tranining data)에 대응하는 오분류 비율 $\hat{r}_1, \ldots, \hat{r}_K$들을 구한다.

$$\hat{k} = \text{argmin}_{1 \le k \le K} \hat{r}_k$$

라 하면 $\hat{\boldsymbol{w}}_{\hat{k}}$를 최종 초기값으로 한다.

3.2 Bumping에 의한 방법

훈련자료 $z = (z_1, \ldots, z_N)$라 하고 관측치 z_i들은 특정 분포 F에서 추출된 임의 표본(random sample)이라고 하자. 각 자료 $z_i = (x^{(i)}, y^{(i)})$는 설명변수 $x^{(i)}$와 분류수준 $y^{(i)}$로 이루어져 있다고 하자.

훈련자료 z에 대한 붓스트랩 표본들을 $z^{*1}, \ldots, z^{*B}$라 하자. 훈련자료 z에 의해 구해진 분류모형을 C라 하고 각 붓스트랩 표본을 훈련자료로 간주하여 생성된 분류모형을

$$C_1 = C(z^{*1}),, \ldots, C_B = C(z^{*B})$$

라 하자.

Bumping은 Bootstrap Umbrella of Model Parameters의 약자로 Tibshirani와 Knight(1999)에 의해 제안된 모형추정 방법이다. 자세한 내용은 Tibshirani와 Knight(1999)를 참조하길 바란다.

신경망을 이용한 분류(classification)에 있어 bumping을 응용한 초기치 탐색 알고리즘을 제안하자면 다음과 같다.

STEP 1 훈련자료 z에 대하여 B개의 붓스트랩 표본 $z^{*1}, \ldots, z^{*B}$을 생성한다.

STEP 2 각 붓스트랩 표본에 대하여 신경망 $C_1, \ldots, C_B$를 적합한다. 이 경우 초기값은 무작위 탐색에서의 STEP 1과 같이 정한다.

STEP 3 훈련자료 z를 분류예측을 위한 신경망 $C_1, \ldots, C_B$에 적용하여 오분류 비율이 가장 작은 모형 C_{b_0}의 최종 수렴한 가중치를 초기치로 정한다.

4 모 독일은행 신용평가자료 분석결과

4.1 실험 및 평가방법

이 절에서는 무작위 탐색방법과 bumping에 의한 탐색방법을 독일 모 은행 신용평가자료에 적용하였다. 자료의 70%는 훈련자료, 30%는 검증(validation)자료로 사용하였고 초기치를 추출하는 구간을 $[-0.5, 0.5]$로 고정하였다. Bumping을 이용한 초기치 탐색의 경우는 20개의 붓스트랩 표본을, 무작위 탐색은 100개의 난수로 부터 발생된 초기치들을 사용하여 신경망을 적합한 후, 훈련자료에 대하여 분류, 예측하였다. 이때 오분류 비율이 가장 작은 모형의 가중치들을 초기치로 선택하여 최종 모형을 적합하였다. 무작위 및 bumping 탐색방법의 성능은 검증자료로 부터의 이득률(gain)과 오분류 비율(MER; Miscalssification Error Rate) 측면에서 평가하였고, 이와 같은 과정을 총 20회 반복하였다. 이득률에 대한 자세한 설명은 구자용, 박헌진, 최대우(2000)을 참조하길 바란다.

본 연구에서 사용된 신경망은 S-PLUS의 라이브러리로 제공되는 알고리즘으로 Vanables와 Ripley(1999)에 의해 구현된 것이다. 그 외 초기치로 부터 가중치를 찾는 반복횟수는 500번으로 고정하였고 설명변수 중 이산형은 가변수(dummy variable)화하여 사용하였다.

본 논문에서는 신경망이 산출하는 스코어에 의해 검증자료를 내림차순으로 정렬한 후 각 $(100 \times p)$%에서의 이득률과 오분류 비율로써 두 초기치 탐색방법을 평가한 것이다.

4.2 자료분석 결과

독일 모 은행의 신용평가 자료는 관측치의 개수가 총 1,000개로 13개의 이산형, 7개의 연속형 자료로 구성되어 있는데, 그 중 5개(이산형 3개, 연속형 2개)의 설명변수를 사용하여 분석하였다.

그림 2은 bumping 탐색방법과 무작위 탐색방법을 이용하여 20회 반복실험한 후 신용불량자 예측에서의 이득률을 도시한 것이다. 초기치는 구간 $[-0.5, 0.5]$에서 추출하였다. Bumping 탐색방법이 무작위 탐색에 비해 이득률 측면에서 안정적인 결과를 제공하는 것을 알 수 있다.

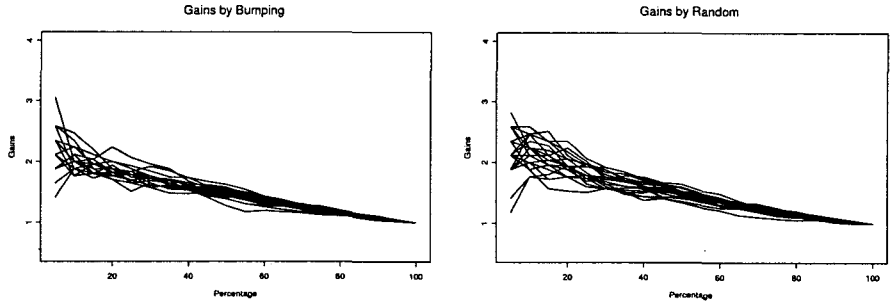

그림 2: 20회의 반복실험을 실시한 결과 (좌측이 bumping 탐색법)

그림 3는 20회의 반복실험 결과에 대해 이득률 평균에 대한 표준편차를 구한 후 이득률 평균을 중심으로 2배의 표준편차를 회색 띠로 도시한 것이다. Bumping 탐색방법이 무작위 탐색방법에 비해 적합 결과에 대한 분산이 작으면서 평균값의 패턴은 거의 비슷함을 알 수 있다.

그림 4은 20회 반복실험에 대해 오분류 비율의 평균과 평균에 대한 2배의 표준편차를 나타낸다. 역시 오분류 비율도 bumping 탐색 방법이 무작위 탐색 방법에 비해 적합결과에 대한 분산이 작고 평균값이 거의 비슷함을 알 수 있다.

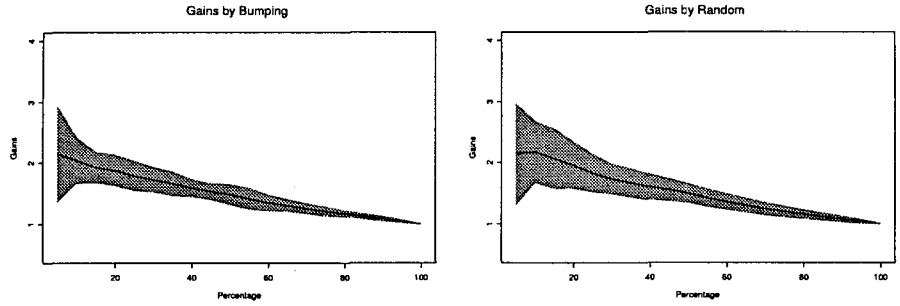

그림 3: 20회 반복실험에 대한 이득률 평균과 평균에 대한 2배의 표준편차 (좌측이 bumping 탐색법)

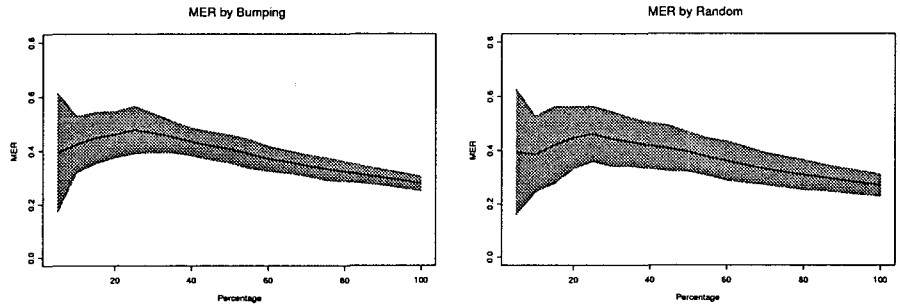

그림 4: 20회 반복실험에 대한 오분류 비율(MER)에 대한 평균과 평균에 대한 2배의 표준편차 (좌측이 bumping 탐색법)

5 결언

살펴 본 바와 같이 제한된 실험이었지만 bumping 탐색법은 무작위 탐색법과 거의 비슷한 성능의 신경망을 제공하였다. 아울러 이득률과 오분류 비율에 대한 표준편차 측면에서 큰 차이는 아니나 bumping 탐색법에 의해, 무작위 방법보다 안정적인 결과를 얻을 수 있었다. 그러나 20회의 붓스트랩 추출과 100회의 무작위 추출의 비교라는 측면에서 bumping 탐색법이 훨씬 효율적이라 할 수 있다.

무작위 탐색법에 비한 bumping의 효율성은 다음과 같은 원인에 의해 나타나는 현상으로 생각된다:

첫째, 다차원의 저주(curse-of-dimensionality)에 의한 무작위 탐색법의 한계를 들 수 있다. 즉, 필요한 개수 만큼의 초기치를 일정 구간 $[-\delta, \delta]$에서 추출하는 과

정을 수 없이 반복하더라도 최적의 초기치를 찾아내기는 거의 불가능하다는 것이다.

둘째로는 붓스트랩 표본이 제공하는 과도적합(over-fitting)의 방지효과를 들 수 있다. 고전적인 무작위 탐색법에서는 최종 초기치 선택에 있어, 훈련자료를 도출된 모형에 다시 적용하여 모형을 평가하므로 과도적합이 발생하는 것으로 생각된다. 이러한 무작위 탐색법에서의 과도적합 현상은 은닉층의 노드 개수가 늘어나 추정하여야 할 모수가 많아지는 경우 더욱 두드러지게 나타난다. 반면 bumping 탐색법에서는 붓스트랩 표본추출에 의해 훈련자료의 각 관측치가 0.632의 확률로 추출된 후 새로운 훈련자료로 사용되고, 원 훈련자료는 검증자료로 사용되기 때문에 훈련자료 자신에 과도하게 적합된 신경망 적합이 방지되는 것이다.

본 연구에서는, 신경망을 중심으로 효율적인 초기치 탐색법을 소개하였다. 제안된 방법을 많은 비선형 최적화의 초기치 선정문제에 적용할 수 있다.

참고 문헌

[1] 구자용, 박헌진, 최대우 (2000). 데이터 마이닝에서의 폴리클래스, 응용통계연구, 13권 2호

[2] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of ideas immanent in neural activity, *Bulletin of Mathematical Biophysics*, vol. 5, 115-133

[3] Tibshirani, R. and Knight, K. (1999). Model search by bootstrap "bumping", *Journal of Computational and Graphical Statistics*, vol. 8, 671-686

[4] Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*, 3rd ed. Springer, New York.