칼만 필터링을 이용한 Mixtures of Experts network 학습

Learning of Mixtures of Experts Network Based on Kalman Filtering

김병관, 최우경, 김성주, 김종수, 서재용*, 전홍태
중앙대학교 전자전기공학부
*한국기술교육대학교 정보기술공학부
전화 : 02-820-5297

Byeongkwoan Kim, Wookyung Choi, Seongjoo Kim, Jongsu Kim *Jaeyong Seo and Hongtae Jeon
School of Electrical and Electronic Engineering, Chung-Ang University
*School of Information and Technology. Korea University of Technology and Education
E-mail : schopenhauer@empal.com

요 약

복잡한 문제 학습을 위해 여러 가지 형태의 모듈라 네트워크의 구조가 제시되어 왔다. 그 중 엑스퍼트 네트워크와 게이팅 네트워크로 구성된 Mixtures of Experts network은 복잡한 문제를 단순한 문제들로 분해하고, 각각의 엑스퍼트 네트워크가 분해된 단순한 문제를 학습하여 결과를 도출함으로써, 국소적 지역해의 위험을 방지하고 보다 정확한 학습을 가능하게 한다. 그러나 엑스퍼트 네트워크의 수렴은 게이팅 네트워크의 수렴에 많은 영향을 받게 되고, 모든 복잡한 데이터에 대한 엑스퍼트 네트워크의 기여도를 학습하는 게이팅 네트워크는 역전파 알고리즘에 의한 학습 방법으로는 수렴 속도가 떨어진다. 본 논문에서는 게이팅 네트워크를 칼만필터로 학습하여 복잡한 문제에 대한 강건성은 유지하고 보다 빠른 수렴이 가능한 방법을 제시하고자 한다.

1. 서론

1940년대 초에 뉴런을 모델로 공학적 응용을 한 이후 학습 이론 및 신경 회로망 이론이 활성화되기 시작 했다. 특히 다층 구조의 신경망이 제시된 이후 신경망의 적용이 활발히 이루어져 왔으나, non-modular neural network의 늦은 학습과 국소적 지역해에 대한 위험 때문에 복잡한 문제에 대한 적용이 어려웠다[1,2,3]. 이를 위해 모듈라 형태의 네트워크 구조가 제시되어 왔고, 그 중 Mixtures of Experts Network(ME) 활발하게 연구되고 있다[4,5,6]. ME는 복잡한 원 문제를 단순한 부문제로 분리하고 학습한 후 다시 통합하는 분할 후 점령 기법(divide-and conquer technique)으로 학습이 진행되며 각 부문제는 여러 개의 엑스퍼트 네트워크(Expert Network:EN)이 학습을 담당하게 되고 게이팅 네트워크(Gating Network:GN)은 각 문제에 대한 기여 정도를 학습하게 된다. 여러 개의 EN과 하나의 GN으로 구성된 ME의 수렴은 EN과 GN의 각각의 수렴에 영향을 받게 되고, 특히 GN은 각 EN에 대한 기여도를 결정하므로 그 수렴의 정도가 전체 네트워크의 수렴에 많은 영향을 끼치게 된다. 이에 비교적 쉬운 부문제를 학습하는 EN은 일반적인 역전파 알고리즘으로 빠른 수렴성이 요구되는 GN은 칼만 필터링 기법으로 학습하므로써 전체 네트워크의 수학적 복잡도와 수렴성에 대하여 균형을 맞추는 방법을 제시하고자 한다.

2. 본론

2.1 Mixtures of Experts Network

신경망은 인공지능을 대표하는 대표적인 알고리즘으로 학습기능으로 많이 사용되어왔다. 그 중 초기 모델인 non-modular neural network 구조는 은닉층에서의 높은 커플링으로 인한 늦은 학습과 overfitting 때문에 복잡한 문제에 대한 수렴성이 떨어진다[1]. 이를 극복하기 위하여 여러 가지의 모듈라 신경망이 제시되었다. 모듈라 신경망을 사용하는데 있어서 핵심은 복잡한 원문제를 간단한 부문제로 분할하는 방법, 각 부문제를 각각의 모듈들에 배당하는 방법, 마지막으로 각각의 모듈이 제공한 결과가 원문제를 재결합하는 문제가 중요하게 작용한다[7]. 지금까지 많은 종류의 분할 후 점령 기법(divide-and conquer technique)에 바탕으로 한 task decomposition methods가 연구되어왔고 대체적으로 Explicit Decomposition, Class Decomposition 그리고 Automatic Decomposition으로 구분할 수 있다[7]. 이 중 마지막 방법은 원문제에 대한 사전 지식이 없어도 decom-position 가능하다는 일반성 때문에 크고 복잡한 문제를 해결하는데 가장 적당하다고 알려져 있다[7].

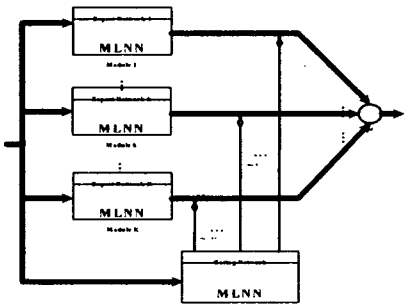

그림 1. Mixture of Experts Network의 구조
Fig. 1. The Structure of Mixture of Experts Network

Automatic Decomposition을 사용하는 구조 가장 잘 알려진 Jacobs와 Jordan[4,5,6]이 제안한 ME의 구조는 그림1과 같다. ME은 EN이라고 불리는 M개의 교사모듈(Supervised Module)과 각각의 EN 사이에 중개역할을 담당하는 GN으로 구성되어 있다.

그림에서 $\mathbf{x}=[x_1 x_2 \cdots x_P]^T$는 P차원의 입력벡터, $\mathbf{y}=[y_1 y_2 \cdots y_Q]^T$는 ME의 최종출력 벡터, $\mathbf{y_k}=[y_k^1 y_k^2 \cdots y_k^Q]^T$는 k번째 EN의 Q*1 출력벡터이다. 입력벡터 x가 EN과 GN에 동시에 인가될 때 ME의 최종 출력은 다음 식과 같다.

$$\mathbf{y}=\sum_{k=1}^{K} g_k \mathbf{y}_k \tag{1}$$

여기서 K는 모듈의 개수이고, $\mathbf{g_k}$는 k번째 GN의 출력이며, $\mathbf{y_k}$는 k번째 EN의 출력 벡터이다. GN의 k번째 가중치(weight)와 입력의 곱의 합이 $\mathbf{u_k}$일 때, 입력벡터 x와 관련된 사전 확률(Priori Probability)인 GN의 최종 출력은 정규화된 soft-max 함수를 사용하여 다음과 같이 표현할 수 있다.

$$g_k=\frac{\exp(u_k)}{\sum_{j=1}^{K}\exp(u_j)} \tag{2}$$

여기서 $\mathbf{g_k}$는 모든 k에 대해서 $0 \le g_k \le 1$과 $\sum_{k=1}^{K} g_k = 1$을 만족해야 한다.

EN의 출력과 관련한 사후확률(Posterior Probability)은 다음과 같이 정의할 수 있다.

$$h_k=\frac{\frac{1}{(2\pi)^{Q/2}}\frac{g_k}{\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}\|\mathbf{d}-\mathbf{y}_k\|^2\right)}{\frac{1}{(2\pi)^{Q/2}}\sum_{j=1}^{K}\frac{g_j}{\sigma_j}\exp\left(-\frac{1}{2\sigma_k^2}\|\mathbf{d}-\mathbf{y}_k\|^2\right)}, \quad k=1,2,\cdots,K \tag{3}$$

여기서 $\mathbf{d}=[d^1 d^2 \cdots d^Q]^T$는 원하는 응답벡터이고 Q는 GN의 출력의 차원, σ_k는 입력벡터의 공분산을 의미한다. $\mathbf{h_k}$는 사전확률과 유사하게 모든 k에 대해서 $0 \le h_k \le 1$과 $\sum_{k=1}^{K} h_k = 1$을 만족해야 한다.

2.2 학습 방법

다층신경망(Multi-Layered Neural Network : MLNN) 을 학습하기 위하여 여러 가지 방법이 제시되어 왔고 그 중 역전파 알고리즘은 네트워크의 가중치를 학습시키기 위해 전형적으로 쓰이는 방법이다. 역전파 알고리즘은 네트워크의 크기가 작거나 단순한 문제를 학습하는 데는 효과적이나 네트워크의 크기가 커지거나 문제가 복잡해 질수록 수렴성이 떨어진다는 단점이 있다. 이는 역전파 알고리즘이 LMS(Least Mean Square) gradient 방법을 사용하기 때문에 과거의 가중치에 대한 정보를 무시하기 때문이다. 이를 해결하기 위해 칼만 필터를 이용한 가중치의 학습 방법이 Sharda Singhal에 의해 제시되

었다[1]. LMS gradient 방법과는 달리 칼만 필터링 방법은 가중치의 최적 값을 과거의 데이터 값에 의해 구하는 방식이다. 그러므로 칼만 필터 방법은 계산상의 복잡도는 증가하지만 과거의 가중치와 연관성 있는 최적 가중치를 구하여 학습하므로 복잡한 문제에 대해서도 수렴성을 보장할 수 있다.

이에 본 논문에서는 단순한 부문제를 학습하는 여러 개의 EN은 역전파 알고리즘으로 학습하고 전체 데이터에 대한 EN의 기여도를 학습하는 GN은 칼만 필터링 방법으로 학습하여 수렴성과 계산량 사이에 균형을 맞추고자 한다.

2.2.1 역전파 알고리즘에 의한 EN 학습

단순한 부문제를 학습하는 EN의 경우 많은 계산량과 메모리를 요구하는 칼만 필터를 경우 EN의 수만큼 계산량과 메모리가 증가한다. 또한 단순한 문제에 대하여 역전파 알고리즘과 칼만 필터링의 수렴 속도는 눈에 띌 만큼 향상되지 않는다고 연구되었다[8]. 그러므로 본 논문에서는 단순한 부문제를 학습하는 EN의 학습은 역전파 알고리즘에 의해서 수행하였다[4].

2.2.2 칼만 필터링 방법에 의한 GN 학습

칼만 필터링에 적용하기 위해 GN은 가중치 조절 과정은 비선형 형태인 다음식과 같이 모델링 될 수 있다[1].

$$\mathbf{w}_{k+1} = \mathbf{w}_k$$
$$\mathbf{h}_k = \mathbf{g}_k + \mathbf{v}_k = \mathbf{f}_k(\mathbf{w}_n, \mathbf{x}_n) + \mathbf{v}_k \quad (4)$$

여기서 $\mathbf{w}_k$는 가중치, $\mathbf{h}_k$는 사후 확률, $\mathbf{g}_k$는 사전 확률, $\mathbf{f}_k$는 활성화 함수 $\mathbf{v}_k$ 관찰 노이즈를 의미하고 $\mathbf{w}_k$와 $\mathbf{v}_k$는다음과 같이 정의 된다.

$$E[\mathbf{w}_0] = \bar{\mathbf{w}}_0, \; E\{[\mathbf{w}_0 - \bar{\mathbf{w}}_0][\mathbf{w}_0 - \bar{\mathbf{w}}_0]^T\} = \mathbf{P}_0$$
$$E[\mathbf{v}_k] = 0, \; E[\mathbf{v}_k \mathbf{v}_n^T] = \mathbf{R}_k \delta_{kn} \quad (5)$$

식(4)에 칼만 필터링의 회귀 과정을 적용하면 다음과 같이 단순화 된다[1].

$$\hat{\mathbf{w}}_{k+1} = \hat{\mathbf{w}}_k + \mathbf{K}_k(\mathbf{h}_k - \mathbf{g}_k)$$
$$\mathbf{K}_k = \mathbf{P}_k \mathbf{F}_k [\mathbf{R}_k + \mathbf{F}_k^T \mathbf{P}_k \mathbf{F}_k]^{-1}$$
$$\mathbf{P}_{k+1} = \mathbf{P}_k - \mathbf{K}_k \mathbf{F}_k^T \mathbf{P}_k \quad (6)$$

여기서 $\mathbf{K}_k$는 칼만 계수, $\mathbf{F}_k$는 비선형 활성화 함수, $\mathbf{R}_k$는 $\mathbf{v}_k$의 공분산, $\mathbf{P}_k$는 $\mathbf{w}_k$에 대한 예측 에러 공분산 (approximate error covariance) 을 의미한다.

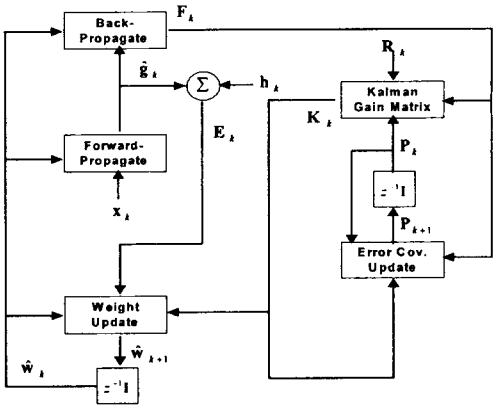

그림 2. GN 학습을 위한 칼만필터링 흐름도
Fig. 2. Signal flow for KF to learn GN

그림2는 칼만 필터링 방법의 과정을 묘사한 것이다. 그림에서 알 수 있듯이 역전파 알고리즘에 칼만 계수 $\mathbf{K}_k$와 $\mathbf{w}_k$에대한 $\mathbf{P}_k$를 구하는 과정이 더해져 수학적 복잡도가 증가하는 것을 알 수 있다.

3. 실험 및 결과

본 논문에서는 단일 은닉층인 4개의 EN과 1개의 GN으로 ME를 구성하였으며, EN은 은닉층은 4개의 노드로 GN의 은닉층은 7개의 노드로 구성되어 있고 각 노드의 활성화 함수로는 sigmoid 함수를 사용하였다. GN에서는 기존의 역전파 알고리즘과 칼만 필터링을 이용한 방법을 비교하기 위하여 다음과 같이 파라미터를 설정하였다. 역전파 알고리즘에서의 학습률은 0.05, 모멘텀을 이용할 경우 모멘텀은 0.9로 설정하였고 칼만 필터링 알고리즘의 계수로 $R_k = \lambda I$로 정의 하였고 λ는 0.01, P_0는 0.1로 초기화 하였다[9]. 학습데이타는 Mobile Robot의 추종 실험을 위해 만들어진 입력 8차원 벡터와 출력 4차원 벡터 쌍 총 1000개를 이용하였다.

그림3은 각 데이터 쌍에 대해 5000번을 반복하여 총 10번 학습된 결과에 대한 RMSE의 평균값이다. 칼만 필터링을 이용한 경우 적은 반복 회수로도 역전파 알고리즘보다 빠른 수렴 속도를 보인다는 것을 알 수 있다.이는 GN에서 사용한 칼만 필터링 방법이 과거의 가중치와 데이터에 대한 값을 저장하여 다음 단계의 학습에서 각 EN에 대한 기여도를 빠르게 결정으로써 각 EN이 학습해야 할 부문제가 좀더 빠르게 결정

될 수 있었기 때문이다.

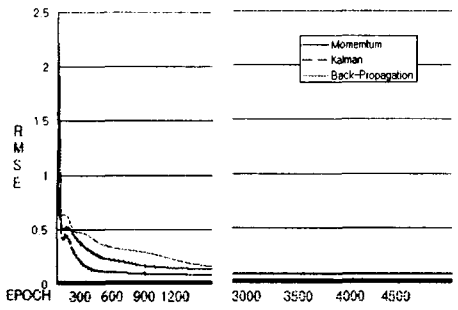

그림 3. 3가지 학습 방법에 따른 RMSE
Fig. 3. RMSEs with 3 different learning method

또한 반복횟수가 증가할수록 역전파 알고리즘과 칼만 필터링은 둘 다 목표치까지 근접해가는 것을 알 수 있으나 초기 수렴 정도는 칼만 필터링에 의한 학습 방법이 월등히 향상됨을 알 수 있다.

4. 결론

본 논문에서는 ME의 GN학습을 역전파 알고리즘 방법과 칼만 필터링 방법을 이용하여 학습하여 비교해 보았다. 칼만 필터링 방법은 과거의 데이터 값을 이용함으로써 보다 빠른 수렴이 가능하였다. 또한 EN은 기존의 역전파 알고리즘을 이용하여 수학적 간결성을 유지하고 GN에만 칼만 필터링 방법을 이용함으로써 복잡한 문제 해결을 위해 EN의 개수를 늘려도 전체 계산량은 유지한 채 수렴 속도를 향상하는 장점이 있다. 그러므로 이 논문에서는 역전파 알고리즘과 칼만 필터링의 장단점을 ME의 학습에 적절히 배치함으로써 수학적 간결성과 수렴속도 사이의 적절한 균형(trade-off)을 제시한 학습 기법을 실험하였다.

감사의 글 : 본 연구는 과학기술부의 뇌신경정보학 연구사업에 의해 지원받았습니다.

5. 참고문헌

[1] Sharad Singhal and Lance Wu, "Training multilayer perceptrons with the extended Kalman algorithm," *Advaces in Neural Information Processing Systems 1*, D. S. Touretzky, Ed. San Matero, CA: Morgan Kaufmann, pp. 133-140, 1989.

[2]Martin Birgmeir, "A Neural Network Trained with the Extended Kalman Algorithm Used for the Equalization of a Binary Communication Channel", *Proc. NNSP IV*, pp. 527-534, 1994.

[3] Youji Iiguni, Hideaki Sakai and Hidekatsu Tokumaru, "A Real-Time Learning Algorithm for a Multilayered Neural Network Based on the Extended Kalman Filter," *IEEE Trans. Signal Processing*, Vol. 40, pp. 959-966, 1992.

[4] Simon Haykin, *Neural Networks-A Comprehensive Foundation*, Macmillian College Publishing Company Inc., 1994.

[5]R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E Hinton, "Adaptive Mixtures of Local Experts", *Neural Computation*, Vol. 6, pp.181-214, 1994.

[6] Craig, J. J, *Introduction to Robotics: Mechanics and Control*, Addison-Wesley, 1988.

[7] Bao-Liang Lu and Massmai Ito, "Task Decompostion and Module Combination Based on Calss Relations: A Modular Neural Network for Pattern Classification", *IEEE Trans. Neural Networks*, Vol. 10 pp. 1244-1255, 1999.

[8] Dennis W. Ruck, Steven K. Rogers, Matthew Kabristky, Peter S. Maybeck and Mar E. Oxely, "Comparative Analysis of Backpropagation and the Extended Kalman Filter for Training Multilayer Perceptrons", *IEEE Trans. Pattern Anlaysis and Machine Intelligence*, Vol. 14, pp.686-691, 1992.

[9]Čerňanský M., Beňušková L. "Simple recurrent network trained by RTRL and extended Kalman filter algorithms,"*Neural Network World, Institute of Computer Science in Czech Academy of Science*, Vol 13, pp. 223-234, 2003.