한국컴퓨터정보학회 동계학술대회 논문집 제21권 제1호 (2013. 1)

의사 샘플 신경망에서 특징 선택 기법

허경용°, 우영운*, 김지홍**, 이임건**, 김남규***
°동의대학교 전자공학과
*동의대학교 멀티미디어공학과
**동의대학교 영상정보공학과
***동의대학교 게임공학과
e-mail: {hgycap°, ywwoo*, arim**, iglee**, ngkim***}@deu.ac.kr

A Feature Selection Method in Pseudo Sample Neural Networks

Gyeongyong Heo°, Young Woon Woo*, Ji-Hong Kim**, Imgeun Lee**, Nam-Gyu Kim***
°Dept. of Electronic Engineering, Dong-Eui University
*Dept. of Multimedia Engineering, Dong-Eui University
**Dept. of Visual Information Engineering, Dong-Eui University
***Dept. of Game Engineering, Dong-Eui University

● 요 약 ●

신경망의 학습은 학습 샘플의 품질뿐만이 아니라 입력으로 사용되는 특징에도 영향을 받으므로 신경망의 출력을 결정하는데 있어 연관성이 높은 특징을 입력으로 사용함으로써 학습된 신경망의 전체적인 성능을 높일 수 있다. 이 논문에서는 신경망의 입력으로 사용되는 특징과 출력의 연관성 파악하고 연관성이 낮은 특징을 학습 과정에서 배제함으로써 신경망의 전체적인 성능을 높일 수 있는 방법을 제시하였다. 토석류 데이터를 위한 의사 샘플 신경망에 제안한 방법을 적용한 경우 연관성이 낮은 특징 하나를 제외함으로써 약 6%의 오류 감소 효과를 얻을 수 있었다.

키워드: 의사샘플 신경망(pseudo sample neural network), 특징 선택(feature selection), 연관성(correlation)

I. 서 론

*신경망은 인간의 두뇌를 모델로 한 방법으로, 인간의 두뇌는 뉴런이 상호 연결되어 있으며 뉴런 사이에서의 신호 전달 과정을 통해 새로운 것을 학습할 수 있다는 이론에 기초하고 있다. 신경망은 입력과 출력 사이의 임의의 관계를 학습할 수 있을 뿐만이 아니라 체계적인 학습 방법이 정립되어 있어 분류기뿐만이 아니라 회귀 분석의 도구로도 그 활용도가 높다[1].

신경망의 학습에 있어 학습 샘플의 중요성은 잘 알려져 있다. 학습 샘플과 더불어 신경망의 성능에 영향을 미치는 중요한 인자 중 하나는 신경망의 입력으로 사용되는 특징(feature)으로 출력 값을 결정하는데 있어 변별력이 높은 특징을 사용하여야 함은 당연하다.

이 논문에서는 신경망의 입력과 출력 사이의 연관성에 기초한 간단하면서도 효과적인 특징 선택 방법을 제안한다. 일반적으로 특징 집합에서 유용한 특징의 부분집합은 반복 실험 과정을 통해 선택되지만 이는 많은 시간을 요하는 작업이므로 문제에 따라서는 적용이 불가능할 수 있다. 제안한 방법을 토석류 데이터를 위한 의사 샘플 신경망에 적용한 경우 오류의 평균 및 분산이 감소함을 실험을 통해 확인할 수 있다.

* 이 논문은 한국콘텐츠진흥원 2010년 선정 문화기술 공동연구센터 3차년도 사업의 연구결과로 수행되었음. [과제명 : 3D 입체영상제작 연구개발, 과제번호: 1-10-7602-001-10002-00-001]

II. 의사 샘플 신경망

토석류 데이터는 산사태에 의한 피해지역 예측을 위해 사용되는 데이터로[2], 5종류의 실측 및 분석 데이터를 입력으로 하고 피해 정도와 범위 예측에 필요한 3가지의 파라미터를 그 출력으로 한다. 토석류 데이터의 경우 데이터의 획득이 용이하지 않아 충분한 학습 데이터를 확보하는데 어려움이 있다. 따라서 이전 연구에서는 기존 샘플을 이용하여 의사 샘플을 생성하고 이를 학습에 이용하는 의사 샘플 신경망을 제안하였다. 의사 샘플 신경망은 학습 샘플을 증가시킴으로써 해공간을 평탄화시키고 학습된 신경망이 국부 최적해에 빠질 확률을 줄여줌으로써 학습된 신경망의 성능을 개선시킬 수 있는 방법이다. N개의 샘플 $X=\{x_1,x_2,...,x_N\}$이

주어진 경우 의사 샘플 X_{PS}는 식 (1)과 같이 주어지며 기존 신경망이 X를 학습에 사용하는 것과 달리 의사 샘플 신경망은 X_{PS}를 학습에 사용한다.

$$X_{PS}=\bigcup_{i=1}^{N}\left\{x_j' \mid x_j' \sim N\left(x_i,\frac{\sigma^2}{N}\right),j=1,...,N_{PS}\right\} \quad (1)$$

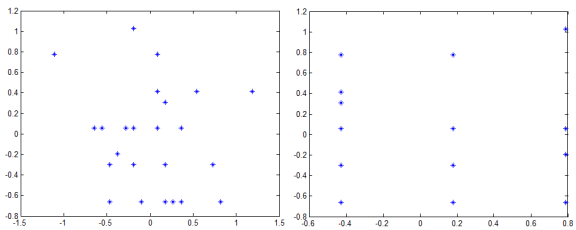
이 때 σ^2는 샘플 X의 분산을 나타내고 N_{PS}는 기존 샘플 하나에서 생성되는 의사 샘플의 개수를 나타낸다. 의사 샘플 신경망은 샘플의 개수가 적은 경우 적용하는 방법으로 기존 샘플로는 샘플의 분포를 추정하기 어려우므로 중심 극한 정리에 따라 가우시안 분포를 가정하였다[4].

III. 특징 선택

신경망의 학습에서 출력 값의 변별력에 미치는 영향이 적은 입력은 사용할 필요가 없음은 당연하며 이는 입력과 출력의 연관성을 통해 판단할 수 있다. 하지만 입력과 출력 사이의 연관성을 판단할 수 있는 일반적인 방법은 없으며 실험적으로 오류가 최대가 되는 특징을 연관성이 적은 것으로 판단하는 것이 일반적이다. 하지만 이러한 방법은 많은 시간을 요하므로 문제에 따라 적용이 불가능할 수 있다. 이 논문에서는 입력과 출력의 연관성을 파악할 수 있는 간단하면서도 효과적인 방법을 제시한다.

입력 벡터와 출력 벡터가 주어진 경우 이를 scatter plot으로 나타낸 것이 그림 1이다. 신경망은 입력과 출력의 비선형적인 임의의 관계도 학습할 수 있지만 scatter plot 상에서 특정 지역에 데이터가 밀집되어 나타나는 경우 이를 분별하기는 쉽지 않다. 따라서 scatter plot 상에서 데이터 포인트들 사이의 거리 분산이 작은 경우 학습에 부정적인 역할을 하는 것으로 생각할 수 있다. 그림 1에서 볼 수 있듯이 그림 1-a는 전체적으로 넓게 분포하는 반면 그림 1-b의 경우는 특정 지역에 밀집된 양상을 보인다. 즉, 그림 1-b에 나타난 특징이 성능 저하의 원인이 되는 것으로 판단할 수 있다.

두 벡터의 상관관계 파악을 위해서는 피어슨의 상관 분석이 많이 이용되지만[5] 선형 관계만을 파악할 수 있고 실제 데이터가 선형 관계를 가지는 경우는 드물기 때문에 사용하지 않았다.



(a) (b)

그림 1. 입출력 값의 상관관계

Fig. 1. Relations between input and output values

입력 $X=\{x_1,x_2,...,x_N\}$와 출력 $Y=\{y_1,y_2,...,y_N\}$가 주어진 경우 scatter plot 상에서 i번째 포인트와 j번째 포인트 사이의 거리는 식 (2)와 같이 주어진다.

$$d_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \quad (2)$$

입력 X와 출력 Y의 연관성 ρ는 거리 $d_{ij}(i=1,...,N,\ j=i+1,...,N)$의 분산으로 나타내질 수 있다. n번째 특징이 m번째 출력과 가지는 연관성을 ρ_{nm}이라 하면 n번째 특징이 출력과 가지는 연관성은 식 (3)과 같이 나타낼 수 있다.

$$\rho_n=\sum_{m=1}^{3}\rho_{nm} \quad (3)$$

토석류 데이터에서 입력이 5개, 출력이 3개이므로 ρ_{nm}은 $n(1\leq n\leq 3)$번째 입력이 $m(1\leq m\leq 5)$번째 출력과 가지는 연관성을, ρ_n은 n번째 입력이 가지는 전체 출력과의 연관성을 나타낸다. 식 (3)의 값이 작은 경우 scatter plot 상에서 샘플 포인트들이 밀집되어 나타나게 된다. 즉, 신경망을 통해 각각의 포인트들을 구별하기가 어려워진다. 따라서 학습 과정에서 식 (3)의 연관성을 계산하고 연관성이 가장 작은 특징을 제외함으로써 전체적인 성능 개선을 가져올 수 있다.

IV. 실험 결과

표 1은 토석류 데이터에 사용된 5개의 특징에 대해 출력값 3개와의 연관성을 계산한 값이다. 표 1에서 볼 수 있듯이 4번째 특징의 연관성이 가장 작다. 즉, scatter plot 상에서 데이터 포인트가 특정 부분에 밀집되어 나타나는 것으로 생각할 수 있다.

전체 5개의 특징을 모두 사용하여 의사샘플 신경망을 학습시킨 경우 오류의 평균은 42.77, 분산은 28.57이었다. 이는 10 fold cross-validation을 시행하고 50회 반복 실험하여 평균한 값이다. 이에 비해 4번째 특징을 제외하고 동일한 방법으로 실험한 경우 오류의 평균은 40.28, 분산은 18.03으로 표 1의 결과와 일치함을 알 수 있다.

표 1. 입력과 출력의 연관성 ρ_n

Table 1. ρ_n between inputs and outputs

특징	연관성 ρ_n
1	4.3020
2	4.5283
3	4.2012
4	4.0719
5	4.1785

표 2. 특징 개수에 따른 오류
Table 2. Errors with respect to the number of features

제거한 특징	오류 평균	오류 분산
-	42.77	28.57
1	44.90	26.67
2	43.45	23.87
3	40.55	20.44
4	40.28	18.03
5	40.40	26.34

표 2는 5개의 입력 특징 중 하나를 제거한 후 동일한 방식으로 실험하여 결과를 비교한 것이다. 표 2의 결과에서도 알 수 있듯이 이 논문에서 제안한 연관성 정도가 특징 선택에 효과적임을 알 수 있다.

V. 결 론

이 논문에서는 신경망을 학습시키는 과정에서 입력과 출력의 연관성을 파악하여 연관성이 낮은 특징은 학습에서 제외함으로써 신경망을 성능을 향상시키는 방법을 제안하였고 이를 토석류 데이터에 적용하여 그 유효성을 보였다. 제안한 특징 선택 방법은 입력과 출력값의 scatter plot에 기초하여 특정 영역에 밀집된 특성을 보이는 특징을 제거함으로써 개선된 성능을 얻을 수 있었다. 하지만 표 2에 나타난 바와 같이 특징 1번이 제거된 경우에는 오류가 증가하였다. 비록 연관성이 큰 값을 가지기는 하지만 제거하지 말아야할 특징임을 결정할 수 있는 임계치를 결정하지는 못하였으며 현재 이에 관해 연구 중에 있다. 이후 성능을 최대화할 수 있도록 다수의 특징을 제거하는 문제로 일반화시킬 예정이다.

참고문헌

[1] C.M. Bishop, “Pattern Recognition and Machine Learning,” 2nd ed. Springer, 2007.

[2] Chang-Woo Lee, Choongshik Woo, and Ho-Joong Youn, “Analysis of Debris Flow Hazard Zone by the Optimal Parameters Extraction of Random Walk Model － Case on Debris Flow Area of Bonghwa County in Gyeongbuk Province,” Journal of Korean Forest Society Vol. 100, No. 4, pp. 664-671, 2011.

[3] Gyeongyong Heo, Chang-Woo Lee, and Choong-Shik Park, “Parameter Estimation in Debris Flow Deposition Model Using Pseudo Sample Neural Networks,” Journal of The Korean Society of Computer and Information, Vol.17, No.11, pp. 11-18, 2012.

[4] J. Rice, “Mathematical Statistics and Data Analysis,“ 2nd ed. Duxbury Press, 1995.

[5] J. Aldrich, “Correlations Genuine and Spurious in Pearson and Yule,“ Statistical Science, Vol.10, No.4, pp. 364-376, 1995.