영평균 정규화와 PCA를 이용한 회귀 신경망의 성능개선

박용수*, 조용현

대구가톨릭대학교 공과대학 컴퓨터정보통신공학부
e-mail : yhcho@cuth.cataegu.ac.kr

Performance Improvement of Regression Neural Networks by Using PCA and Zero-Mean Normalization

Yong-Soo Park*, Yong-Hyun Cho

School of Computer and Information Comm. Eng., Catholic Univ. of Daegu

요약

본 논문에서는 전처리단계로 영평균 정규화 기법과 주요성분분석 기법을 도입하여 다층신경망을 이용한 고신뢰성의 회귀분석 모델을 제안한다. 영평균 정규화 기법은 데이터의 1차적 통계성을 고려하여 알고리즘을 간략화시키며, 주요성분분석 기법은 입력 데이터의 2차적 통계성을 고려하여 독립인 특징들의 집합으로 변환시켜 학습데이터의 차원을 감소시킬 수 있어 고차원의 학습데이터에 따른 회귀분석 모델의 제약을 해결할 수 있었다. 제안된 기법의 신경망을 3개의 독립변수를 가진 암모니아 제조공정문제와 10개의 독립변수를 가진 자동차 연비문제에 각각 적용하여 시뮬레이션한 결과, 단순정규화나 PCA를 적용하지 않는 경우보다 제안된 기법의 학습속도와 회귀성능이 더욱 더 우수함을 확인할 수 있었다.

1. 서론

회귀분석은 하나의 종속변수가 다른 독립변수들에 의해 어떻게 설명 또는 예측되는지를 알아보기 위해 적절한 함수로 표현하여 자료분석을 하는 통계적인 기법이다[1-3]. 이 기법은 주어진 자료들로부터 독립변수와 종속변수의 상관관계에 대한 사전지식을 통하여 회귀분석 방정식의 모델을 설정하며 회귀계수들의 값은 통계적으로 결정한다. 하지만 문제에 따라서는 이러한 모델의 설정이나 계수들의 결정이 매우 힘든 제약들이 여전히 존재한다. 이러한 기존 수치적 기법들이 가지는 제약들을 해결하기 위해서 신경망이 널리 이용되고 있다[3-5]. 신경망을 이용하는 기법 중에는 일반회귀신경망(general regression neural network : GRNN)과 역전파(backpropagation : BP) 알고리즘의 다층신경망(multilayer perceptron : MLP)이 이용된다[3-5].

이들 중 GRNN은 우수한 성능의 회귀분석 신경망[3-5]이나 독립변수 패턴의 고차원 문제, 최적의 평활요소(smoothing factor) 및 중앙값 설정, 회귀분석을 위한 데이터 내 변화를 정확하게 측정하기 위한 많은 학습패턴 요구 등의 제약들이 있다.

한편, MLP는 충분한 뉴런을 가지고 있을 때 어떤 임의의 함수도 근사화할 수 있다고 알려져 있지만 유용한 모델을 얻기 위해서는 많은 학습데이터와 시험데이터가 요구된다. 특히, 역전파 알고리즘으로 이러한 문제를 해결하기 위한 여러 방법들이 연구되어 왔으며[5], 그중 독립변수 패턴의 고차원은 학습시간의 증가와 함께 과학습(overlearning)이 발생되어 회귀분석 성능의 저하를 가져오게 된다[1,2,4].

본 연구에서는 전처리 단계로 입력 회귀자료들을 영평균(zero-mean) 정규화하여 입력데이터로 활용함으로서 알고리즘을 간략화하고[8], 적응적 주요성분분석(principal component analysis : PCA) 기법[7,8]으로 독립변수 벡터의 특징을 추출한 다음 신경망의 입력으로 이용하였다. 제안된 기법의 신경망을 3개의 독립변수를 가진 암모니아 제조공정문제[1]와 10개의 독립변수를 가진 자동차 연비문제[2]에 각각 적용하여 시뮬레이션하여 그 타당성과 성능을 비교고찰 하였다.

2. 영평균 정규화와 주요성분분석

영평균 정규화는 데이터의 1차적인 통계성을 고려한 정규화로 학습알고리즘을 간략화하게 하는 기법이다[8]. 이는 데이터 벡터 X_k 에서 평균값 $\overline{X}_k$을 뺀 차를 구함으로써 데이터의 영평균을 구하고 이를 다시 최대값으로 나누어 구할 수 있다.

한편, 데이터의 2차적 통계성을 고려한 상호간의 의존성을 줄이기 위한 기법으로 whitening이 이용되고 있다[8].

whitening은 데이터 벡터 X_k 의 공분산행렬이 단위행렬 값을 갖도록 함으로써 구할 수 있다. 즉, $E\{ X_k \ X_k^T\} = I$ 가 되도록 한다. 이렇게 함으로써 whiteness된 벡터의 성분들 상호간의 상관성이 줄어들어 강한 독립성분이 된다[8]. 이러한 whitening의 기법으로 PCA가 널리 사용된다. PCA는 공분산행렬의 고유벡터와 고유치를 추정하는 수치적 기법으로 이루어질 수 있다. 하지만 이러한 수치적 기법에서는 대규모의 차원을 가질 때, 공분산행렬의 계산이 매우 증가되고, 그에 따른 고유치 벡터를 찾는 것도 매우 복잡하게 된다.

기존 수치적 기법들은 대규모의 실시간 처리가 요구되는 응용문제에서는 비효율적이므로, 그 대안으로 상관행렬의 고유벡터를 실시간으로 추정하는 적응적 학습알고리즘의 신경망을 이용하는 방법들이 제안되었다[3,5,7]. 여기서는 상관행렬의 추정과정이 요구되지 않으며, 이때 이용되는 신경망은 주로 입력층과 출력층으로 구성된 단층구조이다.

단층구조의 신경망으로 Oja[5]는 정규화된 헤비안규칙(normalized Hebbian rule)의 적응학습 방법을 이용한 단일뉴런모델을 제안하였고, Sanger등[5]은 일반화된(generalized) 헤비안규칙을 이용함으로써 정상과정의 m개의 가장 중요한 주요특징들을 계산하기 위한 다중뉴런모델을 제안하였다. 또한, Foldiak[5,6,7]은 망의 입력과 출력사이의 연결가중치 경신에는 정규화된 헤비안규칙을 이용하고, 망의 출력사이의 측면연결 가중치 경신에는 반 헤비안규칙(anti-Hebbian rule)을 함께 이용한 학습알고리즘을 제안하였다. Foldiak에 의해 제안된 학습이 수렴속도 면에서 더 우수한 것으로 알려 있어[5,7] 본 논문에서는 이 방법을 이용하였다.

n개의 입력뉴런과 m개의 출력뉴런으로 구성된 입력과 출력뉴런간 및 출력뉴런 상호간의 측면연결을 가진 단층 신경망의 구조에서 입력과 출력의 관계를 나타내면 다음과 같다. 즉

$$y_i = \sum_{j=1}^{n} w_{ij} x_j + \sum_{h=1}^{i-1} u_{ih} y_h , \quad (i = 1,2,\cdots,m) \qquad (1)$$

이다. 여기서 w_{ij}는 입력뉴런과 출력뉴런을 연결하는 연결가중치이고, u_{ih}는 출력뉴런 상호간의 측면연결 가중치이다. 이때 w_{ij} 와 u_{ih} 의 경신규칙은 각각

$$w_{ij}(t+1) = w_{ij}(t) + \eta[y_i(t) x_j(t) - w_{ij}(t) y_i(t)^2], \quad (i=1,2,\cdots,m, \ j=1,2,\cdots,n) \qquad (2)$$

$$u_{ih}(t+1) = u_{ih}(t) + \rho y_i(t) y_h(t), \quad (i>h) \qquad (3)$$

이다. 결국 식 (2)와 (3)에 따라 연결가중치를 경신시켜 식 (1)에 대입하면 m 개의 주요특징들을 추출할 수 있다.

따라서 영평균 정규화는 1차적인 통계에 따른 데이터의 정규화로 학습알고리즘을 간략하게 하고, PCA는 2차적인 통계에 따른 데이터 상호간의 의존성을 줄여 독립성을 강하게 함으로써 다음 과정을 쉽게 한다. 또한 PCA는 입력데이터 내에 존재하는 주요특징을 추출하여 감소시킬 수 있어 이를 회귀분석 신경망의 전처리단계로 이용한다면 MLP 입력층의 뉴런개수가 줄어들어 학습을 위한 시간이 감소되고, 학습데이터의 고차원에 따른 독립변수들 상호간의 다중공선성 등으로 인한 회귀성능의 저하와 과학습과 같은 제약들도 함께 해결할 수 있다.

3. 시뮬레이션 및 결과분석

제안된 기법의 회귀 신경망의 성능을 평가하기 위해 Foldiak 학습알고리즘의 단층신경망과 3층의 다층신경망을 각각 구성하였다. 학습은 전체 학습반복수가 20,000이상이거나 전체 오차값이 허용치 10^{-4}이하일 때 종료되도록 하였다. 또한 역전파 학습알고리즘에서 학습율과 모멘트는 각각 0.1과 0.5로 하였다.

3.1 암모니아 제조공정문제

독립변수로 3개의 공정 입력조건 x_1, x_2, x_3와 종속변수로 1개의 출력특성 y를 가지는 21개의 패턴들로 구성되며, 실험에서 16개는 학습패턴으로 나머지 5개는 시험패턴으로 이용하였다. 이용된 MLP에서 입력층 뉴런수는 입력조건인 독립변수의 개수와 동일하게 하고, 은닉층과 출력층 뉴런수는 각각 3개와 1개로 구성하였다.

표 1은 실험에 이용된 자료의 일부를 나타낸 것이다. 실험에서는 전처리 단계로 입·출력 조건을 각각 최대값으로 나누어 1이하의 값으로 단순히 정규화한 데이터를 사용한 경우와 입·출력 조건에서 평균값을 뺀 차를 정규화한 영평균 정규화 데이터를 각각 이용하였다.

표 1. 암모니아 제조공정자료

패턴수	입력조건			출력특성 y
	x_1	x_2	x_3	
1	80	27	89	42
2	80	27	88	37
⋮	⋮	⋮	⋮	
15	50	18	89	8
16	50	18	86	7
⋮	⋮	⋮	⋮	
21	50	20	91	15

표 2는 입력조건의 단순한 정규화와 영평균 정규화 패턴 각각에 대하여 PCA를 적용한 경우와 그렇지 않는 경우에 대한 결과이다. 여기에서 PCA의 적용은 차원의 감소를 위함이 아니고 패턴내의 상관성을 줄이고 분산을 1로 함으로써 입력조건들의 독립성을 증가시키기 위함이다. 표에서는 학습반복수 N_i와 학습에 소요된 CPU 시간 t_{CPU}, 그리고 출력 y와 학습 후 출력 y^와의 절대오차(absolute error : AE, AE = $\sum_{i=1}^{n} | y - y^ |$)로 표현되는 학습오차 E_l, 시험오차 E_t 및 합오차 E_s를 나타낸 것이다. 표에서 보면 단순 정규화와 영평균 정규화 및 PCA의 적용유무에 관계없이 학습의 종료조건은 만족되지 않았다. 이는 이용된 다층신경망의 학습알고리즘으로 역전파 알고리즘을 이

용하기 때문이다. 또한 표에서 절대오차 중 학습오차를 보면, 영평균 정규화 패턴이 단순한 정규화 패턴에 비해 최대 약 2.2배 정도, PCA를 적용한 경우가 적용하지 않은 경우보다 최대 약 1.4배 정도 더 작은 값을 가진다. 이는 동일한 학습반복수에서 영평균 정규화 패턴에 PCA를 적용하면 학습오차 면에서 더욱 우수한 학습성능이 있음을 알 수 있다. 또한 5개의 시험패턴에 대한 시험오차에서도 영평균 정규화 패턴에 PCA를 적용한 경우가 가장 작은 오차를 나타내어 회귀성능도 향상됨을 알 수 있다. 한편 합오차에서도 영평균 정규화 패턴에 PCA를 적용한 경우가 단순 정규화 패턴을 그대로 이용하는 것에 비해 약 1.8배 정도 작은 값을 가진다. 따라서 회귀분석을 위한 신경망의 입력패턴은 영평균 정규화뿐만 아니라 PCA를 함께 적용하면 패턴 상호간의 종속성을 더욱 감소시켜 학습성능과 회귀성능이 개선됨을 알 수 있다.

표 2. 단순 정규화 및 영평균 정규화 패턴에 대한 PCA 적용유무에 따른 학습결과

		Normalized pattern data		Zero-mean normalized pattern data	
		non-PCA	PCA	non-PCA	PCA
학습반복수, N_i		20000	20000	20000	20000
학습시간, t_{CPU}		35	35	35	35
절대오차, AE	학습오차,E_t	0.071116	0.065015	0.040833	0.029748
	시험오차,E_t	0.567587	0.509047	0.371189	0.329392
	합오차,E_s	0.638703	0.574062	0.412022	0.35914

표 3은 회귀 신경망을 대상으로 데이터를 영평균 정규화한 후 PCA를 적용하여 입력조건인 독립변수 개수 N을 조정하면서 학습한 결과를 나타낸 것이다. 표에서 독립변수 수가 3개인 경우는 단순히 패턴의 독립성을 증가시키기 위함으로 PCA가 가지는 차원의 감소는 이루어지지 않은 경우이다. 표에서는 차원감소에 따라 구해진 2개와 1개의 특징들을 독립변수로 이용하는 경우가 3개의 독립변수를 입력조건으로 이용하는 경우보다 합오차가 작아 상대적으로 우수한 회귀성능이 있음을 알 수 있다. 이는 문제에 이용되는 3개의 독립변수 상호간에 선형적인 연관성이 많음을 추측할 수 있다. 또한 표에서는 독립변수가 2개인 경우보다 1개인 경우가 우수한 회귀성능을 가져, 패턴상호간의 상관성은 회귀성능을 저하시키는 결과를 초래함을 알 수 있다. 따라서 회귀 신경망에서 제안된 기법의 PCA

표 3. 입력조건의 변화에 따른 학습결과

독립변수 수, N		3	2	1
학습반복수, N_i		20000	20000	20000
학습시간, t_{CPU}		35	34	33
절대오차, AE	학습오차, E_t	0.029748	0.022531	0.074366
	시험오차, E_t	0.329392	0.301452	0.239795
	합오차, E_s	0.35914	0.323983	0.314161

를 이용하면 신경망의 학습성능도 상대적으로 개선되었다.

3.2 자동차 연비문제

10개의 독립변수에 대하여 1개의 종속변수를 가진 회귀분석 문제이다. 이 문제는 실린더 수(x_1), 입방인치 단위의 배기량(x_2), 마력(x_3), 최종기어비인 최종구동장치비(x_4), 자동차 무게(x_5), 순간가속도 (x_6), V자형(0) 또는 직선형(1)의 엔진형태(x_7), 자동(0)이냐 수동(1)이냐의 기어의 종류(x_8), 기어속도의 수(x_9), 그리고 기화기의 밸브수(x_{10})에 대한 자동차 연비(y)와의 관계를 나타낸 것이다. 본 실험에서는 32개의 패턴데이터 중에서 25개는 학습패턴으로 나머지 7개는 시험패턴으로 이용하였다. 이용된 MLP에서 입력층 뉴런수는 입력조건인 독립변수의 개수와 동일하게 하고, 은닉층과 출력층 뉴런수는 각각 10개와 1개로 구성하였다.

표 4는 실험에 이용된 자료의 일부를 나타낸 것이다. 여기에서도 입·출력 조건을 각각 최대값으로 나누어 1이하의 값으로 단순 정규화한 데이터와 영평균 정규화 데이터를 각각 실험에 이용하였다.

표 4. 자동차 연비자료

패턴수	입력조건										출력 특성
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	6	160	110	3.9	2.62	16.46	0	1	4	4	21
2	6	160	110	3.9	2.875	17.02	0	1	4	4	21
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	8	350	245	3.73	3.84	17.05	0	0	3	4	13.3
25	8	400	175	3.08	3.845	18.9	0	0	3	2	19.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
32	4	121	109	4.11	2.78	18.6	1	1	4	2	21.4

표 5는 입력조건의 단순한 정규화 및 영평균 정규화 패턴 각각에 대해 PCA 적용 유무에 따른 학습결과를 나타낸 것이다. 여기서도 PCA의 적용은 입력조건 상호간의 독립성을 증가시키기 위함이다. 표에서도 표 2에서처럼 학습의 종료조건은 만족되지 않았다. 하지만 학습오차를 살펴보면, 영평균 정규화 패턴은 단순한 정규화 패턴에 비해 최대 약 168.7배 정도, PCA를 적용한 경우가 적용하지 않

표 5. 단순 정규화 및 영평균 정규화 패턴에 대한 PCA 적용유무에 따른 학습결과

		Normalized pattern data		Zero-mean normalized pattern data	
		non-PCA	PCA	non-PCA	PCA
학습반복수,N_i		20000	20000	20000	20000
학습시간,t_{CPU}		37	37	37	37
절대오차, AE	학습오차,E_t	0.042676	0.024126	0.000447	0.000143
	시험오차,E_t	1.952232	0.857857	0.67318	0.425316
	합오차, E_s	1.994908	0.881983	0.673627	0.425459

은 경우에 비해 최대 약 3.1배 정도 더 작은 값을 가진다. 한편 학습오차와 시험오차의 합으로 표현되는 합오차에서도, 영평균 정규화와 PCA를 적용한 패턴을 이용한 경우가 그렇지 않는 경우에 비해 최대 약 4.7배 정도 작은 값을 가진다. 한편 표 2의 암모니아 제조문제와 비교할 때, 문제의 규모가 증가된 연비문제에서 학습성능과 회귀성능의 개선정도는 더욱 더 증가됨을 확인할 수 있다.

표 6은 회귀용 신경망을 대상으로 자동차연비 자료데이터를 영평균 정규화한 후 PCA를 적용하여 독립변수 수 N을 조정하면서 학습한 결과를 나타낸 것이다. 독립변수 9에서 2까지는 차원을 감소시켜 이를 입력조건으로 학습한 경우이다. 차원을 감소시켜 학습시킬 때 좀 더 개선된 학습성능과 회귀성능을 얻을 수 있다. 독립변수의 수가 5개 이하일 때에는 학습조건은 만족되나 시험오차가 증가되어 과학습에 따른 일반화 성능이 감소된 것으로 추측된다. 독립변수의 개수가 9개에서 5개 사이는 학습성능과 회귀성능이 상대적으로 우수하지만 독립변수가 4개 이하인 경우는 상대적으로 저하된 성능을 보인다. 이는 10개의 입력조건들의 속성이 4개 이하에서는 충분히 반영되지 못하기 때문으로 추측된다. 특히 독립변수가 8개와 7인 경우에는 각각 종료조건을 만족하여 이에 따른 학습속도 및 회귀성능의 증대를 확인할 수 있다.

표6. 입력조건의 변화에 따른 제안된 알고리즘의 학습결과

독립변수 수, N		10	9	8	7
학습반복수, N		20000	20000	19957	19881
학습시간, t		37	37	35	34
절대오차, AE	학습오차, E_t	0.000143	0.000128	0.0001	0.0001
	시험오차, E_v	0.425316	0.420921	0.106432	0.091259
	합오차, E_a	0.425459	0.421049	0.106532	0.091359

6	5	4	3	2
20000	20000	20000	20000	20000
34	34	33	33	33
0.000119	0.009311	0.024663	0.517977	0.838667
0.37485	1.295529	1.485278	2.498355	2.148351
0.374969	1.30484	1.509941	3.016332	2.987018

이상의 결과들로부터 PCA에 의한 독립변수의 차원 감소는 각 변수 상호간에 많은 다중공선성을 가지는 회귀분석 문제에서는 우수한 성능이 있음을 추측할 수 있으며, 특히 입력조건 상호간에 선형적 상관성이 많을수록 더욱 더 우수한 성능이 있을 것이다. 이는 PCA 기법이 데이터 상호간의 2차적 통계성에 바탕을 두고 주요성분인 특징을 추출하여 차원을 감소시키기 때문이다. 또한 제안된 기법은 문제의 규모가 증가할수록 학습성능과 회귀성능의 개선폭이 증가됨을 확인할 수 있다.

4. 결 론

본 논문에서는 전처리단계로 데이터의 1차원적 통계성을 고려한 영평균 정규화 기법과 데이터의 2차원적 통계성을 고려한 주요성분분석 기법을 다층신경망의 전단에 이용함으로써 빠른 학습속도의 고신뢰성을 가지는 회귀분석 신경망 모델을 제안하였다.

제안된 기법의 회귀 신경망을 3개의 독립변수를 가진 21개 학습패턴의 암모니아 제조공정문제와 10개의 독립변수를 가진 32개 학습패턴의 자동차 연비문제에 각각 적용하여 시뮬레이션한 결과, 학습패턴의 영평균 정규화 및 PCA를 적용한 경우가 그렇지 않은 경우에 비해 개선된 학습성능과 회귀성능이 있음을 확인하였다. 또한 학습패턴의 차원을 감소시키지 않은 경우와 비교할 때 학습성능과 절대오차로 나타나는 회귀성능도 더욱 우수함을 확인할 수 있었다. 특히 문제의 규모가 증가할수록 제안된 알고리즘의 회귀 신경망은 더욱 더 성능개선의 정도가 증가됨을 확인하였다.

향후 제안된 기법의 신경망을 좀 다양한 회귀분석 문제에 적용하고 회귀검증도 함께 고려되는 연구가 계속되어야 할 것이다.

참고문헌

[1] 허명회 외1, "SAS 회귀분석", 자유아카데미, 1996

[2] 강명욱 외3, "회귀분석 : 모형개발과 진단", 율곡출판사, 1997

[3] S. Haykin, 'Neural Networks : A Comprehensive Foundation', Prentice-Hall, London, 1999

[4] D. F. Specht, "A General Regression Neural Network," IEEE Trans. on Neural Networks, vol.2, no. 6, pp.568-576, Nov. 1991

[5] A. Cichocki and R. Unbehaun,'Neural Networks for Optimization and Signal Processing', John Wiley & Sons., New York, 1993

[6] P. Foldiak, "Adaptive Network for Optimal Linear Feature Extraction," International Joint Conference on Neural Networks, Washington D.C., vol 1, pp.401-406, June 1989

[7] K. I. Diamantaras and S. Y. Kung, 'Principal Component Neural Networks : Theory and Applications, Adaptive and Learning Systems for Signal Processing, Communications, and Control', John Wiley & Sons, Inc., 1996

[8] A. Hyvarinen and E. Oja, " Independent Component Analysis: Algorithms and Applications," Neural Networks, vol. 13, no. 4-5, pp.411-430, 2000

[9] 조용현, 윤중환, 박용수, "조합형 학습알고리즘의 신경망을 이용한 데이터의 효율적인 특징추출", 정보처리학회논문지, 제 8-B권 제 2호, pp.130-136, April 2001