The Korean Journal of Applied Statistics
2021, Vol. 34, No. 2, 167–175

DOI: https://doi.org/10.5351/KJAS.2021.34.2.167
Print ISSN1225-066X / Online ISSN 2383-5818

A study on the performance improvement of learning based on consistency regularization and unlabeled data augmentation

Hyunwoong Kim[a], Kyungha Seok[1,b]

[a]Clinical Trial Center, Haeundae Paik Hospital Inje University; [b]Department of Statistics, Inje University

Abstract

Semi-supervised learning uses both labeled data and unlabeled data. Recently consistency regularization is very popular in semi-supervised learning. Unsupervised data augmentation (UDA) that uses unlabeled data augmentation is also based on the consistency regularization. The Kullback-Leibler divergence is used for the loss of unlabeled data and cross-entropy for the loss of labeled data through UDA learning. UDA uses techniques such as training signal annealing (TSA) and confidence-based masking to promote performance. In this study, we propose to use Jensen-Shannon divergence instead of Kullback-Leibler divergence, reverse-TSA and not to use confidence-based masking for performance improvement. Through experiment, we show that the proposed technique yields better performance than those of UDA.

Keywords: cosistency regularization, Jensen-Shannon divergence, semi-supervised learning, training signal annealing, unsupervised data augmentation

1. 서론

최근 딥러닝(deep learning)에 관한 연구가 폭넓게 진행되고 있는데 목표값(label)이 있는 데이터(labeled data; LD)를 이용하는 지도학습(supervised learning)의 연구가 많은 부분을 차지한다. 딥러닝을 이용한 지도학습에서는 LD의 크기가 아주 중요한 역할을 한다. 그렇지만 목표값을 얻는 것은 많은 시간이 소요될 뿐 아니라 전문적 인 지식이 필요할 때도 있어 대용량 LD를 구하는 것은 쉽지 않다. 이와 같은 현실적인 한계를 극복하기 위해 개발된 방법이 준지도학습(semi-supervised learning)이다. 준지도학습은 LD를 이용하는 지도학습과 목표값이 없는 데이터(unlabeled data; ULD)를 이용하는 비지도학습(unsupervised learning)을 같이 학습하는 방법이다 (Chapelle 등, 2006). 준지도학습에는 다양한 접근방법이 있는데 최근에는 일치성규칙(consistency regularization)을 기반으로 하는 방법이 많은 관심을 받고 있다. 일치성규칙은 입력값 x와 여기에 잡음을 더한 x^*와 유사해지도록 모형을 학습하는 방법이다 (Oliver 등, 2018). 처음으로 제안된 일치성규칙 기반 준지도학습은 Laine과 Aila (2016)이 제안한 *Π*–모형과 시간적조화(temporal ensembling; TE)가 있다. *Π*–모형은 ULD에 각각 다른 잡음을 추가한 데이터를 생성하고 그에 대한 모델의 출력값이 유사해지도록 학습하는 방법이고

This work was supported by the 2019 Inje University research grant.
[1] Corresponding author: Department of Statistics/Institute of Statistical Information, Inje University, 197 Inje-ro, Kimhae 50834, Korea. E-mail: statskh@inje.ac.kr

Published 30 April 2021 / journal homepage: http://kjas.or.kr

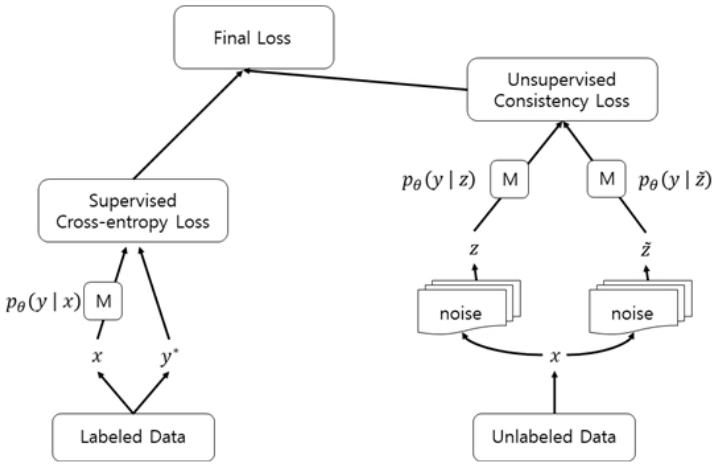

Figure 1: *Structure of Π–model.*

TE는 *Π*–모형에서 이전 결과값을 이용하여 학습하는 방법으로 큰 데이터를 학습할 때 다루기 어렵다는 단점이 있는데 이를 보완하여 평균교습방법(mean teacher) (Tavainen과 Valpola, 2017)이 제안되었다. 그 외 다른 방법으로는 가상대립적훈련(virtual adversarial training) (Miyato 등, 2018) 등이 있다.

Xie 등 (2019)은 최근에 ULD의 데이터 증대를 이용한 준지도학습(semi-supervised learning using unlabeled data augmentation; UDA)을 제안하였는데 이는 ULD에 잡음을 추가하는 방법 대신 증대(data augmentation)를 이용하여 학습을 수행하는 방법인데 아주 좋은 결과를 보여준다. UDA는 목표값이 있는 데이터의 손실함수로는 교차엔트로피(cross-entropy)를 사용하고 없는 데이터의 손실함수로는 KL-정보량 (Kullback-Leibler divergence)을 이용한다. 그리고 UDA는 성능향상을 위해 훈련신호강화(training signal annealing; TSA)와 신뢰기반 마스킹(confidence based masking) 기법을 사용한다.

본 연구에서는 UDA의 성능을 향상시키기 위해 KL-정보량 대신 JS-정보량(Jensen-Shannon divergence)을 제안한다. 그리고 UDA에서 성능향상을 위해 이용하는 TSA를 역으로 이용하고 신뢰기반 마스킹을 제거하는 방법을 제안한다. CIFAR-10 데이터를 통해 제안된 방법이 기존의 UDA보다 더 우수한 성능을 가진다는 것을 확인하였다.

2장에서는 이전에 제안되었던 일치성규칙 기반 방법과 UDA를 소개하고 3장에서는 성능 향상에 대한 방법을 제시한다. 4장에서는 연구에 사용한 이미지 데이터와 분석방법에 관해 설명하고 연구결과를 기술한다. 5장에서는 결론 및 향후 과제를 제시한다.

2. 데이터 증대를 이용한 준지도학습

UDA를 소개하기 전에 먼저 최초의 일치성규칙 기반 준지도학습인 *Π*–모형을 소개한다.

2.1. *Π*–모형

Laine과 Aila (2016)의 *Π*–모형은 최초의 일치성규칙 기반 준지도학습 방법인데 Figure 1은 그 구조를 보여주고 있다. 그림에서 M은 p_θ를 계산하는 심층신경망(deep neural network)을 나타낸다. LD = $\{(x, y^*)\}$는 교차-엔트

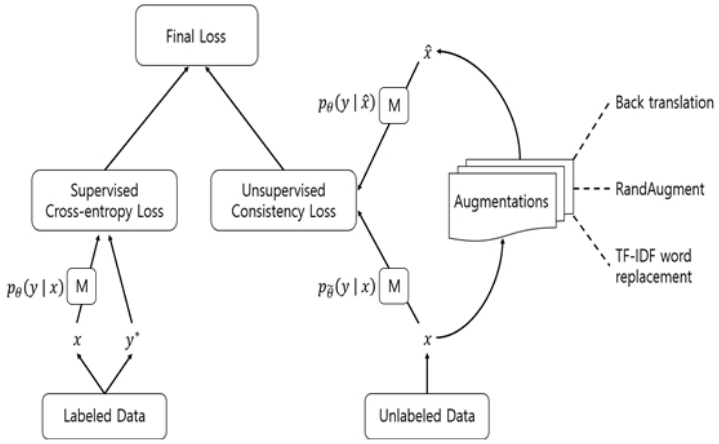

Figure 2: *Structure of UDA*.

로피 손실(cross-entropy loss)을 통해 분류기 $p_\theta(y|x)$의 학습에 사용되는데 교차-엔트로피 손실함수는 다음과 같다.

$$\mathbb{E}_{x,y^* \in L}[-\log p_\theta(y^*|x)], \tag{2.1}$$

여기서 L은 LD 집합을 의미한다. 그리고 $p_\theta(y|z)$와 $p_\theta(y|\tilde{z})$의 차이인 일치성손실(consistancy loss) $(1/n_{\text{ULD}}) \sum \|p_\theta(y|z) -_\theta (y|\tilde{z})\|^2$로 학습한다. 이 때 n_{ULD}는 ULD의 크기이다.

2.2. 데이터 증대를 이용한 준지도학습

UDA는 크기가 작은 크기의 LD와 아주 큰 ULD를 이용하는 학습 방법으로 이미지 데이터와 텍스트 데이터 등 여러 분야에 광범위하게 적용할 수 있다. Figure 2는 UDA 학습과정에 대한 그림으로 *II*–모형과 유사하다. 이미지 데이터에는 랜덤증대(RandAugment) (Cubuk 등, 2019), 텍스트 데이터에는 역번역(back translation) 혹은 단어빈도-문서역빈도(term frequency-inverse document frequency) 단어교체 방법을 적용하여 데이터를 증대한다. 랜덤증대는 자동증대(AutoAugment) (Cubuk 등, 2018)를 보완한 방법으로 회전(rotation), 색 변화(color), 밝기조절(brightness), 자르기(cropping) 등과 같은 변형 방법들 중에서 적용할 개수와 강도를 설정하여 데이터를 증대하는 방법이다.

II–모형처럼 LD는 교차엔트로피 손실함를 통해 지도학습에 이용된다. ULD는 각 사례 x에 대한 라벨 예측값 분포 $p_{\tilde{\theta}}(y|x)$와 증대를 통해 생성된 $\hat{x}$에 대한 라벨 예측값의 분포 $p_\theta(y|\hat{x})$의 차이를 계산하는데 일치성 손실함수에 사용되는데 일치성 손실함수는 아래의 식 (2.2)와 같다.

$$\mathbb{E}_{x \in U} \mathbb{E}_{\hat{x} \sim q(\hat{x}|x)}[D_{\text{KL}}(p_{\tilde{\theta}}(y|x) \| p_\theta(y|\hat{x}))], \tag{2.2}$$

여기서 U는 ULD 집합, q는 데이터 증대 그리고 $D_{\text{KL}}(p\|q)$는 KL-정보량을 나타낸다. KL-정보량은 두 확률분포의 차이를 계산하는 함수로 항상 0 이상의 값을 갖고 0일 때는 두 확률분포가 같다는 것을 의미하며 KL-정보량에 대한 수식은 다음과 같다.

$$D_{\text{KL}}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{2.3}$$

Table 1: α_t of three schedules

로그-스케줄	선형-스케줄	지수-스케줄
$\alpha_t = 1 - \exp\left(-\frac{t}{T} \times 5\right)$	$\alpha_t = \frac{t}{T}$	$\alpha_t = \exp\left(\left(\frac{t}{T} - 1\right) \times 5\right)$

LD와 ULD의 손실함수를 결합하여 아래의 최종 손실함수 (2.4)를 얻을 수 있다.

$$\min_{\theta} J(\theta) = \min_{\theta} \mathbb{E}_{x,y^* \in L}[-\log p_\theta(y^*|x)] + \lambda \mathbb{E}_{x \in U} \mathbb{E}_{\hat{x} \sim q(\hat{x}|x)}[D_{\mathrm{KL}}(p_{\tilde{\theta}}(y|x) \| p_\theta(y|\hat{x}))], \tag{2.4}$$

여기에서 λ는 각 손실의 균형을 맞추기 위해 이용되는 초모수(hyperparameter)다.

2.3. 훈련신호강화(training signal annealing)

LD는 ULD에 비해 크기가 아주 작은 경우가 대부분이다. LD만 이용하는 심층신경망은 빠르게 과적합(overfitting) 될 수 있는데 UDA는 이러한 문제점을 해결하기 위해 TSA를 제안하였다. TSA는 LD를 학습하면서 신뢰도가 임계값보다 낮은 사례만 학습을 진행하는 방법이다. 이는 부스팅(boosting) (Freund과 Schapire, 1997)과 비슷한 의미를 지니는 방법으로, 현재 학습단계 t 에서 LD의 사례 x에 대한 예측값 $p_\theta(y^*|x)$가 임계값 η_t는 $1/K$ (K는 범주의 수)에서 1로 점차 증가하는 값을 가지는데 η_t를 조절하는 방법으로 로그-스케줄(log-schedule), 선형-스케줄(linear-schedule), 지수-스케줄(exponentialschedule) 등이 있다. 각 스케줄은 모델이 과적합이 되기 쉬운지 혹은 어려운지에 따라 다르게 사용될 수 있는데 LD가 매우 적을 때는 지수-스케줄을 이용하고 충분할 때는 로그-스케줄을 이용하는 것이 적합한데 (Xie 등, 2019) 임계값 η_t는 아래의 식 (2.5)와 같이 계산되며 α_t는 각 스케줄에 따라 Table 1과 같이 계산된다.

$$\eta_t = \alpha_t \left(1 - \frac{1}{K}\right) + \frac{1}{K}. \tag{2.5}$$

2.4. Sharpening predictions

UDA는 성능을 저해하는 문제를 해결하기 위해 엔트로피 최소화(entropy minimization) (Grandvalet와 Bengio, 2004)를 이용한다. 엔트로피 최소화는 예측값의 신뢰도를 높이기 위한 하나의 방법으로 준지도학습 방법에 효과적인 것으로 나타났다 (Miyato 등, 2018). 엔트로피가 커질수록 불확실성이 증가하여 확률분포는 평평해지고 작아질수록 확률분포가 집중되는 경향이 있다. UDA에서는 엔트로피 최소화를 위해 차별화 함수(softmax temperature)를 이용하는데 이는 예측값의 최대값과 다른 값들과의 차이를 증폭시키는 역할을 한다. 차별화 함수는 아래의 식 (2.6)과 같은데 τ가 커질수록 예측값들은 더 균등해져서 엔트로피는 더 커진다.

$$\mathrm{st}(p_i, \tau) := \frac{\exp(p_i/\tau)}{\sum_{j=1}^{K} \exp(p_j/\tau)}, \quad i = 1, \ldots, K, \tag{2.6}$$

여기에서 $p_\theta(y|x) = (p_1, \ldots, p_K)'$, K는 범주의 수이다.

이외에도 UDA에서는 성능향상을 위해 신뢰기반 마스킹이라는 방법도 이용한다. 이방법은 LD가 아주 작을 때 ULD를 선별하여 학습에 사용하는 방법이다. 이 방법은 ULD의 사례 x에 대해 모델 예측값이 사전에 정의된 임계값보다 작다면 모델이 분류에 대한 확신이 없다고 볼 수 있기 때문에 이렇게 확신이 없는 데이터는 학습에서 제외하고 임계값보다 큰 데이터들을 학습에 이용하는 방법이다.

3. 성능 향상 방법

3.1. 일칭성 손실 함수

UDA에서 ULD의 일치성 손실함수는 KL-정보량을 이용하는데 KL-정보량은 대칭이 아니므로 확률분포 P 와 Q 의 순서에 따라 정보량이 다르게 계산될 수 있을 뿐 아니라 서포터(support)가 다른 두 분포의 정보량은 항상 ∞ 값이 되므로 손실함수로는 적합하지 않다. JS-정보량은 KL-정보량의 비대칭성을 보완할 수 있는데 아래와 같이 표현된다.

$$\begin{aligned} D_{\mathrm{JS}}(p\|q) &= \frac{1}{2}\left(D_{\mathrm{KL}}(p\|m) + D_{\mathrm{KL}}(q\|m)\right), \quad m = \frac{p+q}{2} \\ &= \frac{1}{2}\int\left(p(x)\log\frac{2p(x)}{p(x)+q(x)} + q(x)\log\frac{2q(x)}{p(x)+q(x)}\right)dx. \end{aligned} \tag{3.1}$$

본 연구에서는 KL-정보량 대신 JS-정보량을 사용하여 UDA의 성능을 향상시키고자 한다.

3.2. RTSA

UDA에서 LD에는 과적합을 막기 위한 목적으로 TSA를 사용한다. TSA에서 사용되는 임계값은 $(1/K)$에서 1까지 증가하는데 모델의 신뢰도가 계산된 임계값보다 낮은 경우에만 학습을 진행한다. 모델을 학습할 때 데이터를 하나씩 학습하는 것이 아닌 배치 단위로 학습을 진행하는데 이때 한 배치에서 어떤 데이터에 대한 모델의 신뢰도가 낮다면 그 배치를 학습하는 단계에서는 모델이 더 나은 방향으로 업데이트되기 어려울 수 있다. 그리고 이미지 데이터에는 신뢰도가 낮은 데이터가 아주 빈번하게 등장할 수 있는데 이는 다음 단계의 학습에도 많은 영향을 미칠 수 있으므로, 본 논문에서는 모델의 신뢰도가 임계값보다 높은 경우에만 학습을 진행하도록 하여 모델의 훈련을 가속화(reverse-TSA; RTSA)하는 것이 성능 향상에 더 도움이 될 것이기 때문에 RTSA를 제안한다. TSA의 임계값을 계산하는 과정은 동일하게 한다. 그리고 이와 같은 이유로 UDA에서 성능향상을 위해 사용한 신뢰기반 마스킹을 제거하는 방법도 고려한다.

4. 실험

4.1. 분석 데이터

분석에 사용된 데이터는 CIFAR-10 (Canadian Institute For Advanced Research)이다. CIFAR-10 데이터는 총 10개의 범주로 이루어진 이미지 데이터인데 크기가 50,000인 훈련용데이터와 크기가 10,000인 시험용데이터로 구성되어있고 각 이미지의 크기는 32픽셀 × 32픽셀이다. Figure 3은 CIFAR-10 데이터의 각 범주에 해당하는 이미지의 예시를 보여준다.

4.2. 분석 방법

기존 UDA에서 ULD는 데이터 증대를 통해 각 이미지 1장당 100장씩 생성하였지만 본 연구에는 제한된 계산 환경 때문에 10장씩 생성하였다. 그리고 본 실험에서는 Python 3.5.2, Torch 0.4.1 그리고 Wide-ResNet-28-2 (Zagoruyko와 Komodakis, 2016) 심층신경망을 이용하였다. LD와 ULD의 배치크기(batch size)는 각각 32와 80이고 실험은 10번 반복하였고, 각 실험은 1000 에폭(epoch) 실행하였다. 먼저 UDA의 성능을 확인하였다.

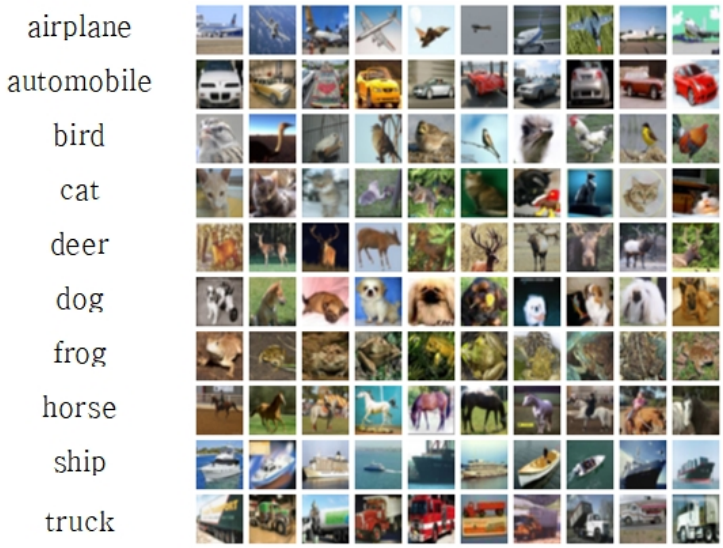

Figure 3: *Example image of CIFAR-10.*

Table 2: Mean and standard deviation of accuracy for SL, UDA and UDA with RTSA

Method	Mean	Standard deviation
SL	0.62	0.0033
UDA	0.71	0.0067
UDA with RTSA	0.78	0.0050

4.3. RTSA와 UDA 성능 확인

UDA의 성능을 확인하기 위해 크기가 4,000인 LD만 이용한 지도학습(SL)과 크기가 46,000인 ULD를 이용한 UDA의 실험결과를 Table 2에 나타내었다. 그리고 UDA에서 RTSA를 사용하였을 때의 결과를 같이 나타내었다. 이 실험에서는 기존의 UDA처럼 $\tau = 0.8$로 결정하였고, 식 (2.4)의 $\lambda = 0.0125$로 경험적으로 결정하였다. 평가는 크기가 10,000인 시험용데이터를 이용하였고 실험은 각 10번씩 실시하였다. 실험 결과는 정확도의 평균과 표준편차로 요약하였다. LD만을 이용하여 학습하였을 때보다 UDA의 정확도의 평균이 약 15% 높은 성능을 보이지만 표준편차가 2배 정도로 커지는 결과를 나타낸다. RTSA 사용하였을 때는 기존의 UDA보다 약 10% 향상된 정확도의 평균과 조금 더 낮은 변동성을 보였다.

4.4. 신뢰기반 마스킹과 JS-정보량

ULD의 일치성 손실함수를 KL-정보량 대신 JS-정보량으로 대체하였을 때, 그리고 기존의 UDA에서 사용한 신뢰기반 마스킹을 제거한 실험을 실행하였다. 신뢰기반 마스킹 제거는 UDA–C로 표현하였다. 평가는 시험용데이터를 이용하였고 실험은 각 10번씩 실시하였다. 기존의 UDA처럼 $\tau = 0.8$로 결정하였고, 경험적으로 식 (2.4)의 $\lambda = 1$로 결정하였다. 실험 결과는 정확도로 측정하여 평균과 표준편차로 요약하여 Table 3에 나타내었다. 신뢰기반 마스킹을 제거하였을 때가 기존의 UDA보다 약 5% 향상된 정확도의 평균과 아주 낮은 표준편차를 보였다. 그리고 신뢰기반 마스킹을 제거하고 손실함수를 JS-정보량으로 대체하였을 때는 약 7.5% 향상된 정확도 평균과 낮은 표준편차를 나타내었다.

Table 3: Mean and standard deviation of accuracy for UDA and without confidence based masking and with JS-divergence

Method	Mean	Standard deviation
UDA	0.80	0.0097
UDA-C	0.84	0.0043
UDA-C : JS	0.86	0.0042

5. 결론 및 제언

본 연구에서는 준지도학습에서 훌륭한 성과를 보이는 UDA의 성능 향상을 위한 세가지 방법을 제안한다. 먼저 ULD의 일치성 손실함수로 사용되는 KL-정보량 대신 대칭성을 가질 뿐 아니라 서포터(support)가 서로 다른 분포에도 적용 가능한 JS-정보량을 제안한다. 그리고 이미지 데이터에는 신뢰도가 낮은 사례가 아주 빈번하게 등장할 수 있음을 고려하여 LD를 사용하는 모델의 신뢰도가 임계값보다 높은 경우에만 학습을 진행하도록 하여 훈련을 가속화하는 RTSA를 제안한다. 또한 신뢰기반 마스킹이 이미지 데이터의 특성을 제대로 반영하지 못하는 것으로 고려되어 신뢰기반 마스킹을 제거하는 것을 제안한다.

CIFAR-10 데이터와 ResNet 심층신경망을 이용한 실험결과로써 RTSA 사용하였을 때는 기존의 UDA보다 약 10% 향상된 정확도 평균과 더 낮은 변동성을 보였다. 그리고 신뢰기반 마스킹을 제거하였을 때는 약 5% 향상된 정확도 평균과 아주 낮은 변동성을 보였고 신뢰기반 마스킹을 제거하고 손실함수를 JS-정보량으로 대체하였을 때는 약 7.5% 향상된 정확도 평균과 낮은 표준편차를 나타내어 제안된 방법의 타당성을 확인할 수 있었다.

실험환경의 한계로 인하여 본 연구에서 더 많은 데이터와 초모수의 선택을 충분하게 고려하지 못한 점은 추후의 연구과제로 남기고자 한다.

References

Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-supervised learning*, Cambridge, Massachusetts: MIT Press.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2018). AutoAugment: Learning augmentation policies from data, `arXiv:1805.09501`.

Cubuk, E. D., Zoph, B., Shlens, J, and Quoc, V. L. (2019). RandAugment: Practical data augmentation with no separate search. `arXiv:1909.13719`.

Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, 529–536.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**, 119-–139.

Laine, S. and Aila, T. (2016). Temporal ensembling for semi-supervised learning. *arXiv:1610.02242*.

Miyato, T., Maeda, S., Ishii, S., and Koyama, M. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979–1993.

Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., and Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms, In *Advances in Neural Information Processing Systems*, 3235–3246.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets

improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 1195–1204.

Xie, Q., Dai, Z., Hovy, E., Luong, M. T., and Le, Q. V. (2019). Unsupervised data augmentation for consistency training, arXiv:1904.12848.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks, arXiv:1605.07146.

Received November 17, 2020; Revised December 31, 2020; Accepted January 4, 2021

일치성규칙과 목표값이 없는 데이터 증대를 이용하는 학습의 성능 향상 방법에 관한 연구

김현웅[a], 석경하[1,b]

[a]인제대학교 해운대백병원 임상시험센터, [b]인제대학교 통계학과

요 약

준지도학습(semi-supervised learning)은 목표값이 있는 데이터와 없는 데이터를 모두 이용하는 학습방법이다. 준지도학습에서 최근에 많은 관심을 받는 일치성규칙(consistency regularization)과 데이터 증대를 이용한 준지도학습(unsupervised data augmentation; UDA)은 목표값이 없는 데이터를 증대하여 학습에 이용한다. 그리고 성능 향상을 위해 훈련신호강화(training signal annealing; TSA)와 신뢰기반 마스킹(confidence based masking)을 이용한다. 본 연구에서는 UDA에서 사용하는 KL-정보량(Kullback-Leibler divergence)과 TSA 대신 JS-정보량(Jensen-Shanon divergene)과 역-TSA를 사용하고 신뢰기반 마스킹을 제거하는 방법을 제안한다. 실험을 통해 제안된 방법의 성능이 더 우수함을 보였다.

주요용어: 데이터 증대, 일치성규칙, 준지도학습, 훈련신호강화, JS-정보량

이 논문은 본 논문은 2019학년도 인제대학교 학술연구조성비 보조에 의한 것임.

[1]교신저자: (50834) 경남 김해시 인제로 197, 인제대학교 통계학과, 통계정보연구소. E-mail: statskh@inje.ac.kr