한국데이터마이닝학회 2021 추계 학술대회

클래스 불균형 데이터 분류를 위한 GEV 활성 함수에 관한 연구

2021.11.25

홍주영[1], 신용관[1], 박혜빈[1], 박정수[1]

전남대학교 수학 / 통계학과

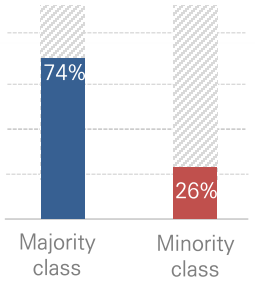
한국데이터마이닝학회 2021 추계학술대회

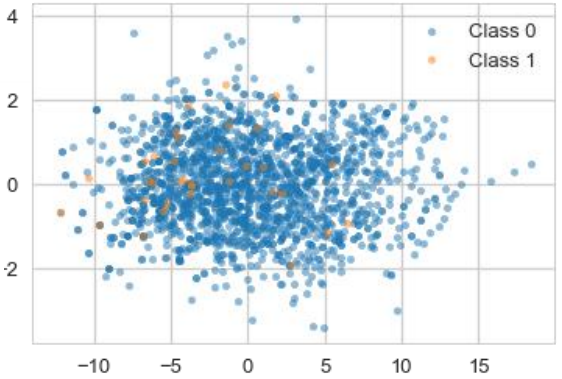
C/O/N/T/E/N/T/S

01 서론

클래스 불균형 (Class imbalance)

특정 클래스(Majority class)가 다른 클래스(Minority class)에 비하여 매우 높은 빈도로 등장하는 경우





- 실생활의 거의 대부분의 분야에서 자주 발생하는 문제

  Ex) 의료 진단, 화재 발생 탐지, 사기 거래 감지 등

- 전통적 알고리즘 적용시, 정확도 향상을 위해 소수 클래스를 무시하는 경향 존재

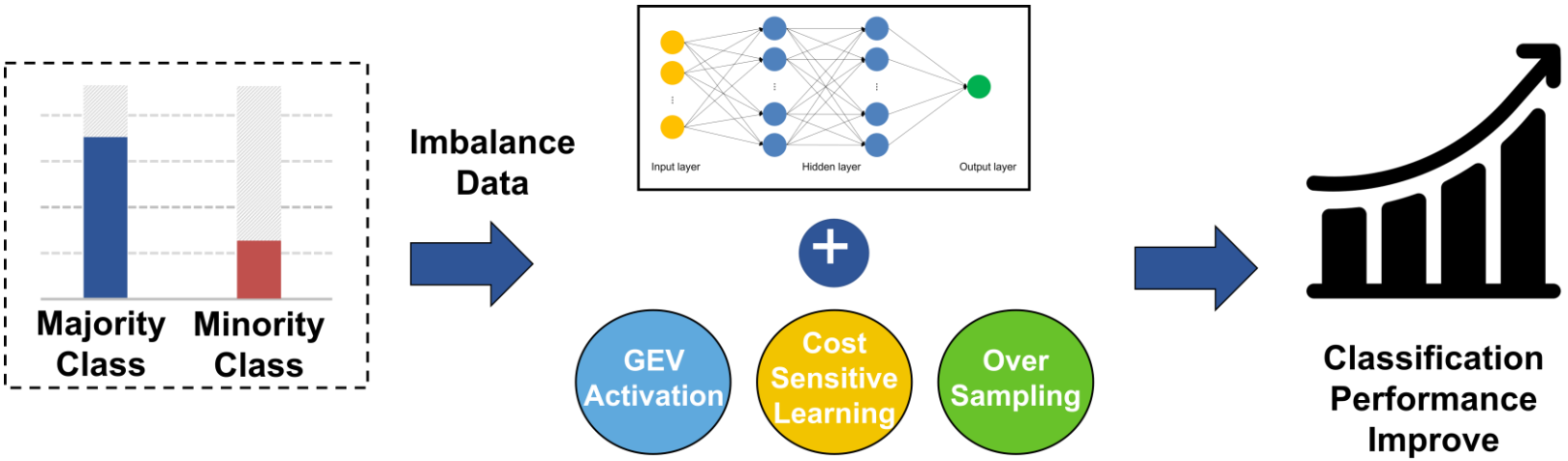
- 대부분 소수 클래스가 우리의 관심, 전체적인 정확도 보다 이를 잘 분류하는 것이 중요

- 데이터의 종류가 다양해지고, 데이터를 수집하는 환경이 복잡해짐에 따라 불균형도가 높아지는 등 여러 문제가 발생

➔ 복잡한 구조의 불균형 데이터를 효과적으로 분류할 수 있는 방안이 꼭 필요

01 서론

연구의 목적



1. 인공신경망에 최근 연구된 GEV-활성함수를 적용하고,
   여기에 기존 방법(오버 샘플링 및 비용민감학습)을 결합하여 더 높은 예측 정확도를 갖는 모형을 제시

2. 제안하는 방법에서의 최적 sampling 비율 탐색 후 제시

➔ 클래스 불균형 데이터에 대한 분류 성능 향상

02 연구 방법

GEV Activation | Cost Sensitive Learning | Over Sampling

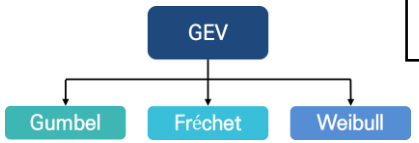
선행 연구

GLM Link ➔ NN Activation

GLM의 비대칭 Link Function으로 GEV(Generalized extreme value) 함수를 적용하여 분류 성능 향상

X. Wang, D.K. Dey, Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption, Ann. Appl. Stat. 4 (4) (2010) 2000–2023, http://dx.doi.org/10.1214/10– AOAS354.

Gumbel ➔ GEV



이진 분류 상황에서 neural network의 활성함수로 Gumbel 분포를 사용하여 분류 성능 향상

Lkhagvadorj Munkhdalai, Tsendsuren Munkhdalai, Keun Ho Ryu. GEV-NN: A deep neural network architecture for class imbalance problem in binary classification. Knowledge-Based Systems. 194. 2020. p. 105534, https://doi.org/10.1016/j.knosys.2020.105534.

코로나(COVID-19) 진단 CNN모델에 GEV활성함수를 적용하여 분류 성능 향상

J. Bridge et al., "Introducing the GEV Activation Function for Highly Unbalanced Data to Develop COVID-19 Diagnostic Models," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2776-2786, Oct. 2020, doi: 10.1109/JBHI.2020.3012383.

➔ Future work : data-efficient methods also needs to be assessed

➔ GEV 활성함수에 Cost-Sensitive Learning 및 Over-sampling을 결합한 방법 제안

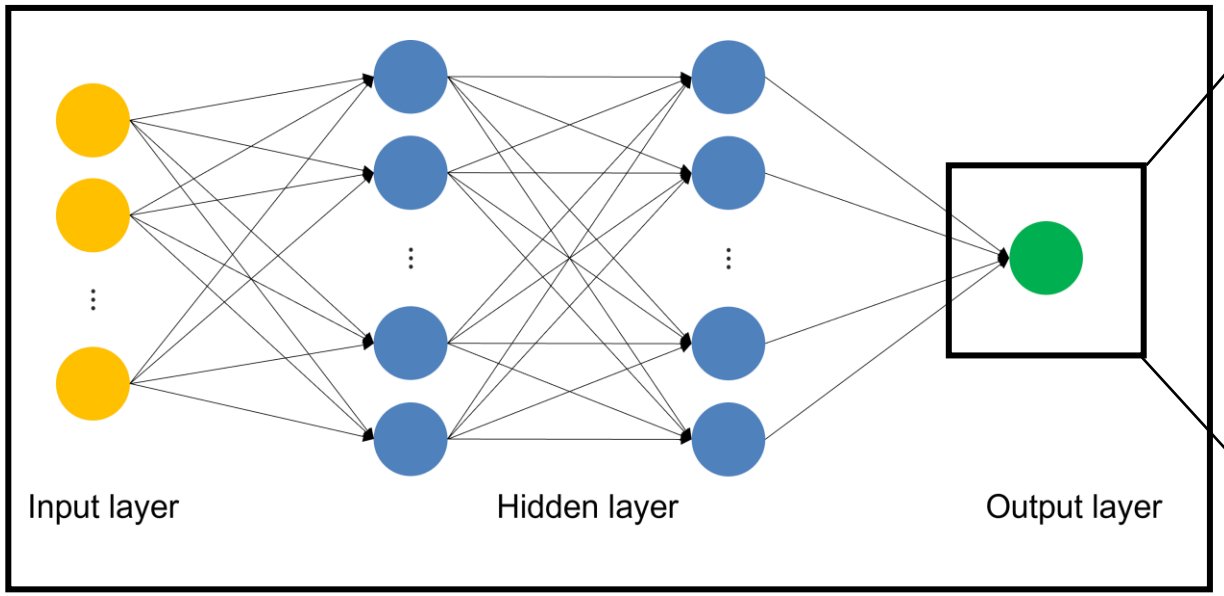
02 연구 방법

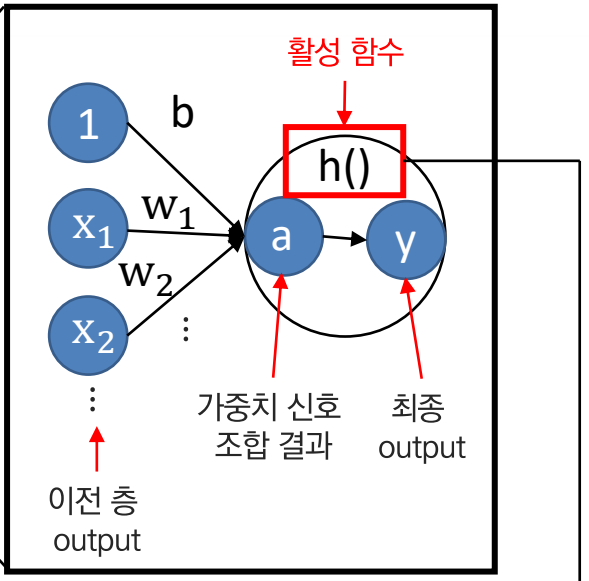
기존 방법

GEV Activation

Cost Sensitive Learning

Over Sampling

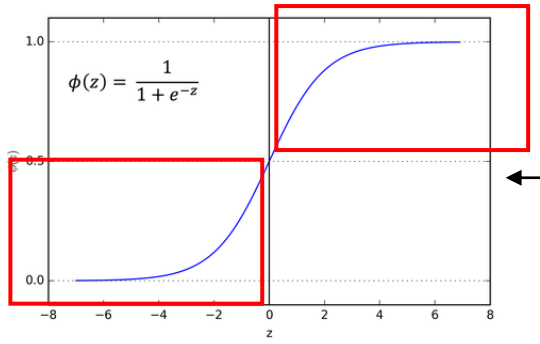




활성함수로 시그모이드 함수를 사용

- 모든 실수 입력을 미분 가능한 0~1 사이 값으로 반환
  (Binary Classification에 적절함)
- 대칭적 구조

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$



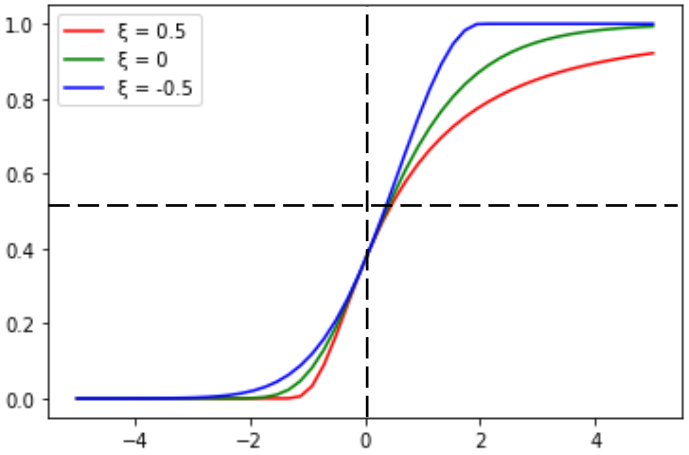
02 연구 방법

제안하는 방법

GEV Activation

Cost Sensitive Learning

Over Sampling



활성함수로 GEV분포의 CDF를 사용
(3개의 모수 μ, σ, ξ 는 오차 역전파 방법으로 다른 가중치와 동일하게 추정)

- 모든 실수 입력을 0~1 사이 값으로 반환
  (Binary Classification에 적절함)

- ξ값에 따라 비대칭 모양을 가지므로,
- 불균형 데이터에 알맞은 결정 경계를 찾아줄 것으로 기대

GEVD (Generalized extreme value distribution)

- If $\xi = 0$

  $G(s) = \exp[-exp\{-s\}], \quad s = \frac{x-\mu}{\sigma}$

  where $-\infty < s < \infty, -\infty < \mu < \infty, \sigma > 0$

- If $\xi \neq 0$

  $G(s) = \exp\{-[1 + \xi(s)]^{-1/\xi}\}, \quad s = \frac{x-\mu}{\sigma}$

  $defined\ on\ \{s : 1 + \xi s > 0\}$, where $-\infty < \mu < \infty, \sigma > 0, -\infty < \xi < \infty$

02 연구 방법

GEV Activation | Cost Sensitive Learning | Over Sampling

Cost Sensitive Learning

소수 클래스에 대한 오분류 비용을 증가시킴 ➔ 전체 오차를 최소화하는 것이 아닌, 총 오분류 비용을 최소화 하도록 학습

α – balanced Focal Loss

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

$$p_t = \begin{cases} p & if\ y = 1\ \text{(positive)} \\ 1-p & otherwise, \end{cases}$$

α_t ➔ pos/neg example의 중요도 균형을 맞춤

γ ➔ Focusing parameter (〉 0)
down weighting 정도를 조절

CV 등으로 구함

선행연구
(α_t : 0.25, γ : 2.0)

Dynamic 하게 Scaling 되는 Cross-Entropy

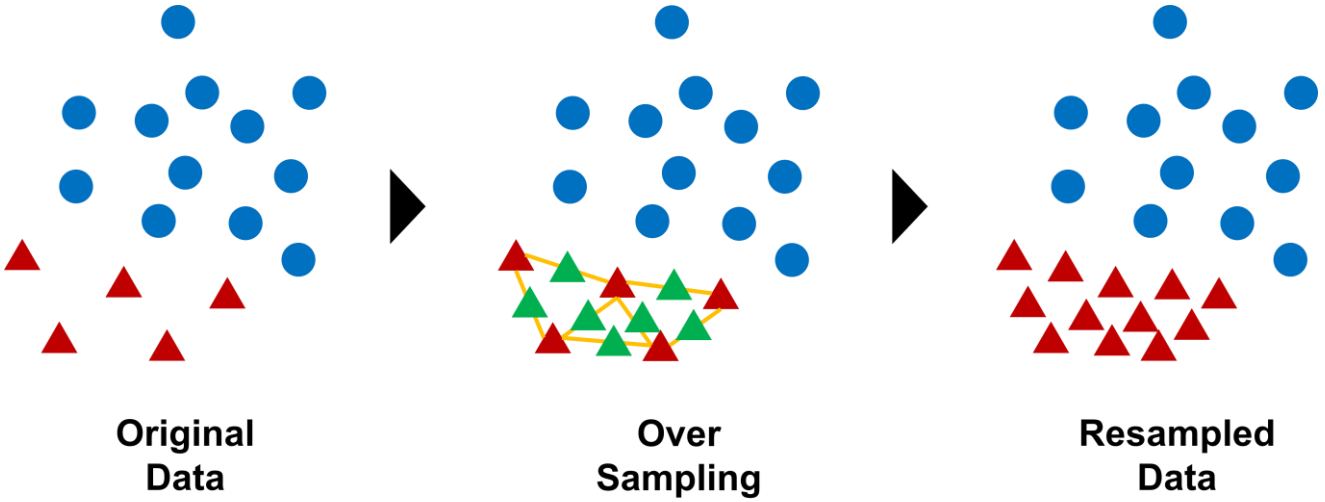
Object detection 에서 foreground와 background클래스 사이 극단적인 불균형이 있을 때 사용

모형이 분류하기 쉬운 샘플에 대해서는 cost를 낮추며, 분류하기 어려운 샘플에 대해서는 cost를 높임

02 연구 방법

GEV Activation | Cost Sensitive Learning | Over Sampling

SMOTE



가장 자주 사용되는 Synthetic minority oversampling technique(SMOTE)를 사용, (Chawla et al., 2002)

데이터 정보 손실이 없으며, 일반적으로 언더샘플링에 비해 분류 정확도가 높음

과적합 문제가 발생 또는 계산시간이 증가하며, 노이즈 또는 이상치에 민감

03 연구 결과

실험 개요

1. Sigmoid Neural Network (base line)
2. GEV활성함수만 사용한 경우
3. GEV활성함수 + Cost-Sensitive Learning (α – balanced Focal Loss)
4. GEV활성함수 + Cost-Sensitive Learning (α – balanced Focal Loss) + Over-Sampling (SMOTE)

다음 4가지 경우를 적용해보고, 결과를 비교

결과의 신뢰도를 위해,
seed를 바꿔가며 동일한 실험조건에서 30번씩 실험 후 ANOVA 및 사후 분석 (Tuckey, Games-Howell) – 5% 유의수준

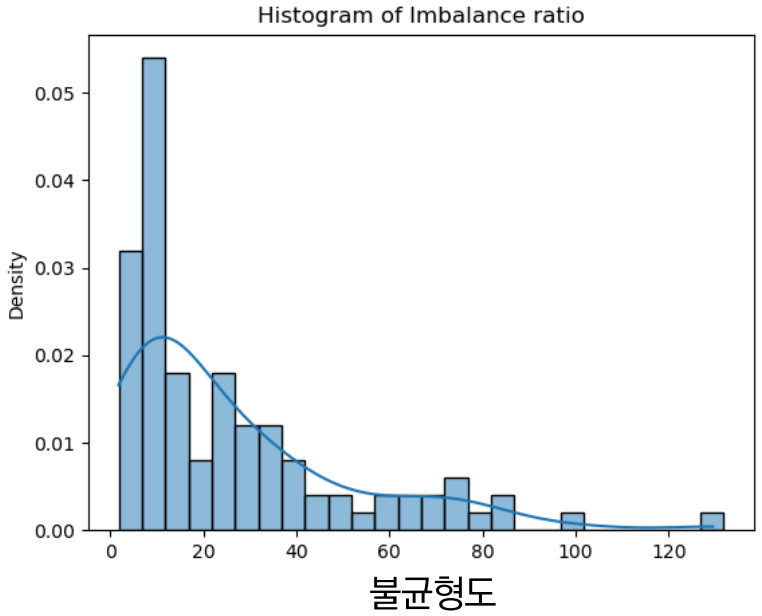
각 자료에 대해 F1-score, Geometric Mean(GM), Area Under Curve(AUC), Balanced Accuracy, Brier Inaccuracy 등

불균형 자료에 적합한 총 5가지 평가지표에 대해 평가 및 평균, 분산 비교

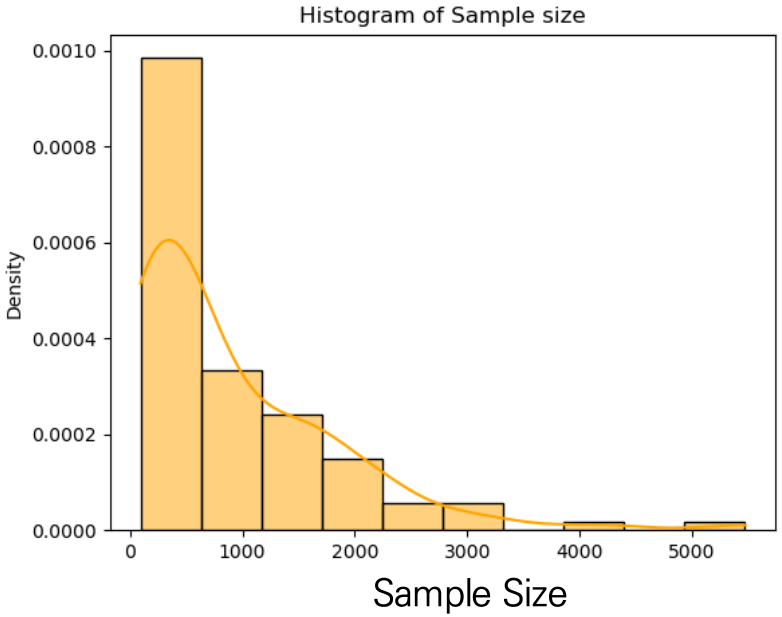
03 연구 결과

실험 개요

Data : KEEL(Knowledge Extraction based on Evolutionary Learning) imbalanced data sets (http://www.keel.es)
총 100개의 dataset을 사용하여 실험
- Imbalance ratio between 1.5 and 9 (22개)
- Imbalance ratio between higher than 9 – Part 1 (22개)
- Imbalance ratio between higher than 9 – Part 2 (22개)
- Imbalance ratio between higher than 9 – Part 3 (34개)



불균형도



Sample Size

03 연구 결과

실험 Setting

Data	Sample Size	입력 변수 (X)	목표 변수(Y)	비대칭도	비고
yeast1	1484	8	(P, N)	2.46	효모 분류(NUC)
shuttle-c0-vs-c4	1829	9	(P, N)	13.87	우주왕복선 실험 (0/4)
⋮					
Winequality-red-4	1599	11	(P, N)	29.17	적포도주 품질 (4)
abalone19	4174	8	(P, N)	129.44	전복 분류 (19)

$$\text{비대칭도 } (\rho) = \frac{\max_i\{|C_i|\}}{\min_i\{|C_i|\}}$$

$|C_i|$ = i번째 클래스의 sample 수

하이퍼 파라미터 (Hyper parameters)

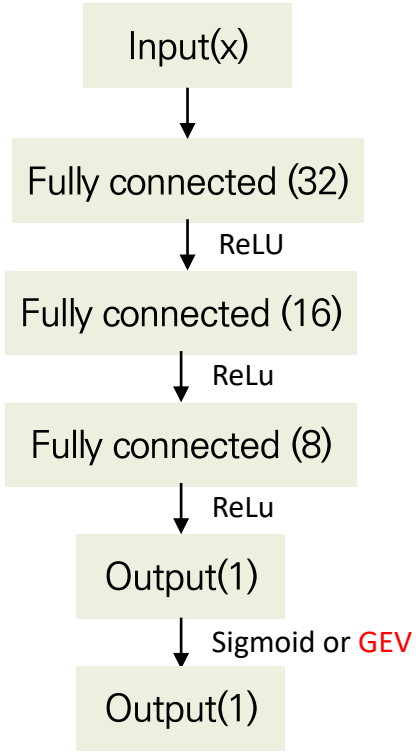
Train , Test = 7:3 비율로 나눔, 나눠진 Train에서 8:2 로 Validation 분할 (56:14:30)

Scaling = MinMaxScaling, batch size = 32, Loss = Binary Cross Entropy (FocalLoss),
Optimizer = Adam, Learning rate = 0.001, epoch = 2000 (Early Stopping 적용, patience = 20)

OverSampling = SMOTE, sampling ratio = [10:1, 20:1, 30:1 ~ 70:1]

03 연구 결과

모델 구조



모델 평가

- F1-score = $\frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}$

  [0(worst) ~ 1(best)]

- Geometric-Mean = $\sqrt{\text{TPR(Recall)} \times \text{TNR(Selectivity)}}$

  [0(worst) ~ 1(best)]

- Balanced Accuracy = $\frac{1}{2} \times (\text{TPR} + \text{TNR})$

  [0(worst) ~ 1(best)]

- Area Under the ROC Curve(AUC) – ROC 곡선의 아래 면적

  [0.5(worst) ~ 1(best)]

- Brier Inaccuracy = $\frac{1}{N}\sum_{I=1}^{N}\sum_{j=0}^{1}\left(\hat{p}(c = j, x^i) - p(c = j, x^i)\right)^2$

  [0(best) ~ 2(worst)]

  x^i : i번째, input vector

  $C \in \{0, 1\}$, class label

  $j \in \{0, 1\}$, possible class label

  $\hat{p}$: 클래스 예측 확률

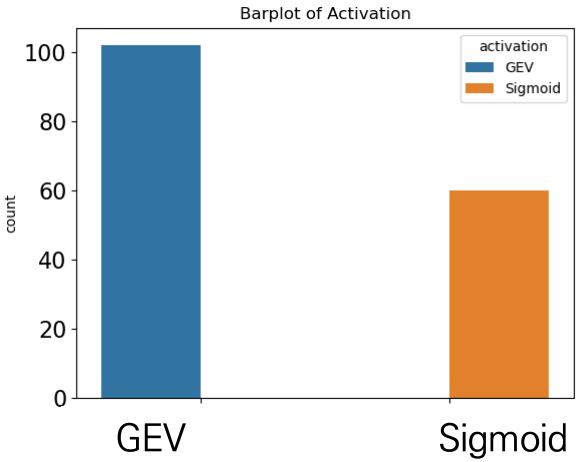
  N : 샘플수

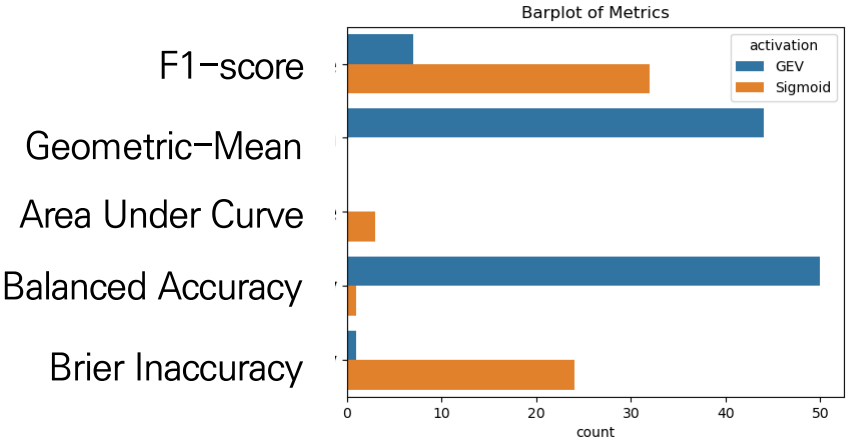
03 연구 결과

분산 분석 (ANOVA)

전체 100개의 dataset에서 5개의 평가지표에 대한 분산 분석 결과
전체의 61.8%가 5% 유의수준 하에서 유의한 차이가 존재 (309 / 500)

유의한 조합에서 사후 분석 (등 분산 : tukey HSD, 이 분산 : Games-Howell) 결과
전체의 32.4%가 5% 유의수준 하에서 유의한 차이가 존재 (162 / 500)
전체적으로 제안하는 방법이 Sigmoid보다 우수하게 나타나며(102:60), 특히, G-Mean과 Balanced-Accuracy에서 좋음



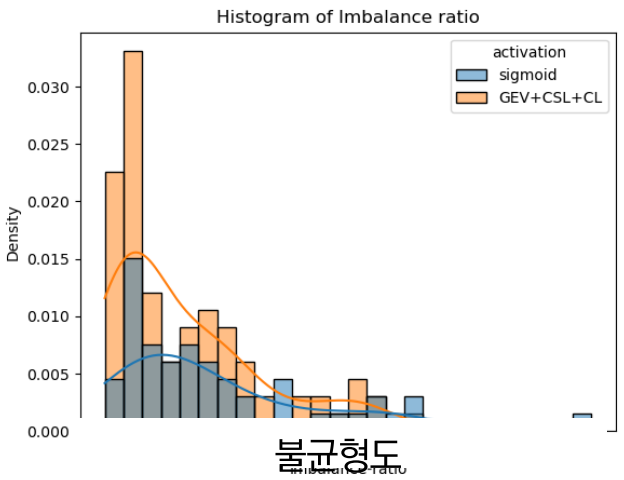


활성함수	F1-score	G-Mean	AUC	Balanced Accuracy	Brier Inaccuracy	합계
Sigmoid	32	0	3	1	24	60
GEV	7	44	0	50	1	102

03 연구 결과

최적 Over Sampling (SMOTE) 비율

8개 데이터 셋 (불균형도 70이상, 샘플사이즈 1000이상)에 대해 오버 샘플링 비율을 바꿔가며 30회 결과의 평균, 표준편차를 비교함
8개 데이터 셋에 대한 평균, 표준편차 비교 결과 불균형도가 약 20일때 제안하는 방법의 성능이 가장 좋음



예시 : abalone19 data(불균형도 : 129.43, 샘플 사이즈 : 4,174)
같은 dataset에 대한 선행 논문의 결과 (AUC : 0.7419, G-Mean : 0.7247) 보다 우수함

불균형도	F1-score	Geometric Mean	Area Under ROC Curve	Balanced Accuracy	Brier Inaccuracy
10.035	0.041(0.016)	0.729(0.15)	0.769(0.081)	0.753(0.073)	0.049(0.023)
20.157	0.044(0.015)	0.762(0.057)	0.781(0.068)	0.77(0.056)	0.039(0.008)
30.104	0.044(0.017)	0.734(0.154)	0.767(0.094)	0.757(0.083)	0.042(0.014)
40.667	0.043(0.019)	0.723(0.164)	0.751(0.126)	0.744(0.101)	0.042(0.014)
50.391	0.041(0.017)	0.722(0.157)	0.751(0.107)	0.746(0.088)	0.042(0.014)
61	0.046(0.018)	0.731(0.155)	0.767(0.099)	0.756(0.084)	0.041(0.015)
70.242	0.042(0.019)	0.722(0.16)	0.743(0.127)	0.747(0.089)	0.039(0.013)

각 지수의 평균, 괄호 안은 표준편차

Lkhagvadorj Munkhdalai, Tsendsuren Munkhdalai, Keun Ho Ryu. GEV-NN: A deep neural network architecture for class imbalance problem in binary classification. Knowledge-Based Systems. 194. 2020. p. 105534, https://doi.org/10.1016/j.knosys.2020.105534.

03 연구 결과

요약 및 결론

클래스 불균형 문제를 해결하기 위한 방안

- GEV 활성함수, Cost-Sensitive Learning (Focal Loss) 및 Over-sampling(SMOTE)을 결합한 방법을 제안

불균형도가 각기 다른 100종의 데이터셋을 사용
3가지 경우 비교 (30회 결과로 ANOVA(5% 유의수준), F1-score 등 5가지 평가지표)

1. Sigmoid (base line)
2. GEV
3. GEV + CSL
4. GEV + CSL + OS

불균형도가 70 이상이며, sample size가 1000 이상인 데이터 셋 8개에 대해 4번 방법 적용 후 최적 샘플링 비율 제시 (20:1)

F1 - score, Brier Inacurracy의 경우 대부분 Sigmoid가 우수하게 나타났으나,
Geometric-Mean, Balanced Accuracy는 제안하는 방법이 우수하게 나타남
AUC의 경우 통계적으로 유의한 차이를 거의 보이지 않음

➔ 5가지 지표를 종합으로 고려하였을 때,
제안한 방법이 기존 방법(Sigmoid) 또는 선행논문의 방법(Gumbel) 보다 우수

04 결론 및 제언

결론 및 제언

KEEL imbalance dataset을 활용한 toy model 실험을 거쳤으나, 강수량 자료, 금융 자료 등 실제 자료에 적용이 필요

평가 지표에 따라 방법의 우수성이 다르게 나타남 (단순한 갯수비교)

➔ 평가지표들의 가중평균 등 모델과 데이터의 특성을 고려한 비교 방법이 필요

$\boldsymbol{\alpha}$ – balanced Focal Loss의 경우,
$\boldsymbol{\alpha}$와 γ를 선택해야 함 (본 연구 – (α_t : 0.25, γ : 2.0))

오버샘플링의 경우 최적 비율이 20:1 정도로 나타났으나,

1. 제시한 비율 (20:1) 보다 더 불균형 해야만 Oversampling을 적용할 수 있음
2. 소수 클래스의 샘플수가 매우 작아 k-nearest neighbor도 선택하지 못한다면, 적용이 어려움
3. 선행논문(SMOTE)에 따르면, 오버샘플링과 함께 언더샘플링 적용시 결과가 더 좋음

현재, 이진 분류의 경우에만 적용하였으나, 향후 다항 분류문제로 확장할 수 있음

클래스 불균형 데이터 분류를 위한 GEV 활성 함수에 관한 연구

〈감사합니다〉

전남대학교 수학/통계학과

홍주영 (hjy_stat@naver.com)

신용관 (syg.stat@gmail.com)

박혜빈 (hyebinpark000@gmail.com)

박정수 (jspark@chonnam.ac.kr)

이 연구는 한국연구재단(NRF-2020R1I1A3069260), 교육부 및 한국연구재단의 4단계 두뇌한국21 사업 (4단계 BK21 사업, NO.5120200913674)의 지원으로 수행되었습니다.