앙상블 학습알고리즘의 일반화 성능 비교
: OLA, Bagging, Boosting

신현정[U] 장민 조성준 이봉기 임용업
서울대학교 산업공학과
{hjshin72, zoon}@snu.ac.kr

Generalization Abilities of Ensemble Learning Algorithms
: OLA, Bagging, Boosting

Hyun-Jung Shin[U] Min Jang Sung-Zoon Cho Bong-Ki Lee Yong-Up Lim
Dept. of Industrial Engineering, Seoul National University

요 약

최근 제안된 관찰학습(OLA: Observational Learning Algorithm)은 committee를 구성하는 각각의 학습 모델들이 다른 학습 모델들을 관찰함으로써 얻어진 가상데이터를 실제 데이터와 결합시켜 학습에 이용하는 방법이다. 본 논문에서는, UCI 데이터 셋의 분류(classification)와 예측(regression)문제에 대하여 다층 퍼셉트론을 학습 모델로 설정하고, 이에 대하여 OLA와 bagging, boosting의 성능을 비교, 분석하였다.

1. 서 론

신경망 모델은 충분한 수의 은닉노드와 학습시간, 특히 충분한 양의 학습 데이터가 주어질 경우, 어떠한 연속함수라도 원하는 정확도 내로 근사(fitting)시킬 수 있는 보편 근사기(general approximator)로 알려져 있다. 즉, 학습에 사용될 샘플의 크기가 크다면, 네트워크 복잡도(은닉노드 수)의 증가는 신경망 모델의 일반화 성능을 향상시킨다. 그러나 현실적으로는, 양질의 충분한 학습 데이터를 얻는다는 것이 비용측면에서 비현실적이거나 아예 불가능할 수도 있다. 충분한 학습 데이터가 주어지지 않은 상황에서 신경망의 복잡도(은닉노드 수) 증가는 과도학습(overfitting)을 유발하게 된다.

이러한 과도학습을 완화시키는 regularization의 한 방법으로 committee 네트워크에 대한 연구가 활발히 진행되어 오고 있다. committee란 동일한 학습 데이터 셋을 여러 네트워크를 이용하여 독립적으로 학습시킨 후, 이들을 결합(ensemble)하여 최종결과를 얻는 방법이다[2]. committee를 구성하는 네트워크들의 알고리즘과 조합방법에는 제한이 없으나, 예를 들어, 다층 퍼셉트론(multi-layer perceptron)으로만 committee를 구성하는 경우, 각 네트워크들은 초기 연결 가중치(initial weight)나 네트워크 구조를 서로 다르게 하여 구성하는 방법이 보편적으로 사용되고 있으며, 이는 분류(classification)나 예측(regression) 문제에 있어서 유의한 성능 향상을 보인다[6,8].

신경망의 일반화 성능 향상과 관련된 다른 측면의 방법으로는 학습 데이터 셋을 어떻게 구성하느냐라는 주제와 연관 지을 수 있다. 즉, 주어진 데이터 셋을 여러 개의 학습 데이터 셋이 되도록 샘플링하는 방법에 의해 신경망의 일반화 성능이 향상될 수 있다. 따라서, committee 네트워크의 일반화 성능 향상은 크게 초기 연결 가중치나 네트워크 구조를 달리함으로써 얻어질 수도 있고, 각각의 committee 구성원들을 학습시킬 데이터 셋을 샘플링하는 방법에 의해서도 얻어질 수 있다[7].

본 연구에서는 학습 데이터 셋을 샘플링 방법에 따라 세 개의 committee 네트워크, OLA(Observational Learning Algorithm), bagging 앙상블, boosting 앙상블을 선정하고 분류와 예측문제에 대하여 이들간의 성능을 비교, 실험하였다. OLA는 네트워크을 구성하는 각각의 학습 모델들이 다른 학습 모델들을 관찰함으로써 얻어진 가상데이터를 실제 데이터와 결합시켜 다음 학습에 이용하는 방법이다[2,3]. bagging 앙상블은 주어진 학습 데이터를 복원추출(bootstrapping : resampling randomly with replacement)하여 여러 개의 학습데이터로 만들고 각각에 대하여 학습된 모델들의 결과를 결합하는 방법이다[1]. Boosting 앙상블은 학습데이터의 패턴 중, 학습이 잘 안되는 패턴(hard example)에 가중치를 두어 단계별로 샘플링한 후, 다음 단계에서 이들을 반영시키는 방법으로서, 최종적으로 각 단계별 네트워크 결과를 결합하는 방법이다[4,5]. 원래의 데이터 셋으로부터 각 committee의 학습데이터 셋을 어떻게 구성하느냐의 관점에서 이들을 간단히 비교하면 다음과 같다. OLA는 원래 데이터 셋에 가상데이터 셋을 추가함으로써 학습 데이터에 포함될 원 데이터의 패턴을 확장시키는 방법이고, bagging은 샘플링에서 보편적으로 쓰이는 방법으로, 원래의 데이터 셋을 어떻게 나누느냐의 방법이다. boosting은 원래 데이터 셋으로부터 학습데이터 셋을 구성하는 과정에서, 학습 데이터에 포함될 원 데이터의 패턴이 일부 특정 패턴으로 축소, 강조되는 방법이다.

본 논문의 2절에서는 OLA, bagging 앙상블, boosting 앙상블에 대하여 각각의 알고리즘을 보다 자세히 기술하였다. 3절에서는 본 연구의 실험방법 및 적용문제에 대하여 기술하였고, 4절에서는 이에 대한 실험결과를 요약하였다. 마지막 절에서는 향후 연구방향에 대하여 제시하였다.

2.관련연구

2.1 OLA(Observational Learning Algorithm:관찰학습)

OLA는 원래의 학습데이터로부터 인공적으로 가상데이터를 생성하여 학습에 참여시키는 방법이다. 이 때, 각 committee 네트워크들의 앙상블은 단계별로는 가상데이터의 출력패턴을 생성하는 데에도 이용되며, 최종적으로는 결과산출에도 이용된다. 세부 내용은 다음과 같다.

크기가 N인 학습 데이터 셋 $D=\{(x_k^*, y_k^*)|k=1,\ldots,N\}$와 L개의 네트워크 f_i, $(i=1,\ldots,L)$가 주어졌다고 하자. L개의 각 네트워크들에 대한 초기 학습 데이터 셋은 D로부터 bootstrapping을 이용하여 N개씩 L개가 추출된다. 각 네트워크들은 매 epoch 마다 학습단계(T-step)와 관찰단계(O-step)를 수행하게 된다. 현재의 epoch을 t라하면, 학습단계는 f_i^t들이 D_i^t를 학습하는 과정으로 각 네트워크의 연결가중치들이 수정된다. 관찰단계는 학습된 네트워크들의 앙상블 f_{com}^t을 이용하여 크기 N인 L개의 가상 데이터 셋 V_i^t를 생성하는 단계이다. 이 때, 가상 데이터 셋 V_i^t의 입력패턴은 원래의 데이터 셋 D_i의 입력패턴에 정규분포를 따르는 잡음(noise)을 추가하여 생성된 것이며, 이에 대한 V_i^t의 출력패턴은, i번째 네트워크 f_i^t를 제외시킨 $L-1$개 네트워크들의 앙상블(f_{com}^t)로 채워진다. 따라서, $t+1$에서의 학습 데이터 셋은 $D_i^{t+1} = D_i + V_i^t$로 $2N$개가 된다. 단, $t=0$에서는 가상 데이터가 없기 때문에 원래의 데이터 셋 D_i로만 학습이 수행된다. 이러한 과정을 G epoch 만큼 반복한 후, L개의 네트워크(f_i^G)들을 가중치 α_i를 이용해 결합함으로써 최종결과를 얻는다[3]. <그림1>은 OLA를 정리한 것이다.

Initialize

Let $\{f_i \mid i=1,\ldots,L\}$ *be a set of networks in the ensemble.*
Let $\{D_i \mid i=1,\ldots,L\}$ *be bootstrapping replicates of original data set D.*
Let $\{V_i \mid i=1,\ldots,L\}$ *be virtual data sets to be generated.*

Do for $t=1,\ldots G$

[T-step] *train each network*

$$f_i^t(\vec{x}, \vec{y}),\quad i=1,\ldots,L,\quad (\vec{x}, \vec{y}) \in D_i^t$$

[O-step] *for each network, virtual data sets are created*

$$V_i^{t+1} = \{(\vec{x}', \vec{y}') \mid \vec{x}' = \vec{x} + \vec{\varepsilon},\ \vec{\varepsilon} \sim N(0, \Sigma),\ \vec{y}' = f_{com}^t(\vec{x}'),\ f_{com}^t(\vec{x}') = \sum_{j=1, j\neq i}^{L} \beta_j f_j^t(\vec{x}'),\ \vec{x} \in D_i\}$$

$$D_i^{t+1} = D_i \bigcup V_i^t$$

End

Final output *The networks are combined with weight* α_i*'s,*

$$f_{com}(\vec{x}) = \sum_{i=1}^{L} \alpha_i f_i^G(\vec{x})$$

<그림1> OLA(Observational Learning Algorithm)

2.2 Bagging

Bagging(bootstrap aggregating)은 Breiman에 의하여 개발된 방법으로, 학습 알고리즘을 여러 개의 복사본으로 만든 후, 이들을 각각 학습시키고 그 결과를 결합하는 방법이다[1]. 즉, bagging은 bootstrapping에 의하여 만들어진 L개의 데이터 셋에 대하여 L개의 복사된 학습 알고리즘들이 각각을 학습한 후, 그 결과들이 majority voting이나 simple averaging과 같은 방법에 의하여 결합된다. bagging은 데이터 내의 작은 변화가 학습 알고리즘의 결과에 미치는 영향이 큰 경우, 즉 불안정(unstable)한 학습 알고리즘을 사용할 경우에 효과적인 성능을 나타내는 방법이다. <그림2>는 bagging 알고리즘을 정리한 것이다.

Initialize

Let $\{f_i \mid i=1,\ldots,L\}$ *be a set of networks in the ensemble.*
Let $\{D_i \mid i=1,\ldots,L\}$ *be bootstrapping replicates of original data set D.*

Do for $t=1,\ldots G$

train each network : $f_i(\vec{x_j}, \vec{y_j}),\quad i=1,\ldots,L,\quad (\vec{x_j}, \vec{y_j}) \in D_i$

End

Final output *The networks are combined with weight* α_i*'s*

$$f_{com}(\vec{x}) = \sum_{i=1}^{L} \alpha_i f_i^G(\vec{x})$$

<그림2> Bagging

2.3 Boosting

AdaBoost(Adaptive Boosting)는 학습 데이터 셋 중, 학습결과의 성능 저하(loss)를 유발하는데 기여도가 큰 특정 패턴에 가중치를 주어 학습시킨 후, committee의 결과를 단계별로 결합하는 방법이다. Boosting 알고리즘은 주로 weak learning algorithm들을 결합시키는데 많이 사용되나, AdaBoost는 decision tree나 신경망에 대한 앙상블에 적용되고 있다. 또한, AdaBoost는 최근 다양한 문제별로 여러 version이 개발되어 오고 있다.

본 연구에서는 예측(regression)문제에 AdaBoost.R2를, 분류(clas-sification)문제에 대해서는 AdaBoost.M2를 사용하였다[4,5]. <그림3>은 AdaBoost.R2의 알고리즘을 정리한 것이다.

Initialize

Let f *be a WeakLearner (=a network)*
Let $D=\{(x_1, y_1),\ldots,(x_n, y_n)\}$ *be original data set with* $y \in Y=[0,1]$.
Let $dist^1(j) = \frac{1}{n}$, $(j=1,\ldots,n)$ *be distribution of each pattern in D.*

Do while $\bar{L} < \frac{1}{2}$

D^t *is constructed with* $dist_t(j)$ *from D.*
train the network : $f^t(\vec{x}, \vec{y}),\quad (\vec{x}, \vec{y}) \in D^t$
calculate a loss : $L_j = \frac{|f^t(x_j, y) - y_j|}{L_{\max}}$
$(L_{\max} = \sup|f^t(x_j, y) - y_j|$, over all $j)$
calculate weighted loss : $\bar{L} = \sum_{j=1}^{n} L_j\, dist^t(j)$
set $\beta_t = \frac{\bar{L}}{(1-\bar{L})}$
update distribution : $dist^{t+1}(j) = \frac{dist^t(j)\,\beta_t^{(1-L_j)}}{Z_t}$,
Z_t *is a normalization constant.*
$t = t+1$

end

Final output $f_{com} = \inf[\, y \in Y : \sum_{(t \mid h_t \geq y)} \log(\frac{1}{\beta_t}) \geq \frac{1}{2} \log(\frac{1}{\beta_t})\,]$

<그림3> Boosting(AdaBoost : Adaptive Boosting)

3. 실험방법

본 연구에서는 UCI repository의 4개의 실제 데이터에 대하여 OLA, bagging 앙상블, boosting 앙상블의 성능을 비교하였다. <표1>은 실험에 사용된 데이터 셋을 요약한 것이다.

HHD(Hungarian Heart Disease)와 Ionosphere는 분류(classification) 문제에 대하여, Ozone과 BH(Boston Housing)는 예측(regression) 문제에 대하여 적용되었다. 또한, 학습 데이터 셋의 크기가 신경망의 일반화 성능에 어떻게 영향을 미치는가를 실험하기 위하여, 각 데이터 셋별로 학습 데이터 셋의 크기를 몇 가지 수준으로 나누어 실험하였다.

<표 1> Data Set Summary

	# of Total Data	# of Training Data (N)	# of Test Data	# of Variables
HHD	196	50,100,146	50 (fixed)	10
Ionosphere	351	100, 200	52 (fixed)	34
Ozone	330	50,100,150	150 (fixed)	8
BH	506	100,200	406,306	12

신경망의 은닉층은 하나로 설정하였고 은닉노드의 수는 학습데이터 셋으로부터 임의추출된 검증(validation) 데이터 셋에 의하여 실험적으로 결정되었다. 신경망을 학습시키는 알고리즘으로는 Levenberg-Marquardt 방법이 사용되었다. 세 가지의 앙상블 알고리즘들은 각각, 초기 연결가중치가 다른 25개의 committee들로 동일하게 구성되었으며, 4개의 데이터 셋에 대한 10개의 실험에서 매회 일정 epoch 수만큼 총 25회씩 반복 실험하였다

4. 실험결과

<표2>와 <표3>은 분류와 예측문제에 대한 세 가지 앙상블 알고리즘들의 실험결과를 각각 요약한 것이다. 25회의 실험결과에 대하여, 분류문제에 대해서는 오분류율(classification error rate)에 대한 평균(μ)과 분산(σ^2)을, 예측문제에 대해서는 MSE(mean squared error)에 대한 평균과 분산을 측정하였다.

<표 2> Classification Error Rate

		Bagging		Boosting		OLA	
Data	**N**	μ	σ^2	μ	σ^2	μ	σ^2
HHD	50	28.48	31.43	35.56	78.43	24.96	29.04
	100	21.28	14.63	26.22	46.02	19.04	10.04
	146	20.40	9.00	35.71	40.11	18.80	5.66
Ionosp-here	100	13.85	5.57	25.60	21.69	13.54	9.98
	200	9.77	5.80	13.46	9.85	10.31	13.83

<표 3> Mean Square Error

		Bagging		Boosting		OLA	
Data	**N**	μ	σ^2	μ	σ^2	μ	σ^2
Ozone	50	24.38	18.48	31.32	83.97	24.74	11.04
	100	19.69	2.16	27.33	31.10	19.58	2.80
	150	18.46	1.18	22.79	9.04	19.12	0.87
BH	100	22.58	9.72	28.49	68.92	18.97	9.12
	200	16.68	4.28	17.37	5.16	15.59	0.77

실험결과를 이용하여 각 데이터 셋별로 세 가지 앙상블 알고리즘의 성능을 비교하면 다음과 같다.

- **HHD :** ***OLA > Bagging > Boosting***
  OLA에 의한 신경망의 일반화 성능이 bagging에 비해 약 11.5% 정도 개선되었으며, 각 알고리즘별 실험결과에 대한 평균의 차이는 유의함이 검증되었다(유의수준 1%).
- **Ionosphere :** ***OLA > Boosting , Bagging > Boosting***
  Boosting의 성능이 OLA나 bagging에 비해 좋지 않았으며, OLA와 bagging,두 알고리즘 간의 성능에 대한 평균차이는 유의하지 못했다(유의수준 1%).
- **Ozone :** ***OLA > Boosting , Bagging > Boosting***
  OLA와 bagging이 성능에 있어서 차이가 있다고는 볼 수 없었으며, boosting의 성능은 나머지 두 알고리즘에 비해 좋지 않았다(유의수준 1%).
- **BH :** ***OLA > Boosting , OLA > Bagging***
  OLA가 나머지 두 알고리즘에 비해 성능면에서 약 19.4% 정도로 우수하게 검증되었으며, bagging과 boosting의 성능차이는 유의하지 않았다(유의수준 1%).

또한, OLA의 성능이 뛰어났던 HHD에 대하여, 학습 데이터 셋의 크기 수준에 대한 성능을 Bagging과 비교해 본 결과, 50, 100, 146에 대하여 각각 14.1%, 11.8%, 8.5%로, 학습 데이터의 수가 적을수록 OLA의 성능이 향상됨을 보였다.

5. 결론

본 연구에서는 학습 데이터 셋을 구성하는 샘플링 방법에 따라 세 개의 앙상블 네트워크, OLA, bagging 앙상블, boosting 앙상블을 선정하고, 4개의 UCI repository 데이터 셋에 대하여, 이들간의 성능을 비교, 실험하였다. 실험결과, 가상샘플을 생성하여 학습에 추가하는 OLA의 성능이 bagging과 boosting에 비하여 우수하였다. 또한, OLA의 특성상 학습 데이터가 적을수록 다른 알고리즘에 대한 성능 개선비가 커짐을 관찰할 수 있었다.

본 연구에서 실험한 세 가지 앙상블 네트워크는 샘플링 방법에 따른 신경망의 일반화 성능을 비교한 것이므로, 적용되는 데이터 셋의 특성-데이터의 잡음(noise)정도, 변수의 개수 및 종류 등-에 따라 그 성능의 우열이 달라질 수 있다. 따라서, 데이터 셋의 특성을 규정하는 기준을 마련하고 이에 따른 비교실험이 추가되어져야 할 것이다.

6. 참고문헌

[1] Breiman, L., "Bagging Predictors". *Technical Report 421*, Department of Statistics, University of California, 1994

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

[3] Cho, S., Jang, M. and Chang, S., "Virtual Sample Generation using a Population of Networks", *Neural Processing Letters*, 83-89, Netherland, 1997.

[4] Drucker, E., "Boosting Using Neural Networks", In Amanda J. C. Sharkey (Eds), *Combining Artificial Neural Nets :Ensemble and Modular Learning*, 51-77 , Springer-Verlag, 1999

[5] Freund,Y., Schapire,R.E.,"Experiments with a New Boosting Algorithm", *Machine Learning : Proceedings of the Thirteenth International Conference*, 1996

[6] Parmanto,B., Munro, P.W., *et al.*, "Neural network classifier for hepatoma detection", *Proceeding of the World Congress of Neural Networks*, San Diego, 1994

[7] Parmanto,B., Munro, P.W. & Doyle, H. R.,"Improving committee diagnosis with resampling techniques", In D. Touretsky, M. Mozer & M. Hasselmo (Eds), *Advances in Neural Information Processing Systems-8*, Cambridge, MA: MIT Press.

[8] Perrone, M.P., "Improving regression estimation : averaging methods for variance reduction with extension to general convex measure optimization", PhD Thesis, Department of Physics, Brown University, Providence, RI, 1993.