ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 June: 28(2): 19~31 http://dx.doi.org/10.13088/jiis.2022.28.2.019

데이터 증가를 통한 선형 모델의 일반화 성능 개량 (중심극한정리를 기반으로)

황두환
육군3사관학교 국방시스템과학과
(kkoo6103@gmail.com)

기계학습 모델 구축 간 트레이닝 데이터를 활용하며, 훈련 간 사용되지 않은 테스트 데이터를 활용하여 모델의 정확도와 일반화 성능을 판단한다. 일반화 성능이 낮은 모델의 경우 새롭게 받아들이게 되는 데이터에 대한 예측 정확도가 현저히 감소하게 되며 이러한 현상을 두고 모델이 과적합 되었다고 한다. 본 연구는 중심극한정리를 기반으로 데이터를 생성 및 기존의 훈련용 데이터와 결합하여 새로운 훈련용 데이터를 구성하고 데이터의 정규성을 증가시킴과 동시에 이를 활용하여 모델의 일반화 성능을 증가시키는 방법에 대한 것이다. 이를 위해 중심극한정리의 성질을 활용해 데이터의 각 특성별로 표본평균 및 표준편차를 활용하여 데이터를 생성하였고, 새로운 훈련용 데이터의 정규성 증가 정도를 파악하기 위하여 Kolmogorov-Smirnov 정규성 검정을 진행한 결과, 새로운 훈련용 데이터가 기존의 데이터에 비해 정규성이 증가하였음을 확인할 수 있었다. 일반화 성능은 훈련용 데이터와 테스트용 데이터에 대한 예측 정확도의 차이를 통해 측정하였다. 새롭게 생성된 데이터를 K-Nearest Neighbors(KNN), Logistic Regression, Linear Discriminant Analysis(LDA)에 적용하여 훈련시키고 일반화 성능 증가정도를 파악한 결과, 비모수(non-parametric) 기법인 KNN과 모델 구성 간 정규성을 가정으로 갖는 LDA의 경우에 대하여 일반화 성능이 향상되었음을 확인할 수 있었다.

주제어 : 일반화, 정규성, 중심극한정리, 기계학습

논문접수일 : 2022년 3월 5일 논문수정일 : 2022년 3월 28일 게재확정일 : 2022년 4월 7일
원고유형 : Regular Track 교신저자 : 황두환

1. 개요

통상 우리는 Machine Learning(ML) 모델 구축 간 트레이닝 데이터를 활용하며, 훈련 간 사용되지 않은 테스트 데이터를 활용하여 모델의 정확도와 일반화 성능을 판단한다. 일반화 성능이 낮은 모델의 경우 새롭게 받아들이게 되는 데이터에 대한 예측 정확도가 현저히 감소하게 되며 이러한 현상을 두고 모델이 과적합(overfitting) 되었다고 한다 (Gareth et al., 2017). 과적합이 발생하는 이유에는 크게 두가지 경우가 있다. 첫째, 훈련용 데이터의 수가 적거나 대표성을 띄는 데이터의 수가 적어 모델이 학습을 진행함에 있어 해당 데이터가 지니고 있는 노이즈에 대해 학습하는 경우 과적합이 발생할 수 있다(Ying, 2019). 이 경우 훈련용 데이터의 수를 늘리거나, 대표성이 있는 데이터에 대해 학습이 가능하도록 알고리즘을 구성하는 것이 중요하다 (Paris et al., 2003). 둘째, 모델이 너무 많은 가정을 내포하고 있거나 수많은 변수들을 포함하고 있어 복잡도가 높을 경우 모델은 훈련용 데이터에 한하

여 높은 정확도를 보일 수 있다(Paris et al., 2003). 과적합이 발생된 모델의 경우 새로운 데이터가 입력되었을 때 전혀 다른 결과를 낳을 수 있다. 이러한 과적합을 방지하기 위해서 1.모델이 훈련용 데이터에 대하여 완전히 훈련을 하지 못하도록 훈련을 조기에 종료시키거나(Early Stopping), 2.모델의 변수 개수를 축소시키거나(Dimensionality Reduction), 3.훈련용 데이터의 수를 증가시키거나, 4.여러 방법의 규제(Regularization)를 적용하는 등의 방법을 이용한다(Ying, 2019). 본 연구는 여러 방법 중 훈련용 데이터의 양을 증가시켜 정규성을 증가시키고 이를 바탕으로 모델의 일반화 성능을 증가시키는 방법에 관한 것이다. 모델을 훈련시킴에 있어 훈련용 데이터의 양과 질은 굉장히 중요하다(Sarker, 2021). 여러 ML 알고리즘 중 Linear Discriminant Analysis(LDA), 선형회귀(Linear Regression) 등 정규성 가정을 내포하고 있는 알고리즘의 경우 훈련용 데이터가 정규분포를 따를 경우 여러 이점을 얻을 수 있다(Abhishek, 2020). 그러나 우리가 현실에서 다루는 데이터의 경우 그 분포가 다양하며 때로는 왜곡된 경우가 많다. 따라서 본 연구에서는 중심극한정리를 활용하여 훈련용 데이터의 수를 증가시켜 기존의 데이터가 갖는 비정규성(Non-Normality)을 감소시키고 이를 통해 모델의 일반화 성능을 증가시킬 수 있는 방법에 대해 알아보았다.

2. 연구방법

본 연구의 목적은 모델의 일반화 성능을 증가시킴에 있으며 아래의 순서에 따라 연구가 진행되었다.

1. 기존에 보유한 데이터의 각 특성 별 모평균 및 표준편차를 구한다.
2. 중심극한정리의 성질을 활용하여 특정 표본의 수(t)에 따른 표본평균 및 표준편차를 구한다.
3. 2.에서 구한 표본평균 및 표준편차를 따르는 p개의 데이터를 생성한다.
4. 기존의 데이터(existed data)와 생성된 데이터를 통합하여 새로운 훈련용 데이터를 생성하고 이를 newly data라고 한다.
5. 기존 데이터와 새로운 훈련데이터의 정규화 정도를 비교하기 위하여 Kolmogorov-Smirnov 통계량을 산출한다.
6. 기존데이터와 새로운 훈련데이터를 활용하여 여러 ML모델을 각각 훈련시킨다.
7. 기존데이터와 새로운 훈련데이터를 통해 훈련시킨 모델의 일반화 성능정도를 비교한다.

2.1. 중심극한정리(Central Limit Theorem, CLP)의 정의와 성질

중심극한정리는 모집단의 분포가 어떤 분포라도 추출되는 표본의 수가 충분히 커진다면 표본평균의 분포는 정규분포에 근사함을 의미한다(Kwak et al., 2017). 이때 모집단의 평균과 분산이 각각 μ, σ^2 이라고 할 때, 표본의 개수가 n이면 표본평균의 평균과 분산은 μ, $\frac{\sigma^2}{n}$ 이다.

2.2. 정규성 검정 방법

본 연구의 목적은 CLP를 기반으로 데이터를 생성하고 이를 훈련용 데이터에 추가하여 훈련용 데이터의 정규화 정도를 증가시키고, 이를 바탕으로 모델의 일반화 성능을 증가시키는 것이다. 따라서 정규성의 정도가 기존의 데이터보다 새로운 훈련용 데이터에서 얼마만큼 향상되었는지 파악하는 절차가 필요했다. 대표적인 정규성 검정의 방법에는

Kolmogorov-Smirnov(K-S) 정규성 검정이 있다 (Asghar et al., 2012). 한 표본에 대한 K-S 정규성 검정은 관측된 데이터의 분포와 가정된 분포 사이의 적합성의 정도를 파악하는 방법으로 아래와 같은 절차를 거쳐 진행된다(Lilliefors, 1967).

1. 관측된 데이터를 크기 순으로 나열한다.
2. 각 데이터에 대한 누적확률을 구한다.
3. 이를 가정된 분포의 누적확률과 비교한다.
8. 두 누적확률 값 간의 차이의 최대값이 임계값보다 작을 때 가정된 분포의 모집단에서 표본을 추출했다는 귀무가설을 채택할 수 있다. 따라서 K-S 정규성 검정의 경우 통계량(K-S statistics)의 수치가 작을수록 데이터가 가정된 분포와 유사하다고 판단할 수 있다(Massey, 1951). 본 연구에서 가정된 분포는 결국 정규분포를 의미한다.

2.3. 머신러닝 분류 모델

중심극한정리를 기반으로 새롭게 생성되어 정규성 정도가 높아진 훈련용 데이터가 모델의 일반화 성능에 어떤 영향을 미치는지 확인하기 위하여 비모수기법(non-parametric) 중 하나인 KNN알고리즘과 정규성을 기본 가정으로 하는 Linear Discriminant Analysis(LDA), 그렇지 않은 로지스틱 회귀분석(Logistic Regression)에 새로운 훈련용 데이터를 적용해보았다.

2.3.1. K-Nearest Neighbor Classifier (KNN)

KNN 알고리즘은 non-parametric 기법으로 분류 및 회귀 문제에 모두 적용 가능하다(Gareth et al., 2017). 새로운 데이터가 입력되었을 경우 이는 인접한 k개의 훈련용 데이터와의 유사성에 따라 분류된다(Zhang, 2016). KNN 알고리즘에서 사용되는 매개변수는 k이며, 이는 분류를 위해 고려할 인접 데이터의 수를 의미한다. k가 작아질수록 모델이 과적합되는 경향이 있다 (Zhang, 2016).

2.3.2. Logistic Regression

로지스틱 회귀분석의 목적은 종속변수와 독립변수의 관계를 특정 함수로 표현하고, 일반 회귀분석과 동일하게 향후 예측 모형에 사용하는 것이다 (Fleiss et al., 1986). 로지스틱 회귀분석은 선형 결합으로 종속 변수를 설명할 수 있다는 점에서 선형 회귀분석과 유사하다. 하지만, 종속 변수가 범주형 데이터이며 입력 데이터가 주어질 때 데이터의 결과가 여러 그룹으로 나뉘기 때문에 로지스틱 회귀는 회귀가 아닌 분류 기법으로 분류된다(Köküer et al., 2007). 또한 선형 회귀분석과는 달리 선형성 및 정규성과 같은 주요 가정을 만족시키지 않아도 만족스러운 결과를 얻을 수 있다(Leung, 2022).

2.3.3. Linear Discriminant Analysis (LDA)

LDA는 차원축소를 활용한 지도학습 알고리즘이다(Balakrishnamam et al., 1998). LDA에서는 각 군집을 분류하고 군집 별 결정경계를 구성하기 위한 정보를 유지함과 동시에 훈련용 데이터가 갖는 차원을 감소시키기 위한 알고리즘이다. LDA에서는 종속변수의 군집끼리 최대한 분리될 수 있는 축을 찾고, 데이터를 이 축에 정사영(projection)시켜 구분하려는 범주의 평균은 멀고, 각각의 분산이 작아질 수 있는 차원을 찾는다(Malik, 2022). 또한 LDA에서는 훈련용 데이터의 각 특성 값들이 정규분포를 따르며 랜덤으로 추출되었음을 가정한다(Pohar et al., 2004).

〈Table 1〉 Overall information of dataset

Feature Variables											Target Variable
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6.0
7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
8.1	0.22	0.43	1.5	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6.0

2.4. 데이터 소개 및 전처리

2.4.1. 데이터 소개

본 연구에서 이론을 적용하는데 사용된 데이터는 UCI Machine Learning Repository 출처의 Wine Quality Dataset이다. 데이터는 11개의 특성 값을 가지고 있으며, 레드 와인 4,898건, 화이트 와인 1,599건으로 구성 되어있다. 종속변수(Target Value)는 와인의 질로써 1부터 10까지의 수치를 갖는다(Cortez, 2003). 본 연구에서는 레드 와인과 관련된 데이터만 사용하였으며 아래 Table 1.은 본 연구에서 사용한 데이터 중 10개의 데이터를 나타낸 것이다.

2.4.2. 데이터 전처리

모델을 훈련시킴에 있어 보유하고 있는 모든 데이터를 사용할 경우 모델은 과적합되기 쉬우며, 새로 입력된 데이터에 대한 예측 정확도가 현저히 감소하게 된다(Gareth et al., 2017). 따라서 4,898건의 전체 데이터를 3,673개의 훈련용 데이터와 1,225개의 테스트 데이터로 구분하여 모델 훈련을 진행하였다. 또한 연구결과를 명확히 보이기 위하여 기존 1~10의 범주로 구분되어 있던 데이터를 1~7까지의 범주는 0으로, 8~10까지의 범주는 1로 변환하여 기존 다중분류 문제를 이진분류 문제로 변환하였다.

2.4.2. 데이터 생성

중심극한정리를 바탕으로 생성 및 기존의 데이터에 추가하여 새로운 훈련용 데이터를 구성하였으며 이 데이터가 가지는 평균 및 분산을 파악하였다. 이를 통해 추가하고자 하는 데이터의 표준편차의 정도 및 데이터의 개수를 판단하고자 하였다. 기존에 가지고 있었던 데이터의 수를 n, 평균 및 분산을 각각 μ, 라고 가정하자. 중심극한정리의 성질에 따라,

t개의 샘플을 추출할 경우 표본평균의 평균 및 표준편차는 각각 μ, 이다. 표본평균의 표준편차에 k를 곱하여 생성되는 데이터의 분포 정도를 조절할 수 있도록 하였다. 따라서 생성되는 데이터(p개)는 평균과 표준편차를 각각 μ, 로 하는 정규분포를 따르도록 설계되었다. 이렇게 생성된 p개의 데이터와 기존 보유하고 있었던 n개의 데이터를 합하여 (n+p)개의 훈련용 데이터를 구성하였으며 이 훈련용 데이터의 평균과 분산은 각각 아래의 Equation 1, 2와 같다.

$$\mu_{new} = \frac{1}{n+p}\left[(x_1 + x_2 + \cdots + x_n) + (x_{s,1} + x_{s,2} + \cdots + x_{s,p})\right]$$

$$= \frac{1}{n+p}(n \cdot \mu + p \cdot \mu) = \mu$$

〈Equation 1〉

$$\sigma_{new}^2 = \frac{1}{n+p}\left[\sum_{i=1}^{n}(x_i - \mu)^2 + \sum_{j=1}^{p}(x_{s,j} - \mu)^2\right]$$

$$= \frac{1}{n+p}\left(n \cdot \sigma^2 + p \cdot \frac{k^2\sigma^2}{t}\right)$$

$$= \left(\frac{n \cdot t + p \cdot k^2}{n \cdot t + p \cdot t}\right) \cdot \sigma^2$$

〈Equation 2〉

새롭게 생성된 데이터를 활용하여 모델 훈련을 진행함에 있어 과적합을 방지하기 위해서는 기존의 데이터가 가지고 있는 분산보다 새롭게 생성된 데이터의 분산을 크게 설정하여 모델이 학습할 데이터의 범주를 넓게 부여해야 한다는 가정을 통하여 아래의 Equation 3을 수립하였다.

$$\sigma^2 < \sigma_{new}^2$$

$$\rightarrow \sigma^2 < \left(\frac{n \cdot t + p \cdot k^2}{n \cdot t + p \cdot t}\right) \cdot \sigma^2$$

$$\rightarrow n \cdot t + p \cdot t < n \cdot t + p \cdot k^2$$

$$\rightarrow t < k^2$$

〈Equation 3〉

따라서 본 연구에서는 다양한 30이상의 t와 k값을 활용하여 새로운 훈련용 데이터를 생성하고 이를 바탕으로 모델의 일반화 성능이 얼마나 향상되었는지를 확인해보도록 하겠다.

3. 결과

3.1. 정규성(Normality) 증가

아래의 Table 2.는 기존의 훈련용 데이터에 대한 K-S통계량과 p=1000으로 고정된 상태에서 여러 t와 k값을 적용한 새로운 훈련용 데이터에 대한 K-S통계량을 나타낸다.

Table 2.와 Figure 1.에서 보면 기존 훈련용 데이터(existed data)에 비하여 새로운 훈련용 데이터에 대한 K-S통계량이 감소했음을 확인할 수 있다. 이것은 새로운 훈련용 데이터가 기존 데이터보다 더 높은 정규성 가지고 있음을 의미한다.

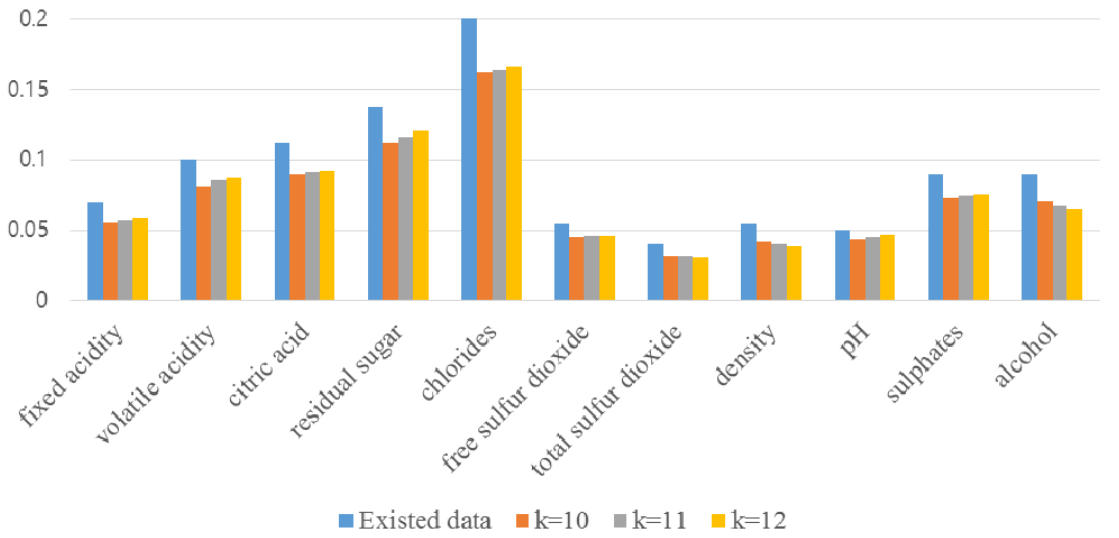
3.2. 일반화 성능 측면

Table 2.를 살펴보면 t가 30이상일 경우 K-S통계량에는 큰 변화가 없음을 확인할 수 있으므로, 모델의 일반화 성능측면을 고찰하기 위해 t를 50으로 고

〈Table 2〉 The Kolmogorov-Smirnov Statistic of existed and newly training data

p=1000	t	30			50			100		
	k	8	9	10	10	11	12	14	15	16
Variables	The Kolmogorov-Smirnov Statistics									
	existed data	newly data								
fixed acidity	0.069	0.056	0.057	0.061	0.055	0.057	0.058	0.055	0.056	0.057
volatile acidity	0.100	0.082	0.086	0.090	0.081	0.085	0.087	0.082	0.083	0.086
citric acid	0.112	0.090	0.092	0.094	0.089	0.091	0.092	0.089	0.091	0.091
residual sugar	0.137	0.113	0.119	0.123	0.112	0.116	0.120	0.112	0.114	0.118
chlorides	0.201	0.163	0.165	0.169	0.162	0.164	0.166	0.162	0.163	0.165
free sulfur dioxide	0.054	0.046	0.046	0.046	0.045	0.046	0.046	0.045	0.045	0.046
total sulfur dioxide	0.040	0.032	0.031	0.032	0.032	0.032	0.031	0.032	0.032	0.032
density	0.054	0.041	0.038	0.036	0.042	0.040	0.038	0.042	0.041	0.039
pH	0.050	0.044	0.046	0.051	0.043	0.045	0.047	0.043	0.044	0.046
sulphates	0.089	0.073	0.074	0.076	0.073	0.074	0.075	0.073	0.073	0.074
alcohol	0.089	0.069	0.067	0.063	0.070	0.068	0.065	0.070	0.068	0.067

정하고 다양한 k값과 p값을 적용하여 새로운 훈련 사용되지 않은 새로운 데이터에 대해 얼마나 정확



〈Figure 1〉 Decreasing K-S statistic of each variable at fixed p=1000, t=50

용 데이터를 생성하고 이를 활용해 여러 알고리즘을 훈련시켜보았다. 일반화는 알고리즘이 훈련에 한 예측을 할 수 있는지에 대한 성능을 의미하며 (Takano, 2021), 본 연구에서는 전체 데이터 중 일

〈Table 3〉 The result of KNN

The type of parameter			The existed data			The newly data		
t	k	p	training accuracy	test accuracy	difference	training accuracy	test accuracy	difference
50	10	1000	85.598	82.939	2.659	84.699	83.510	1.189
		2000				84.893	82.204	2.689
		3000				84.715	82.286	2.429
	11	1000				84.849	82.449	2.400
		2000				84.664	82.367	2.297
		3000				84.475	82.694	1.781
	12	1000				84.785	82.286	2.499
		2000				84.347	83.347	1.000
		3000				84.400	82.776	1.624

부를 분리하여 테스트용 데이터(1225건)로 구분하였으며 이를 활용해 일반화 성능을 측정했다. 일반화 성능은 훈련된 모델에 트레이닝 데이터와 테스트용 데이터를 각각 적용했을 경우 도출되는 정확도의 차이를 기반으로 산정할 수 있다(Theodoridis, 2015). 즉, 두 정확도의 차이가 작을수록 일반화 성능이 높다고 할 수 있다. 두 정확도의 차이를 작게 만드는 요인에는 1. 훈련용 데이터에 대한 정확도가 낮아지는 경우 2. 테스트 데이터에 대한 정확도가 높아지는 경우가 있다. 가장 이상적인 결과는 트레이닝 데이터에 대한 정확도와 테스트 데이터에 대한 정확도가 거의 유사한 상황이다. 따라서 여러 알고리즘을 훈련시켜보고 위와 같은 기준을 바탕으로 일반화 성능을 측정하였다.

3.2.1. KNN모델의 결과

KNN 모델 훈련 간 매개변수 k는 10으로 설정하였으며 기존 훈련용 데이터 및 새로운 훈련용 데이터를 적용했을 때의 결과값은 Table 3.과 같다.

Table 3.에서 보면 전체적으로 새로운 훈련용 데이터를 활용하여 모델을 훈련하였을 때, 훈련용 데이터에 대한 정확도와 테스트용 데이터에 대한 정확도의 차이가 기존의 데이터를 활용했을 때 보다 작은 것을 확인 할 수 있다. 이러한 결과는 테스트 데이터에 대한 정확도가 높아져서 발생한 결과이기보다는 훈련용 데이터에 대한 정확도의 감소로 인한 결과로 해석된다. 이는 KNN 알고리즘의 작동원리에 기인한 것으로, 인접한 k개의 데이터를 선정하고 다수결을 통해 분류를 진행함에 있어 기존 훈련용 데이터보다 새로운 훈련용 데이터의 개수가 증가하였고 이로 인해 모델 훈련 간 강건한 학습이 이루어졌다고 판단된다. 따라서 정규성이 증가된 훈련용 데이터를 KNN알고리즘에 적용하였을 경우, 테스트용 데이터에 대한 정확도가 증가하는 경향은 없지만 훈련용 데이터에 대한 정확도의 감소로 인해 전체적인 모델의 일반화 성능은 증가한다고 결론내릴 수 있다.

〈Table 4〉 The result of Logistic regression

The type of parameter			The existed data			The newly data		
t	k	p	training accuracy	test accuracy	difference	training accuracy	test accuracy	difference
50	10	1000	81.378	78.449	2.929	81.126	78.939	2.187
		2000				80.327	79.102	1.225
		3000				81.612	77.959	3.653
	11	1000				80.226	79.02	1.206
		2000				79.834	78.041	1.793
		3000				81.252	77.632	3.62
	12	1000				79.799	78.857	0.942
		2000				79.623	77.959	1.664
		3000				80.938	77.061	3.877

3.2.2. 로지스틱 회귀모형의 결과

아래의 Table 4.는 로지스틱 회귀모형에 기존 훈련용 데이터와 다양한 매개변수(k, p)을 기반으로 생성된 새로운 훈련용 데이터를 각각 적용한 결과를 나타낸다.

Table 4.를 보면, 새로운 훈련용 데이터를 적용하여 로지스틱 회귀모형을 훈련시킨 뒤 얻은 훈련용 데이터에 대한 정확도와 테스트용 데이터에 대한 정확도의 차이가 감소하는 경우와 증가하는 경우가 혼재되어 있음을 확인할 수 있다. 따라서 로지스틱 회귀 모형의 경우는 정규성이 증가된 훈련용 데이터를 적용하여 훈련을 진행하였을 때와 기존의 데이터를 활용해 훈련을 진행한 경우에 있어 큰 차이가 없다고 결론 내릴 수 있다.

〈Table 5〉 The result of LDA

The type of parameter			The existed data			The newly data		
t	k	p	training accuracy	test accuracy	difference	training accuracy	test accuracy	difference
50	10	1000	80.942	78.857	2.085	81.061	78.857	2.204
		2000				80.98	78.612	2.368
		3000				81.358	78.531	2.827
	11	1000				81.083	79.187	1.896
		2000				80.927	78.939	1.988
		3000				80.788	78.776	2.012
	12	1000				80.89	79.184	1.706
		2000				80.592	78.939	1.653
		3000				80.489	78.857	1.632

3.2.3. LDA 모형의 결과

Table 5.는 LDA에 기존 훈련용 데이터와 다양한 매개변수(k, p)를 적용하여 생성된 새로운 훈련용 데이터를 활용하여 훈련 및 테스트한 결과를 나타낸다.

Table 5를 보면 k가 증가할수록 동일한 p에 대한 일반화 성능이 좋아지고 있음을 확인할 수 있다. 뿐만 아니라 k와 p가 증가함에 따라 전체적인 일반화 성능 또한 개량되었음을 확인할 수 있다. 이 중, k=12일 때 훈련용 데이터에 대한 정확성은 감소함과 동시에 테스트용 데이터에 대한 정확성은 증가하여 두 정확성 간의 차이가 최소화됨을 확인할 수 있었다. 따라서 LDA모형의 경우 중심극한정리 기반의 정규성이 증가된 훈련용 데이터를 적용하여 훈련을 진행하였을 때 모델의 일반화 성능이 좋아진다는 결론을 내릴 수 있다.

4. 결론

본 연구는 중심극한정리를 기반으로 훈련용 데이터를 생성하고 이를 기존의 훈련용 데이터와 결합하여 정규성을 증가시킴과 동시에 모델의 일반화 성능을 증가시키는 방법에 대한 것이다. 이를 위해 중심극한정리의 성질을 활용해 데이터의 각 특성별로 표본평균의 평균 및 표준편차를 활용하여 데이터를 생성하고 이를 기존의 훈련용 데이터와 합쳐 새로운 훈련용 데이터를 구성하였다. 새로운 훈련용 데이터를 기존 훈련용 데이터와 비교하였을 때 얼마나 정규성이 증가했는지를 파악하기 위하여 Kolmogorov-Smirnov 정규성 검정을 실시하였으며 검정결과 새로운 훈련용 데이터의 정규성이 증가했음을 확인할 수 있었다. 새로운 훈련용 데이터가 실제 모델의 일반화 성능을 증가시킬 수 있을 것인지에 대해 검증하기 위하여 비모수(non-parametric) 방법인 KNN과 모델 구성 간 정규성 가정의 유무에 따른 로지스틱 회귀분석과 LDA 모형을 새로운 훈련용 데이터와 기존의 훈련용 데이터를 이용하여 각각 훈련시켰다. 일반화 성능을 측정하는 척도로 훈련용 데이터에 대한 정확도와 테스트용 데이터에 대한 정확도의 차이 값을 사용했다. KNN의 경우 테스트용 데이터에 대한 정확도가 증가하는 경향은 없지만 훈련용 데이터에 대한 정확도의 감소로 인해 전체적인 모델의 일반화 성능은 증가한다고 결론내릴 수 있었다. 로지스틱 회귀 모형의 경우, 정규성이 증가된 훈련용 데이터를 적용하여 훈련을 진행하였을 때와 기존의 데이터를 활용해 훈련을 진행한 경우에 있어 일반화 성능 측면에서 큰 차이가 없다고 결론을 내렸는데 이는 모델 구성 간 정규성을 가정으로 두지 않는 로지스틱 회귀모형의 특성에 의한 것으로 생각된다. 마지막으로 LDA의 경우, 중심극한정리 기반의 정규성이 증가된 훈련용 데이터를 적용하여 훈련을 진행하였을 때 모델의 일반화 성능이 향상되었음을 확인할 수 있었다. 이는 훈련용 데이터가 정규분포를 따를 경우 동작을 더 잘하는 LDA의 특성에서 기인했다고 생각된다. 위의 결과들을 통해 보았을 때, 중심극한정리를 바탕으로 훈련용 데이터를 생성하고 이를 기존 데이터와 합칠 경우 데이터의 정규성이 증가하고 이를 활용하여 모델을 훈련시킬 경우 비모수 방법이나 정규성 가정을 기반으로 두는 알고리즘의 경우에 한하여 모델의 일반화 성능이 증가한다고 결론내릴 수 있다. 최근 이미지 분류를 위한 딥러닝 기법 중에는 훈련용 이미지 증가를 위한 여러가지 시도가 있었으나 머신러닝 분야에서는 데이터 증가를 위한 연구가 상대적으로 미비했다는 점에서 본 연구가 의미가 있다고 생각된다. 그러나 본 연구에서는 비

모수적 방법이나 선형 모델에 한하여 일반화 성능이 향상됐다는 점에서 한계점이 있으며 이를 보편적인 모형에도 적용할 수 있는 방안을 향후 연구 방향으로 설정토록 하겠다.

참고문헌(References)

Abhishek B, *"Normal Distribution and Machine Learning",* Available at https://medium.com/analytics-vidhya/normal-distribution-and-machine-learning-ec9d3ca05070 (Accessed 20, February, 2022).

Cortez, P., A, Cerdeira., F, Almeida., T, Matos., "Modeling wine preferences by data mining from physicochemical properties", *Decision support systems,* Vol.47, No 4(2003), 547-553.

Paris, G., D, Robilliard., C, Fonlupt., "Exploring overfitting in genetic programming", In *International Conference on Artificial Evolution (Evolution Artificielle),* Springer, Berlin, Heidelberg, 267-277.

Asghar, G., S, Zahediasl., "Normality tests for statistical analysis: a guide for non-statisticians." *International journal of endocrinology and metabolism,* Vol.10, No 2 (2012), 486.

Gareth, J., D, Witten., T, Hastie., R, Tibshirani., *"An introduction to statistical learning",* Vol.112, New York, springer, 2013.

Fleiss, J.L., W, J.B., D, A.F., "The logistic regression analysis of psychiatric data", *Journal of Psychiatric Research,* Vol 20, No 3 (1986), 195-209.

Massey Jr,. "The Kolmogorov-Smirnov test for goodness of fit." *Journal of the American statistical Association,* Vol 46, No 253 (1951), 68-78.

Leung K., *"Assumptions of Logistic Regression, Clearly Explained",* Available at https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290#b004 (Accessed 20, February, 2022).

Lilliefors, H.W., "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", *Journal of the American statistical Association,* Vol 62, No 318 (1967), 399-402.

Pohar, M., M, Blas., S, Turk., "Comparison of logistic regression and linear discriminant analysis: a simulation study", *Metodoloski zvezki,* Vol1, No 1 (2004), 143.

Malik U., *"Linear Discriminant Analysis (LDA)_in python with Scikit-Learn",* Available at https://stackabuse.com/implementing-lda-in-python-with-scikit-learn, (Accessed 20, February, 2022).

Köküer, M., RNG, Naguib., P Jančovič., H, Banfield Younghusband., and G, Roger., "Towards automatic risk analysis for hereditary non-polyposis colorectal cancer based on pedigree data." In *Outcome Prediction in Cancer,* (2007), Elsevier, 319-337

Zhang, S., X, Li., M, Zong., X, Zhu., "Efficient kNN classification with different numbers of nearest neighbors", *IEEE transactions on neural networks and learning systems,* Vol 29, No 5(2017), 1774-1785.

Balakrishnama, S., A, Ganapathiraju., "Linear discriminant analysis-a brief tutorial", *Institute for Signal and information Processing,* Vol 18 (1998), 1-8.

Sarker, I. H., "Machine learning: Algorithms, real-world applications and research

directions", *SN Computer Science,* Vol 2, No 3 (2021), 1-21.

Kwak, S. G., J.H. Kim., "Central limit theorem: the cornerstone of modern statistics", *Korean journal of anesthesiology,* Vol 70, No 2 (2017), 144.

Takano, S., *"Thinking Machine: machine learning and its hardware implementation",* 1stEdition, Elsevier, 2021.

Wayne, W., MD, Lamorte., "The Role of Probability", Available at https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html#:~:text=The%20central%20limit%20theorem%20states,will%20be%20approximately%20normally%20distributed, (Accessed 20, February, 2022).

Theodoridis, S., *"Machine learning: a Bayesian and optimization perspective".* Academic press, 2015.

Ying, X., "An overview of overfitting and its solutions", *Journal of Physics: Conference Series,* Vol 1168, No. 2 (2019), 22.

Zhang, Z., "Introduction to machine learning: k-nearest neighbors". *Annals of translational medicine,* Vol 4, No 11 (2016).

Abstract

Improvement of generalization of linear model through data augmentation based on Central Limit Theorem

Doohwan Hwang*

In Machine learning, we usually divide the entire data into training data and test data, train the model using training data, and use test data to determine the accuracy and generalization performance of the model. In the case of models with low generalization performance, the prediction accuracy of newly data is significantly reduced, and the model is said to be overfit. This study is about a method of generating training data based on central limit theorem and combining it with existed training data to increase normality and using this data to train models and increase generalization performance. To this, data were generated using sample mean and standard deviation for each feature of the data by utilizing the characteristic of central limit theorem, and new training data was constructed by combining them with existed training data. To determine the degree of increase in normality, the Kolmogorov-Smirnov normality test was conducted, and it was confirmed that the new training data showed increased normality compared to the existed data. Generalization performance was measured through differences in prediction accuracy for training data and test data. As a result of measuring the degree of increase in generalization performance by applying this to K-Nearest Neighbors (KNN), Logistic Regression, and Linear Discriminant Analysis (LDA), it was confirmed that generalization performance was improved for KNN, a non-parametric technique, and LDA, which assumes normality between model building.

Key Words : Generalization, Normality, Central Limit Theorem, Machine Learning

Received : March 5, 2022 Revised : March 28, 2022 Accepted : April 7, 2022
Corresponding Author : Doohwan Hwang

* Corresponding author: Doohwan Hwang
Korea Army Academy at Yeong-Cheon
2, Daman 9-gil, Yeongcheon-si, Gyeongsangbuk-do, Republic of Korea
Tel: *** - **** - **** E-mail: kkoo6103@gmail.com

저 자 소 개

황 두 환
육군3사관학교 국방시스템과학과 조교수
Texas A&M University Industrial engineering 석사
관심분야: Machine Learning, Algorithm, Data Science