전문화된 네트워크들의 결합에 의한 앙상블 학습 알고리즘

신현정[U] 이형주 조성준
서울대학교 산업공학과
{hjshin72, impatton, zoon}@snu.ac.kr

Ensemble Learning Algorithm of Specialized Networks

Hyunjung Shin[U] Hyoungjoo Lee Sungzoon Cho
Dept. of Industrial Engineering, Seoul National University

요 약

관찰학습(OLA: Observational Learning Algorithm)은 앙상블 네트워크의 각 구성 모델들이 다른 모델들을 관찰함으로써 얻어진 가상 데이터와 초기에 bootstrap된 실제 데이터를 학습에 함께 이용하는 방법이다. 본 논문에서는, 초기 학습 데이터 셋을 분할하고 분할된 각 데이터 셋에 대하여 앙상블의 구성 모델들을 전문화(specialize)시키는 방법을 적용하여 기존의 관찰학습 알고리즘을 개선시켰다. 제안된 알고리즘은 bagging 및 boosting과의 비교 실험에 의하여, 보다 적은 수의 구성 모델로 동일 내지 보다 나은 성능을 나타냄이 실험적으로 검증되었다.

1. 서 론

관찰학습(OLA: Observational Learning Algorithm)은 앙상블 네트워크의 각 구성 모델들이 다른 모델들을 관찰함으로써 얻어진 가상 데이터와 초기에 bootstrap된 실제 데이터를 학습에 함께 이용하는 방법이다[1][2](그림1 참조). 관찰학습에 사용되는 가상데이터는 overfitting을 방지하고 구성 네트워크간의 총론(consensus)를 유도하는 역할을 한다. 이는 bagging 및 boosting과의 비교실험을 통하여 성능의 우수성에 대한 유의함이 실험적으로 검증된 바 있다[3]. 앙상블 학습에서는 구성 네트워크의 학습에러간에 상관성이 작을수록 더 좋은 결과를 산출한다[5]. 구성 네트워크간의 독립성은 네트워크별로 그 구조나 학습데이터 셋을 다르게 만듦으로서 얻어질 수 있는데, 기존의 관찰학습에서는 bootstrapping을 통한 후자의 방법을 택하고 있다. 그러나 bootstrapping으로 얻어진 서브 데이터 셋들은 서로 동일하지는 않더라도 동일한 데이터 셋으로부터 추출되었으므로 확률적으로는 동일하다. 따라서 각 학습 데이터 셋을 확률적으로도 다르게 구성하고 이를 각 구성 네트워크에 상호 배제적으로 분담 학습시키는 방법을 구사할 수 있다. 전문화(specialization)된 네트워크들은 앙상블 알고리즘의 학습 성능을 보다 많이 개선시킬 수 있다. 본 연구에서는 clustering을 통하여 서브 학습 데이터 셋을 구성하고 각 cluster의 중심과 주어진 입력 벡터와의 거리를 네트워크 결합의 가중치 계산에 반영하는 방법을 제안하였다. 제안된 방법은 기존 관찰학습의 가상데이터 생성과정과 recall 과정에 구현되었다.

[INITIALIZE] ***Bootstrap D into L replicates*** $D_1, \ldots, D_L$

[TRAIN]

Do For $t = 1, \ldots, G$

[T-step] ***Train each network :***

Train j^{th} network f_j^t with D_j^t for each $j \in \{1, \ldots, L\}$

[O-step] ***Generate virtual data set*** V_j ***for network*** j **:**

$$V_i^t = \{(x', y') \mid x' = x + \varepsilon,\ \varepsilon \sim N(0, \Sigma),\ x \in D_j,\ y' = \sum_{j=1}^{L} \beta_j f_j^t(x'),\ where\ \beta_j = 1/L\}$$

Merge virtual data with original data :

$$D_j^{t+1} = D_j \cup V_j^t$$

End

[FINAL OUTPUT] ***Combine networks with weighting factor*** β **:**

$$f_{com}(x) = \sum_{j=1}^{L} \beta_j f_j^T(x),\quad where\ \beta_j = 1/L$$

<그림1> OLA(Observational Learning Algorithm)

본 논문의 2절에서는 제안된 알고리즘이 보다 자세히 소개되었다. 3절에서는 인공함수 추정 및 실제 문제에 대하여 제안된 알고리즘 및 bagging, boosting과의 비교 실험이 기술되었고, 마지막 절에서는 향후 연구방향에 대하여 제시하였다.

2. Specialized OLA

본 연구의 주된 아이디어는 앙상블 네트워크의 각 구성 멤버를 전문화(specialization)시키고 이를 네트워크의 결과 결합에 이용하는 것으로서 다음의 두 단계로 이루어진다.

2.1 Clustering을 이용한 학습 데이터 셋 분할

원 학습 데이터 셋 D는 K-means clustering 또는 Self Organizing Feature Map(SOFM)을 이용하여 K개의 cluster 즉, K개의 서브 데이터 셋으로 분할된다. 각 서브 데이터 셋마다 하나의 구성 네트워크가 할당되므로 K개의 멤버로 이루어진 앙상블 네트워크가 만들어진다(그림2의 [INITIALIZE] 참조). 이러한 학습 데이터 셋의 분할은 원 데이터가 갖고 있는 분포의 성질을 반영하는 데 도움이 된다. 또한 각 구성 네트워크에 대한 이들의 배제적(exclusive) 할당은 주어진 문제를 divide-and-conquer한다라는 점에서 각 네트워크에 부여된 학습의 부담을 경감시킬 수 있다. 학습 데이터 셋 분할은 또한 앙상블 알고리즘의 구성 멤버 수(population size) 결정시 적절한 근거를 제시한다. 앙상블의 구성 멤버 수를 cluster의 수와 일치하게 설정함으로써 구성 멤버 수 결정에 관한 어려움을 효율적으로 피할 수 있다.

[INITIALIZE]

1. Cluster D into K clusters, with K-means algorithm or SOFM,
$D_1, D_2, \ldots, D_K$ ***with centers located at*** $C_1, C_2, \ldots, C_K$ ***respectively.***

2. Set the ensemble size L equal to the number of clusters K.

[TRAIN]

Do For $t = 1, \ldots, G$

[T-step] ***Train each network :***

Train j^{th} network f_j^t with D_j^t for each $j \in \{1, \ldots, L\}$

[O-step] ***Generate virtual data set for each network*** j **:**

$$V_j^t = \{(x', y') \mid x' = x + \varepsilon,\ \ \varepsilon \sim N(0, \Sigma),\ \ x \in D_j,$$
$$y' = \sum_{j=1}^{L} \beta_j f_j^t(x'),\ \ where\ \beta_j = 1/d_j(x')$$
$$\text{and } d_j(x') = \sqrt{(x' - C_j)^T \Sigma_j^{-1} (x' - C_j)}\}$$

Merge virtual data with original data :

$$D_j^{t+1} = D_j \cup V_j^t$$

End

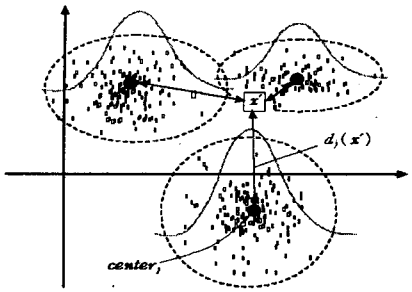
[FINAL OUTPUT] ***Combine networks with weighting factor*** β **:**

$$f_{com}(x) = \sum_{j=1}^{L} \beta_j f_j^T(x),\ \ where\ \beta_j = 1/d_j(x')$$

<그림2> Specialized OLA : OLAβ

2.2 "친근도(familiarity)"에 따른 네트워크의 가중치 결합

각 구성 네트워크들의 결과를 어떻게 결합하는가 하는 문제는 앙상블 알고리즘에서 거론되고 있는 또 하나의 커다란 이슈다. 구성 네트워크의 "전문화"는 이를 해결할 수 있는 방법을 제공한다. 즉, 각 네트워크가 주어진 입력 벡터에 얼마나 친근한가(familiar)의 정도를 이용하여 각 네트워크 결과의 신뢰도(confidency)를 가늠할 수 있다. 그림3 에서 보듯이 입력벡터 x'에 대해서는 우측 상단의 네트워크 결과값이 가장 신뢰할 만하다. 신뢰도는 네트워크 결과 결합시 가중치(weighting factor)로 활용된다. 각 네트워크의 신뢰도는 서브 데이터 셋의 중심 즉, 각 cluster 중심과 제시된 입력벡터 x'의 거리에 반비례하게끔 설정된다. 학습 데이터 셋에 대한 확률 밀도 함수(PDF)의 추정은, 대개의 경우 이에 대한 통계적 사전 정보가 주어지지 않으므로, 본 연구에서는 mixture gaussian kernel을 이용하였다[6][7][8].



<그림3> 입력벡터에 대한 친근도와 네트워크 결과의 신뢰도

따라서 입력 x'에 대한 j^{th} kernel 함수의 친근도는 다음과 같이 정의된다 ($j = 1, \ldots, L$) .

$$\theta_j(x') = \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(x' - C_j)^T \Sigma_j^{-1} (x' - C_j)\} \qquad 식(1)$$

식(1)에 자연로그를 취하면 다음을 유도할 수 있다.

$$\log \theta_j(x') = -\frac{1}{2} \log|\Sigma_j| - \frac{1}{2}(x' - C_j)^T \Sigma_j^{-1} (x' - C_j) \qquad 식(2)$$

여기서, $\Sigma_j = \Sigma_{j'}$ (단, $j \neq j'$)로 가정하면 친근도 $d_j(x')$를 측정하는 다음의 관계를 얻을 수 있다(단, $d_j(x') = \sqrt{(x' - C_j)^T \Sigma_j^{-1} (x' - C_j)}$).

$$\log \theta_j(x') \propto d_j^2(x') \qquad 식(3)$$

위의 관계로부터 주어진 입력벡터와 각 네트워크와의 친근도는 이 입력벡터와 각 cluster 중심과의 mahalanobis 거리에 반비례함을 알 수 있다. 입력벡터 x'에 가까운 벡터들로 학습한 네트워크는 결과 결합시 더 많은 가중치를 부여받게 된다. 가중치 β_j는 $1/d_j(x')$으로 정의되며 이는 그림2의 [O-STEP]과 [FINAL OUTPUT]에서 이용된다. 기존의 관찰학습에서는 β_j를 단순평균 즉, $1/L$로 설정하였었다(그림1 참조).

3. 실험방법 및 결과

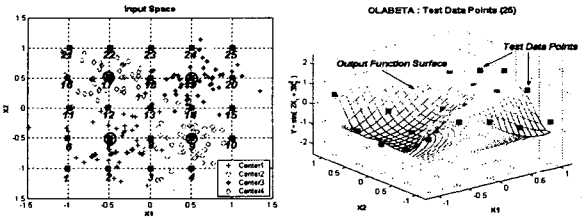
3.1 인공함수 추정문제에 대한 실험

제안된 알고리즘(OLAβ)은 우선 $y = \sin(2x_1 + 3x_2^2) + \varepsilon$, $\varepsilon \sim N(0, 0.05^2 I)$로 정의된 인공함수 추정문제에 적용되었다. 각각의 입력벡터 x는 4개의 gaussian 분포 $N(C_i, 0.3^2 I)$ 중 하나로부터 생성되었다(단, $\{(x_1, x_2) \mid (-0.5, -0.5), (0.5, -0.5), (-0.5, 0.5), (0.5, 0.5)\}$). 각 cluster마다 80개의 데이터를 생성하고 구성 네트워크의 수는 4로 설정하였으므로 총 320개의 데이터가 실험에 사용되었다(그림4 참조). 4개의 2-5-1 MLP 네트워크들은 Levenberg-Maquardt 알고리즘으로 5 epoch 동안 훈련되었다. Generation의 수는 4회로 설정되었으며 각 generation마다 80개의 가상데이터가 생성되어 80개의 원래 학습데이터와 합쳐졌다. 단, 매 generation마다 이전 generation에서 생성되었던 가상데이터는 제거되었으므로 각 네트워크는 160개의 학습데이터 셋으로 훈련되었다. 이 실험에서는 학습 데이터를 인위적인 cluster 중심으로부터 발생시켰으므로 별도의 데이터 셋 분할 작업은 수행되지 않았다.

성능비교의 목적으로 기존의 OLA, 단순평균 bagging, adaboost.R2 알고리즘들을 실험에 추가하였다[4]. Bagging은 구성 네트워크 수가 4, 15, 25인 세 종류를 사용하였다. Boosting은 매 반복마다 생성되는 구성 네트워크의 수가 변화하므로 이를 평균하여 나타내면 총 50회 반복동안 평균 36개의 네트워크를 사용하였다고 볼 수 있다.

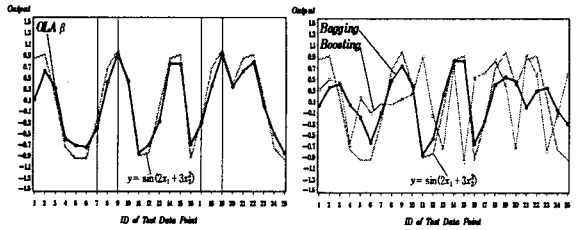
테스트 셋으로는 2가지 종류의 데이터 셋이 사용되었는데 하나는

제안한 알고리즘의 작동을 검토해 보기 위한 목적으로, 다른 하나는 정확한 평가를 위한 목적으로 고안되었다. 첫 번째 테스트 데이터 셋은 자신의 고유한 ID를 갖고 있는 25개의 데이터들로 이루어졌다(그림 4의 [LEFT] 참조). 그림4의 [RIGHT]에 표시된 mesh는 인공함수의 표면이고 사각점들은 25개 테스트 데이터들이 입력되었을 때 제안된 알고리즘에 의해 추정된 값들이다. 그림5는 25개 테스트 데이터의 입



<그림4> [LEFT] 4개의 gaussian으로부터 생성된 인공데이터와 25개의 테스트 데이터, [RIGHT] 25개 테스트 입력에 대한 추정 결과.

력과 추정된 값들을 일련의 테스트 번호대로 정렬하고 제안된 알고리즘과 bagging, boosting의 결과를 나타낸 것이다. 제안된 알고리즘은 각 cluster의 중심에 해당하는 테스트 데이터의 입력에 대해서 (7,9,17,19) 상대적으로 높은 정확도를 보여준다는 점이 특징적이다. 이는 해당 cluster를 학습한 네트워크의 결과가 네트워크 결합시 가장 비중있게 반영되었기 때문이다.



<그림5> 제안된 알고리즘은 각 cluster의 중심 부근(7,9,10,13)에서 다른 알고리즘에 비해 상대적으로 상당히 정확한 추정치를 산출한다.

다른 하나의 테스트 데이터 셋은 400개의 데이터들로 구성되었다. 표1은 50회의 반복실험에 대한 결과를 MSE에 대한 평균과 편차로 요약한 것이다. MSE의 평균만을 고려해 보면 기존의 관찰학습(OLA-25)이 가장 좋은 성능을 보이나 구성 네트워크의 수를 함께 고려한다면 제안된 알고리즘이 훨씬 적은 수(4 networks)의 네트워크로 이와 유사한 정확도를 보임을 알 수 있다. 앙상블 학습은 여러 개의 네트워크를 훈련시켜야 하는 부담을 안고 있고 이는 구성 네트워크의 수와 밀접한 관계를 가지므로, 제안된 알고리즘이 초기에 clustering을 수행해야 함에도 불구하고 효율적인 방법임을 알 수 있다. Bagging은 제안된 알고리즘 네트워크 수의 약 4배 가량의 구성 네트워크를 가진 후에야 유사한 성능을 보였다. Boosting은 비교된 다른 알고리즘들에 비해 성능이 저조하였다. 표1의 마지막 행은 제안된 알고리즘과 다른 알고리즘과의 MSE에 대한 단측 T-test 결과를 p-value로 기록한 것이다. 유의수준을 5%로 한다면 제안된 알고리즘은 OLA-4 및 bagging-4, boosting 보다 우수함이 검증된 것이다.

<표1 : 인공함수 추정에 대한 실험결과>

50 runs	OLAβ	OLA			Bagging			Boosting
Ensemble Size	4	4	15	25	4	15	25	Avg(36)
Avg MSE (10^{-3})	5.4	6.5	4.8	4.7	7.3	5.4	5.4	10.5
Std MSE (10^{-3})	4.0	2.0	0.9	0.8	3.0	2.7	1.3	9.2
P-value (T-test)	-	0.04	0.87	0.90	0.01	0.99	1.00	0.00

3.2 실제 문제에 대한 실험

제안된 알고리즘은 실제 예측문제에도 적용되었다: Boston Housing [9], Ozone[10]. 데이터 셋은 K-means clustering을 이용하여 각각 10개, 9개의 서브 데이터 셋으로 분할되었다. 10개의 13-10-1 MLP와 9개의 8-10-1 MLP는 모두 LM 알고리즘으로 훈련되었다. 이에 대한 실험결과가 표2에 제시되어 있다. Boston Housing 문제에서는 제안된 알고리즘이 bagging 및 boosting 보다 월등히 좋은 예측력을 보여 주었으나 Ozone 문제에서는 bagging이 가장 우수한 성능을 나타냈다.

<표2 : 실제 문제에 대한 실험결과>

Tr/Val/Test 30 runs	Boston Housing 200/106/100			Ozone 200/30/100		
	OLAβ	Bagging	Boosting	OLAβ	Bagging	Boosting
Ensemble Size	10	25	Avg(48)	9	25	Avg(49)
Avg MSE (10^{-2})	9.3	10.3	10.9	19.2	19.0	21.0
Std MSE (10^{-2})	0.89	0.96	1.39	0.94	0.65	0.82
P-value (T-test)	-	0.00	0.00	-	0.79	0.00

4. 결론

본 연구에서는 K-means clustering을 이용하여 초기 학습 데이터 셋을 여러 개의 서브 학습 데이터 셋으로 분할한 후, 이를 각 구성 네트워크에 상호 배제적으로 분담 학습시킴으로써 전문화된 네트워크들의 결합에 의한 앙상블 학습방법을 제안하였다. 제안된 알고리즘의 네트워크 결합방법은 제시된 입력벡터와 각 서브 학습 데이터 셋간의 거리를 반영한 가중 평균을 통해 이루어진다.

제안된 알고리즘은 인공함수 추정 및 실제 문제에 대한 bagging 및 boosting과의 비교 실험에 의하여, 보다 적은 수의 구성 네트워크로 동일 내지 보다 나은 성능을 나타냄이 실험적으로 검증되었다.

그러나 본 연구의 보다 일반화된 타당성을 검증받기 위해서는, 여러 실제 문제에 대한 확장 실험이 후행되어져야 한다. 또한, 알고리즘의 세부적인 측면으로는, clustering이외의 다양한 분할방법을 통한 서브 학습 데이터 셋의 구성과 보다 정교한 계산방법에 의한 네트워크 결합 가중치 설정 등이 연구주제로 남아있다.

본 연구는 뇌 과학 및 공학 연구 프로그램과 Brain Korea21에 의해 지원되었다.

5. 참고문헌

[1] Cho, S. and Cha, K., "Evolution of neural network training set through addition of virtual samples," *International Conference on Evolutionary Computations*, 685-688 (1996)

[2] Cho, S., Jang, M. and Chang, S., "Virtual Sample Generation using a Population of Networks," *Neural Processing Letters*, Vol. 5 No. 2, 83-89 (1997)

[3] Jang, M. and Cho, S., "Observational Learning Algorithm for an Ensemble of Neural Networks," submitted (1999)

[4] Drucker, H., "Improving Regressors using Boosting Techniques," *Machine Learning: Proceedings of the Fourteenth International Conference*, 107-115 (1997)

[5] Perrone, M. P. and Cooper, L. N., "When networks disagree: Ensemble methods for hybrid neural networks," *Artificial Neural Networks for Speech and Vision*, (1993)

[6] Platt, J., "A Resource-Allocating Network for Function Interpolation," *Neural Computation*, Vol 3, 213--225 (1991)

[7] Roberts, S. and Tarassenko, L., "A Probabilistic Resource Allocating Network for Novelty Detection," *Neural Computation*, Vol 6, 270--284 (1994)

[8] Sebestyen, G. S., "Pattern Recognition by an Adaptive Process of Sample Set Construction," *IRE Trans. Info. Theory IT-8*, 82--91 (1962)

[9] http://www.ics.uci.edu/~mlearn

[10] http://www.stat.berkeley.edu/users/breiman