김일성종합대학학보
수 학

주체110(2021)년
제67권 제2호

적은 개수의 훈련자료로부터 심층신경망을 학습시키기 위한 한가지 자료증식방법

리상민, 리명철

심층신경망을 학습시키고 그 성능을 향상시키는데서 나서는 한가지 문제는 훈련자료가 충분해야 한다는것이다. 특히 의학부문에서는 환자자료를 충분히 얻을수 없는것으로 하여 학습된 신경망의 일반화성능이 떨어져 널리 리용되지 못하고있다.

선행연구[3]에서는 전자병력서로부터 얻은 환자자료로 학습시킨 심층신경망을 리용하여 환자를 진단하는 방법을 제기하였다. 그러나 신경망을 mimic자료기지와 같은 공개된 자료기지에서만 훈련하고 검사를 진행하므로 일반화성능이 높지 못하고 잡음견딜성이 약한 결함이 있다. 이로부터 이미 림상실천에서 쓰이는 전문가체계로부터 자료를 생성하고 증식된 훈련자료를 리용하여 학습하는 방법[4]이 제기되였다.

현실에서는 선행연구[3]에서와 달리 주어진 환자자료가 수천명정도밖에 되지 않는것으로 하여 신경망을 효과적으로 학습시킬수 없으며 진단정확도도 높일수 없다. 또한 전문가체계가 구축되여있지 않으므로 선행연구[4]에서와 같이 전문가체계로부터 자료를 생성할수도 없고 이 방법이 림상실천에서는 효과성이 높지 못하므로 실정에 맞는 자료증식방법을 연구하여 훈련자료수를 늘이고 신경망을 학습시켜야 한다.

이로부터 론문에서는 적은 병력서자료로부터 자료증식을 진행할수 있는 환자자료모의 및 생성방법을 새롭게 제안하였으며 이 방법을 리용하여 생성한 자료에서 신경망을 학습시키고 실험을 통하여 제안한 방법의 효과성을 검증하였다.

1. 환자자료의 모의와 생성

림상실천에서 실험검사자료를 리용할 때에는 수값자료 그자체보다도 그것을 여러 등급으로 리산화한 값을 쓴다. 이로부터 환자자료를 모의하기 위하여 실험검사자료와 같은 련속적인 수값자료들을 먼저 리산화하였다.

1) 예측구간을 리용한 수값형자료의 리산화

수값형지표에 대하여 병력서로부터 얻어진 자료들로부터 직접 리산화를 진행하면 환자자료가 적고 모의하는 자료수가 주어진 자료수에 비해 아주 많기때문에 타당하지 못하다. 이로부터 매 수값형지표들에 대하여 미래다중관측의 예측구간을 구하고 리산화를 진행한다.

환자의 백혈구수와 같은 실험검사지표들이 정규분포에 따르므로 정규모집단의 미래표본평균, 미래표본표준편차의 량측예측구간을 얻었다.[1, 2]

먼저 5 300명의 환자자료로부터 수값형지표에 해당한 자료들을 뽑아내고 α를 0.01로 정하였으며 모의하여 얻으려는 환자수는 40가지 질병에 대하여 병별로 1만명씩 총 40만명으로 하였다.

실례로 $\alpha=0.01$, $n=5\ 300$, $m=400\ 000$ 인 경우 백혈구수에 대한 미래40만건의 실험자료의 평균, 표준편차의 수준이 0.99인 예측구간은 다음과 같다.[1, 2]

$\bar{Y}$의 량측예측구간 (6.24, 6.52), S_Y의 량측예측구간 (2.98, 3.34)

자료를 모의하기 위하여 모든 수값형지표를 $(-\infty, \bar{Y}-S_Y)$, $[\bar{Y}-S_Y, \bar{Y}+S_Y)$, $[\bar{Y}+S_Y, \infty)$의 세 구간으로 리산화를 진행하였다.

환자진단에 쓰이는 모든 지표들을 리산화한데 기초하여 환자자료를 모의하여 생성하였다.

2) 환자자료의 모의와 생성

매 지표의 리산화된 값들을 의학소견들로 리용하여 환자자료를 모의하고 생성하였다.

ㄱ) 환자자료의 모의

매 질병에 대하여 의학소견들의 발생확률을 결정하고 그 발생확률에 따라 매 의학소견들을 발생시키면 환자의 병력서자료와 진단을 모의할수 있다. 이로부터 매 질병별 의학소견들의 발생확률을 결정하였다.

우선 자료로부터 의학소견들의 발생확률을 추정하였다. 병 D에 걸린 환자가 의학소견 g를 가질 확률을 p라고 하자. 그러면 일반적으로 병 D에 걸린 환자가 의학소견 g를 가지는가 마는가 하는것은 2항분포 $Bi(1, p)$에 따른다. 병 D에 걸린 환자가 의학소견 g를 가질 확률 p를 크기 k인 표본 $x_1, x_2, \cdots, x_k$의 표본평균 $\hat{p}=\bar{x}=\frac{1}{k}\sum_{i=1}^{k}x_i$로 추정하고 환자의 해당 의학소견을 분포 $Bi(1, \hat{p})$에 따라 발생시킨다.

환자자료가 적은 병에 대하여 믿음성을 높이기 위하여 론문에서는 의사지식으로부터 결정론적으로 매 질병에 대한 의학소견들의 발생확률을 결정하였다. 이렇게 하여 믿음성이 낮은 질병의 부족점도 보충하고 모의하는 자료에 의사지식도 배합하였다.

우의 두가지 방법을 리용하여 발생확률을 결정하는데서 먼저 병별로 성별, 나이와 같은 인구통계학정보가 담긴 의학소견들에 대하여 결정하고 병명과 인구통계학정보가 주어진 조건하에서 나머지 의학소견들의 발생확률을 결정하였다.

론문에서는 환자수를 기준으로 하여 환자수가 N명이상인 경우에는 자료로부터 추정한 발생확률을 리용하고 그 미만일 때에는 의사지식으로부터 결정한 발생확률을 리용하였다. 기준 N을 5, 10, 15, 20, 25, 30으로 설정하여 자료생성을 진행하고 모형의 성능을 평가한 결과 $N=20$인 경우 모형의 성능이 제일 높았다.

ㄴ) 환자자료의 생성

N가지의 질병을 각각 $D_1, D_2, \cdots, D_N$이라고 하고 M개의 의학소견을 각각 $g_1, g_2, \cdots, g_M$이라고 하자.

환자의 자료생성을 다음과 같은 단계로 진행한다.

① 평등분포에 따라 N가지 질병중의 하나 D_i $(i\in\{1, 2, \cdots, N\})$를 발생시킨다.

② 병이 주어진 조건에서 성별, 나이와 같은 인구통계학정보가 담긴 의학소견들을 발생확률에 따라 발생시킨다.

③ 병과 인구통계학정보가 주어진 조건에서 나머지 의학소견들을 발생확률에 따라 발생시킨다.

④ 발생한 자료가운데서 동시에 발생할수 없는 의학소견들을 제외한다.

⑤ 제외한 동시에 발생할수 없는 소견들가운데서 하나를 다항분포를 리용하여 발생시킨다.

2. 계 산 실 험

효과성검증을 위하여 40가지 병진단을 위한 신경망모형을 작성하고 5 300명의 환자자료를 리용하여 세가지 방법으로 학습시켜 성능을 비교하였다.

모형 1은 5 300명의 환자자료에 대하여 학습한 모형이다.

모형 2는 제안한 방법을 리용하여 생성한 40만건의 자료에서 학습한 모형이다.

모형 3은 증식된 자료에서 학습된 모형 2의 무게결수를 초기무게결수로 하여 5 300명의 환자자료에 대하여 전이학습을 진행한 모형이다.

훈련은 일괄크기 128, 반복회수 50으로 진행하였다. 모형들의 최종적인 성능검사는 자료증식에 리용되지 않은 1 346명의 환자자료에 대하여 진행하였다. 모형 1, 2, 3의 *Top k* 정확도를 비교하는 방법으로 성능평가를 진행하였다.

$$Top\ k = \frac{\sum_{t=1}^{T}\sum_{j=1}^{j=k}[y^{(t)} = \hat{y}^{(t)}[j]]}{T}$$

여기서 T는 시험모임의 환자수,

$$[a=b] = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$$

이다. $\hat{y}^{(t)}[j]$는 시험모임의 t번째 환자자료에 대하여 모형이 예측한 최대 j번째 클라스이다.

모형들의 성능을 보여준 실험결과는 표와 같다.

표. 실험결과

방법	훈련자료수	검증자료수	*Top*-1정확도	*Top*-3정확도	*Top*-5정확도
환자자료로 학습한 신경망	5 300	1 346	90.13%	93.62%	96.15%
증식된 자료로 학습한 신경망	400 000	1 346	90.45%	95.89%	98.75%
전이학습한 신경망	5 300	1 346	93.27%	97.91%	99.79%

우의 실험결과는 론문에서 제안한 방법을 리용하여 증식한 자료에서 학습한 모형 2와 전이학습한 모형 3이 모형 1에 비하여 성능이 높다는것을 보여준다.

참 고 문 헌

[1] 한광룡; 통계적구간, 고등교육도서출판사, 3～19, 주체104(2015).

[2] 리상민, 한광룡; 조선민주주의인민공화국 과학원통보, 3, 5, 주체109(2020).

[3] A. Rajkomar et al.; arXiv, vol.801.07860, 1～16, 2018.

[4] A. Kannan; Proceedings of Machine Learning Research, 85, 1, 2018.

주체109(2020)년 12월 5일 원고접수

A Method of Data Augmentation to Train a Deep Neural Network with a Few Training Data

Ri Sang Min, Ri Myong Chol

In this paper, we propose a method of data augmentation to train a deep neural network with a few training data and verify it via tests. The experimental results showed the efficiency of our method.

Keyword: data augmentation