IES 2022

GEV activation function for binary classification of class imbalance data

Hyebin Park*, Juyong Hong

Department of Mathematics and Statistics
Chonnam National University, Korea

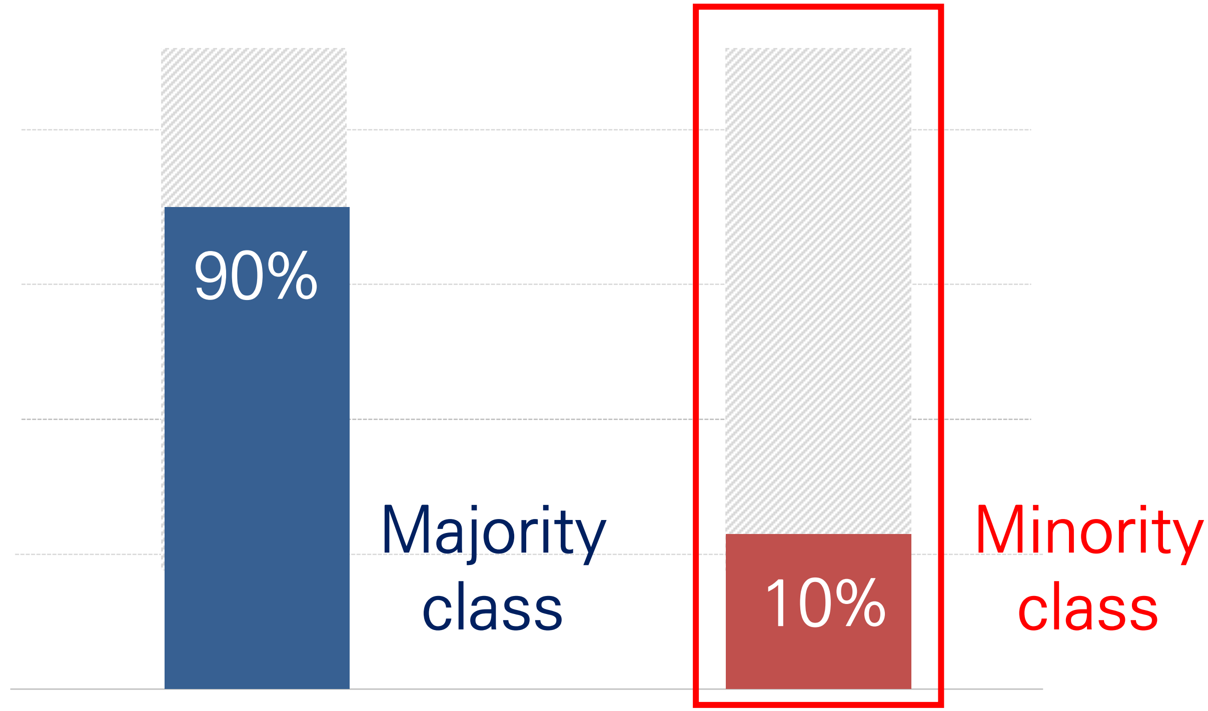
2022.01.27
Italy, capua

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

01 Introduction

imbalance in binary classification

A situation where a specific class appears at a very high frequency compared to other class



01 Introduction

Minority class is not be classified well by existing methods.
This is called an imbalance problem

➔ We need a way to classify complex imbalance data effectively

Confusion matrix

〈Existing method〉

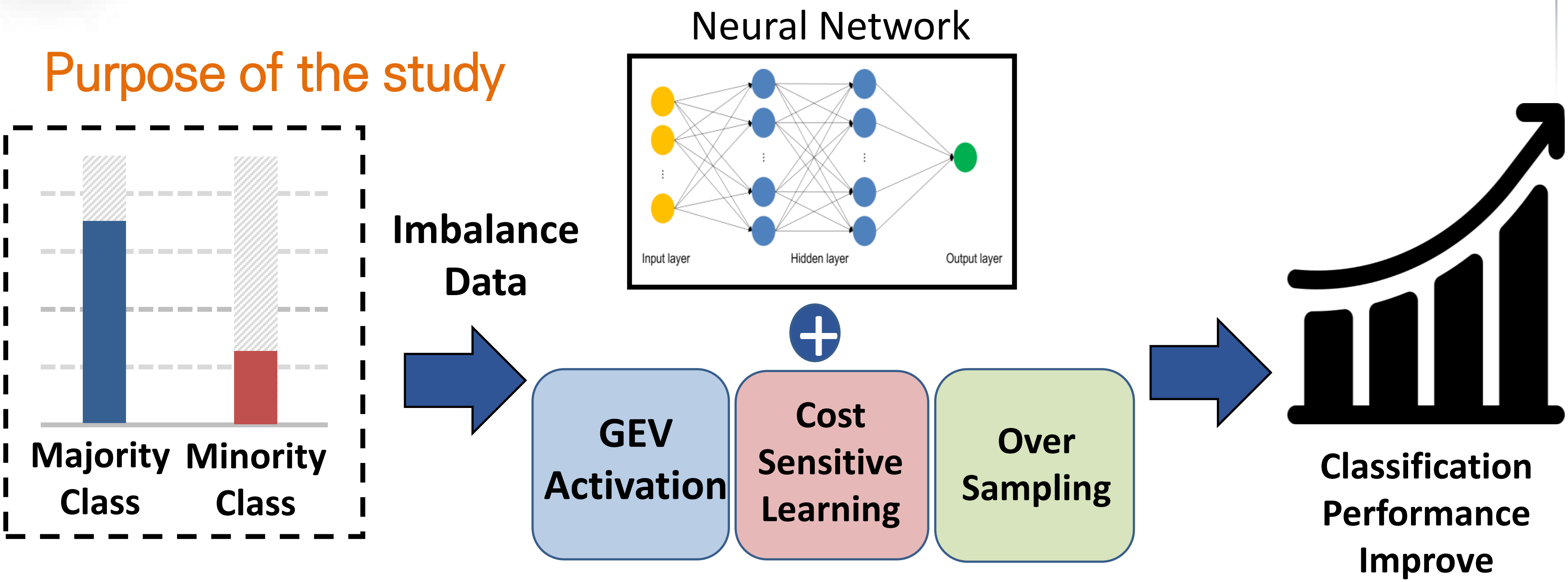
TN=1212	FP=4
FN=37	TP=0

〈proposed method〉

TN=1204	FP=12
FN=13	TP=24

01 Introduction

Purpose of the study



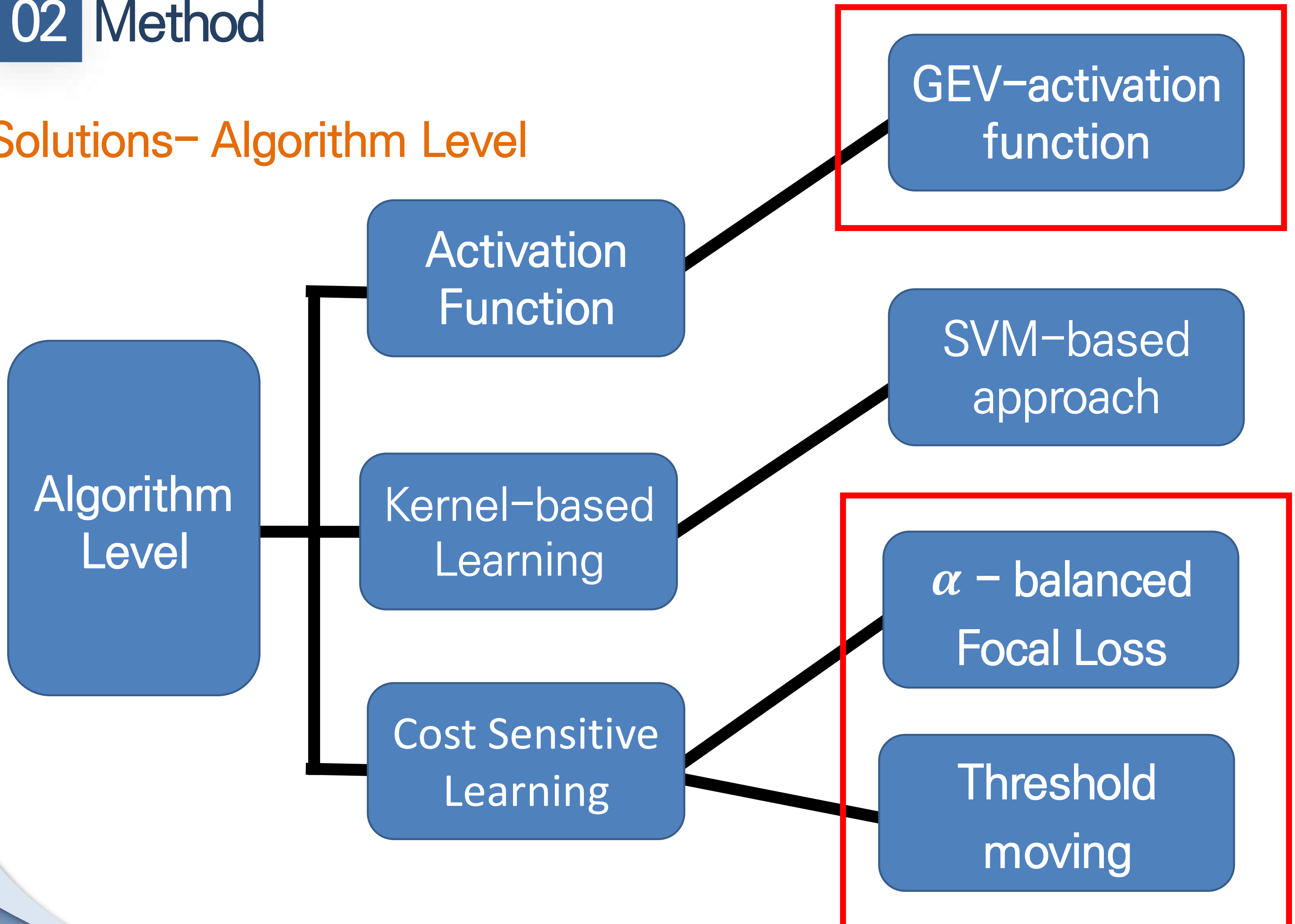
➔ To improve classification performance for class imbalance data

We combined 3 methods

(GEV-activation function, cost-sensitive learning, oversampling)

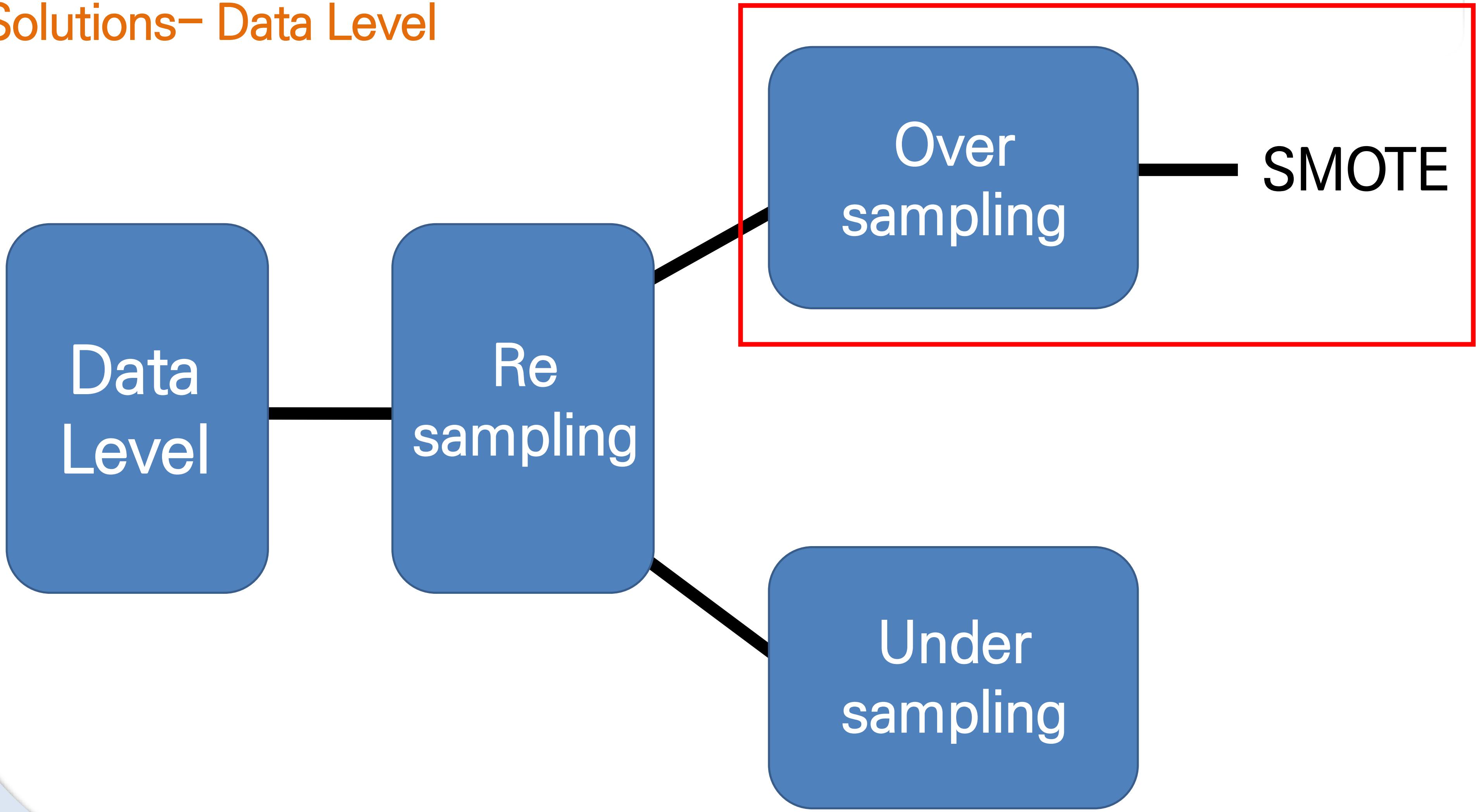
02 Method

Solutions- Algorithm Level



02 Method

Solutions- Data Level



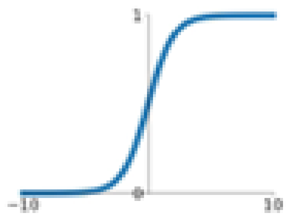
02 Method

Activation Function

We use activation function to add non-linearity to neural network models.

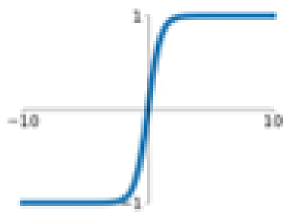
Sigmoid

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



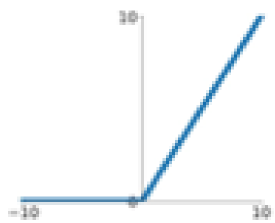
tanh

$\tanh(x)$



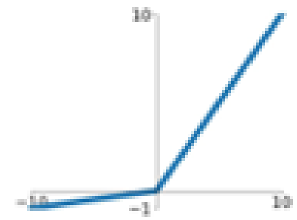
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

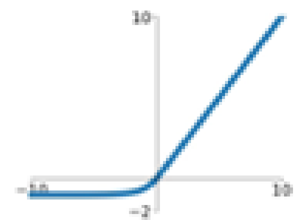


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

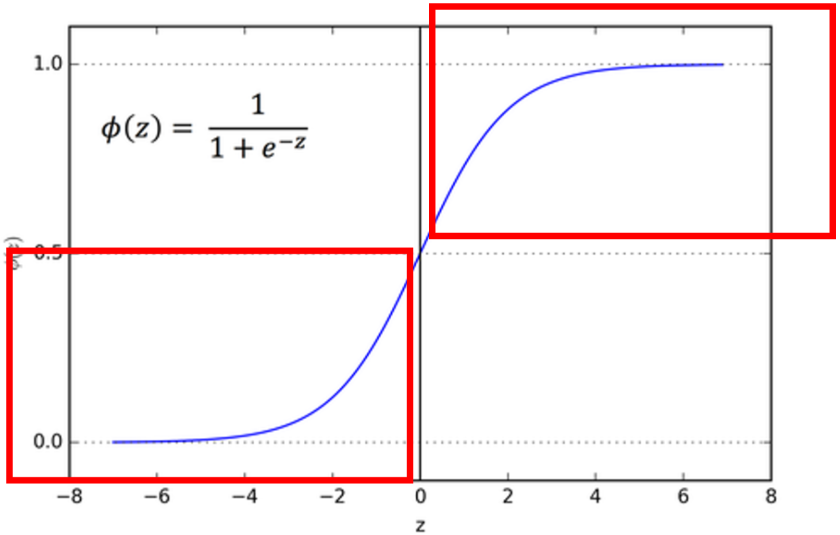
ELU

$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$



02 Method

The existing method – Sigmoid



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

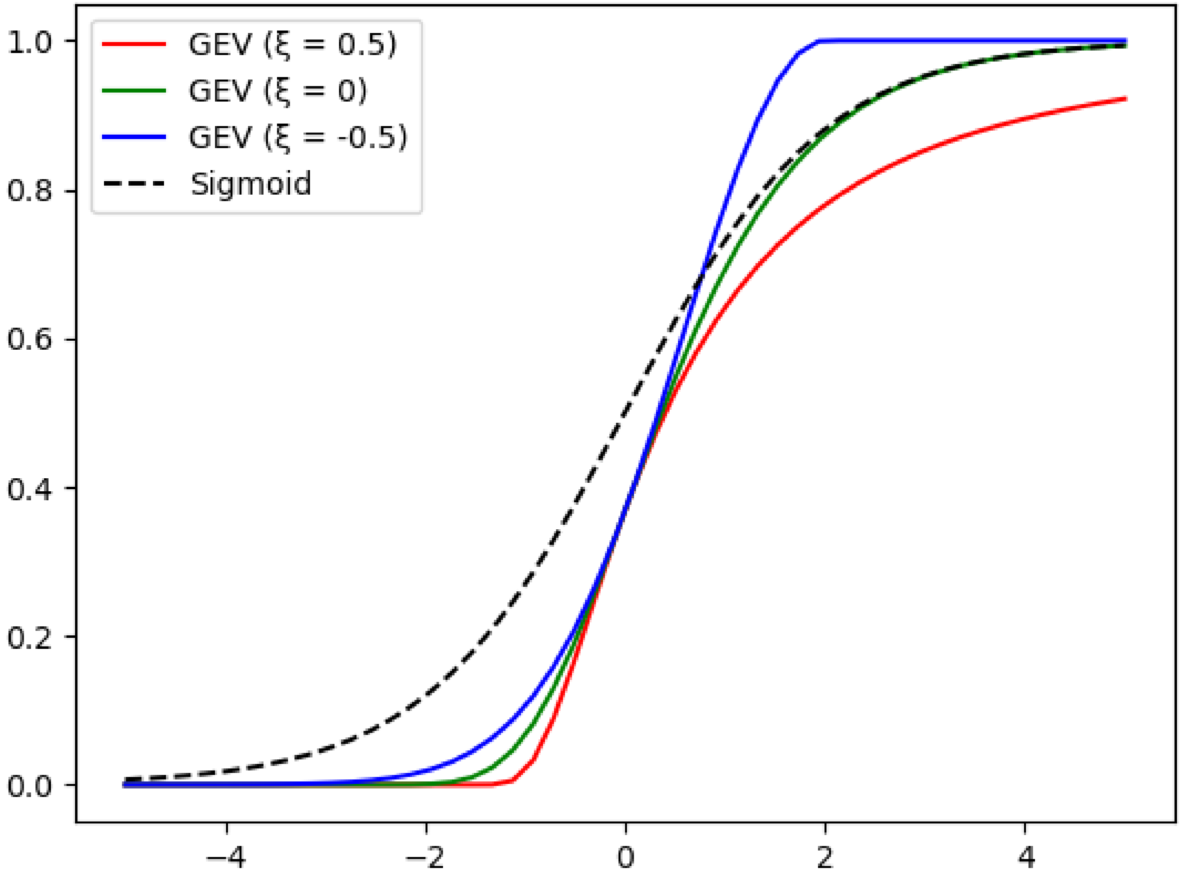
–Sigmoid returns the values between 0 and 1 from all inputs
Appropriate for binary classification

–Symmetric structure

02 Method

GEVD (Generalized extreme value distribution)

It is generally used to model extreme values



$$\mathbf{s} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

$$\boldsymbol{G(s)} = \mathbf{exp}\left\{-\left[\mathbf{1} + \boldsymbol{\xi(s)}\right]^{-\mathbf{1}/\boldsymbol{\xi}}\right\}$$

02 Method

Algorithm Level– GEV (CDF of GEVD)

- Since GEV has an asymmetric shape depending on the ξ value, it is expected to find a suitable decision threshold for imbalanced data

Use the CDF of the GEV distribution as an activation function (Bridge et al., 2020)

02 Method

Cost Sensitive Learning

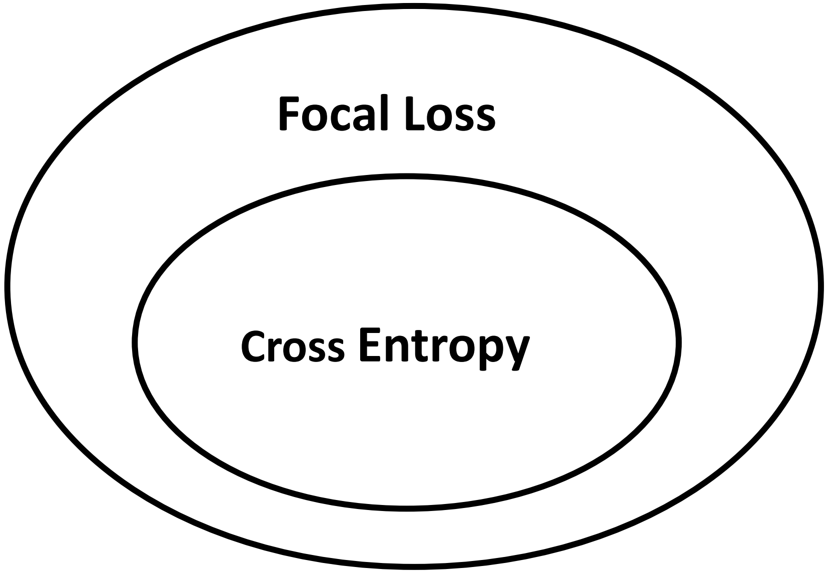
Increase the cost of misclassification for minority class
➔ Learning to minimize the total cost of misclassification

- **Threshold moving (Thresholding, Threshold Optimization)**

  Method for searching optimal threshold that minimizes misclassification costs.
  In this study, we used threshold that maximizes Geometric-Mean.

- α – balanced Focal Loss

  (Lin et al., 2017)



02 Method

α – balanced Focal Loss

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

$$p_t = \begin{cases} p & probability\ y\ belongs\ to\ minor\ class \\ 1 - p & otherwise, \end{cases}$$

α➔ Balance the importance of pos/neg example

γ ➔ Focusing parameter (〉0)
Adjusting the degree of down weighting

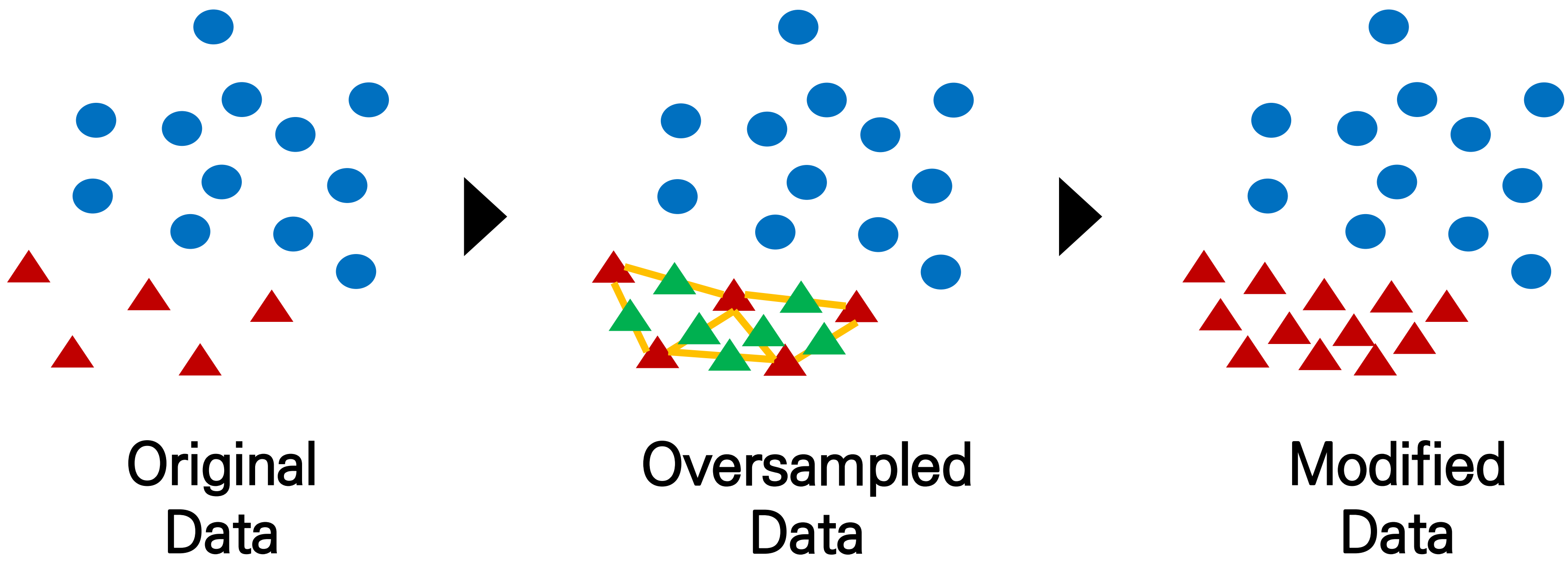
Previous study
(α_t : 0.25, γ : 2.0)

※The details can be found in (Lin et al., 2017)

02 Method

Oversampling

A method of securing sufficient data for learning
by increasing data of a minority class



02 Method

Oversampling

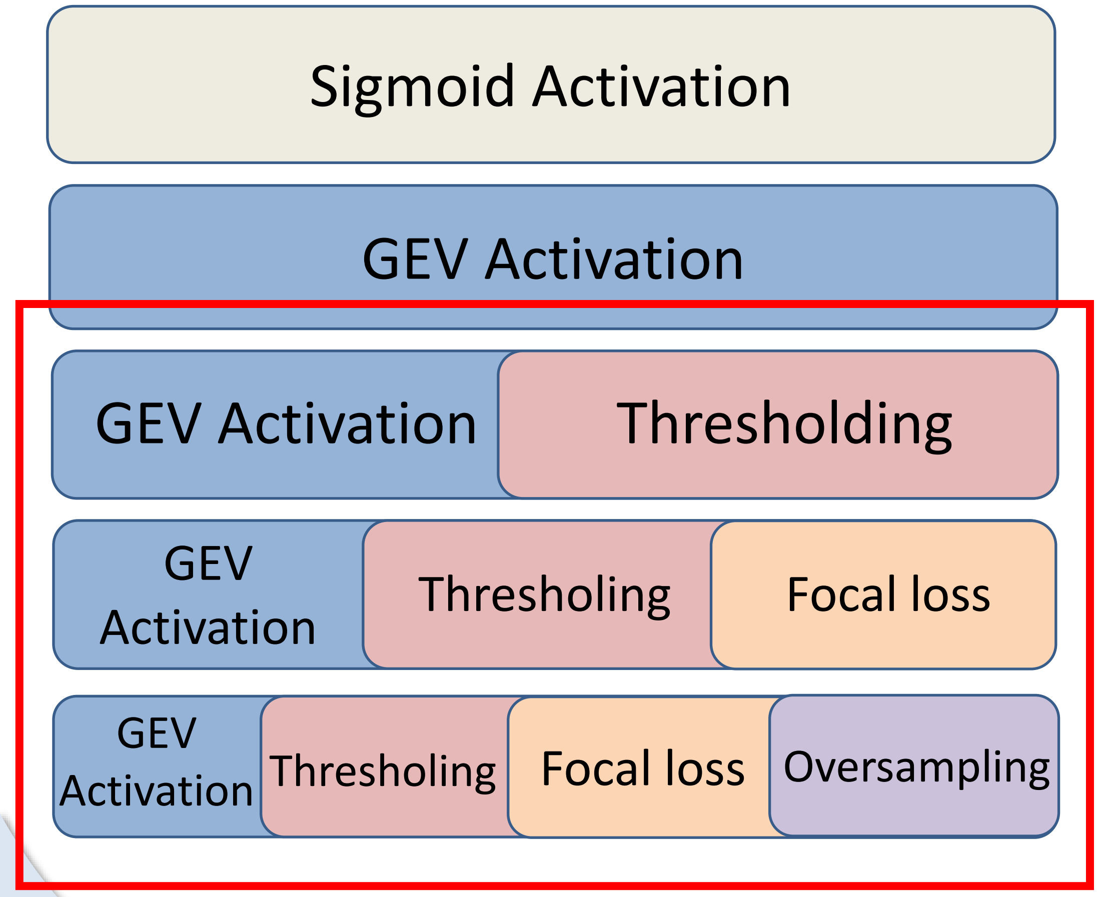
Advantage:
- does not cause data information loss

Disadvantage:
- causes overfitting problems and sensitive to noise or outliers
- adds uncertainly due to random sampling

→ We used so-called SMOTE, oversampling techniques
(Chawla et al., 2002)

02 Method

5 models for comparison



*my proposal

(1): sigmoid activaition function

(2): GEV activation function

(3): GEV + Thresholding

(4): GEV+ Thresholding + Focal loss

(5): GEV+ Thresholding
+ Focal loss + oversampling

03 Testing

Data Set

Data : KEEL(Knowledge Extraction based on Evolutionary Learning)
imbalanced data sets (http://www.keel.es)

Imbalance ratio (ρ) = $\dfrac{\max_i\{|C_i|\}}{\min_i\{|C_i|\}}$

$|C_i|$ = The number of samples of the i-th class

Using 65 datasets

Imbalance ratio range 10 to 130

Sample size range 92 to 5,472

03 Testing

Test Setting

divide each data set into the following percentages

*Validation set: used for early stopping and prevent overfitting

03 Testing

Test Setting

〈Hyper parameters〉

Hyper Parameter Type	Hyper Parameters
Data Scaling	Min Max Scaling (0 ~ 1)
Batch Size(for training)	32
Loss Function	Binary Cross Entropy $\alpha - balanced\ Focal\ Loss$
Optimizer	Adam (Adaptive Moment Estimation)
Learning Rate	0.001
Epoch	2000 Early Stopping (patience = 20)
Over-Sampling	SMOTE

Set hyperparameters for learning!

03 Testing

Model structure

Input layer (X)
Normalized input

↓

Hidden layer : 32 neurons

↓ **ReLU**

Hidden layer : 16 neurons

↓ **ReLU**

Hidden layer : 8 neurons

↓ **ReLU**

Hidden layer : 1 neurons

↓ **Sigmoid** or **GEV**

Output layer ($\hat{y}$)
Predicted Class

The prediction Network is consist of 4 hidden layers.

We used ReLU(Rectified Linear Unit) AF between hidden layers.

On the end of the prediction Network,
We used Sigmoid AF in Experiment (1)
We used GEV AF in Experiments (2) ~ (5)

* AF : activation function

03 Testing

Model evaluation measures

<The bigger, the better>

- 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) [0.5(worst) ~ 1(best)]
- 2−Brier Inaccuracy = $2 - \left(\frac{1}{N}\sum_{I=1}^{N}\sum_{j=0}^{1}\left(\hat{p}\left(c=j,x^{i}\right) - p\left(c=j,x^{i}\right)\right)^{2}\right)$ [0(worst) ~ 2(best)]

x^{i} : i−th, input vector
$C \in \{0, 1\}$, class label
$j \in \{0, 1\}$, possible class label
$\hat{p}$: Class prediction probability
N : number of sample

03 Testing

Test result (Average and standard deviation)

example : "poker data" (imbalance ratio : 85.88, sample size : 1,477)

	Type of experiment	F1-score	Geometric Mean	Area Under ROC Curve	Balanced Accuracy	2-Brier Inaccuracy
(1)	Sigmoid	0.86 (0.314)	0.868 (0.311)	0.94 (0.181)	0.923 (0.165)	1.996 (0.007)
(2)	GEV	0.6 (0.401)	0.669 (0.421)	0.908 (0.184)	0.809 (0.223)	1.989 (0.009)
(3)	GEV + FL	0.137 (0.246)	0.961 (0.151)	0.908 (0.184)	0.92 (0.145)	1.989 (0.009)
(4)	GEV + FL + TH	0.291 (0.272)	0.992 (0.042)	0.993 (0.036)	0.993 (0.037)	1.994 (0.014)
(5)	GEV + FL + TH + OS	0.4 (0.237)	0.998 (0.011)	0.999 (0.005)	0.998 (0.011)	1.997 (0.006)

Orange: best result
Yellow: second best result

03 Testing

Test result (Average and standard deviation)

example : "abalone19 data" (imbalance ratio : 129.43, sample size : 4,174)

	Type of experiment	F1-score	Geometric Mean	Area Under ROC Curve	Balanced Accuracy	2-Brier Inaccuracy
(1)	Sigmoid	0.0 (0.0)	0.0 (0.0)	0.794 (0.084)	0.5 (0.0)	1.984 (0.0)
(2)	GEV	0.0 (0.0)	0.0 (0.0)	0.659 (0.143)	0.5 (0.0)	1.81 (0.013)
(3)	GEV + FL	0.033 (0.021)	0.662 (0.16)	0.659 (0.143)	0.687 (0.105)	0.81 (0.013)
(4)	GEV + FL + TH	0.045 (0.023)	0.733 (0.165)	0.76 (0.119)	0.757 (0.095)	1.963 (0.012)
(5)	GEV + FL + TH + OS	0.044 (0.015)	0.762 (0.057)	0.781 (0.068)	0.770 (0.056)	1.961 (0.008)

Orange: best result
Yellow: second best result

03 Testing

<50 data sets - Counting the best result, second best result>

imbalance ratio < 50

	F1-score	Geometric Mean	AUC	Balanced Accuracy	Brier Inaccuracy	Total
(1)	**37**	7	34	7	36	121
(2)	18	22	17	20	**38**	115
(3)	6	22	17	20	**38**	103
(4)	22	**61**	55	**62**	24	**224**
(5)	28	56	**49**	59	11	203

– evauation results of (4) and (5) are good

03 Testing

<15 data sets - Counting the best result, second best result>

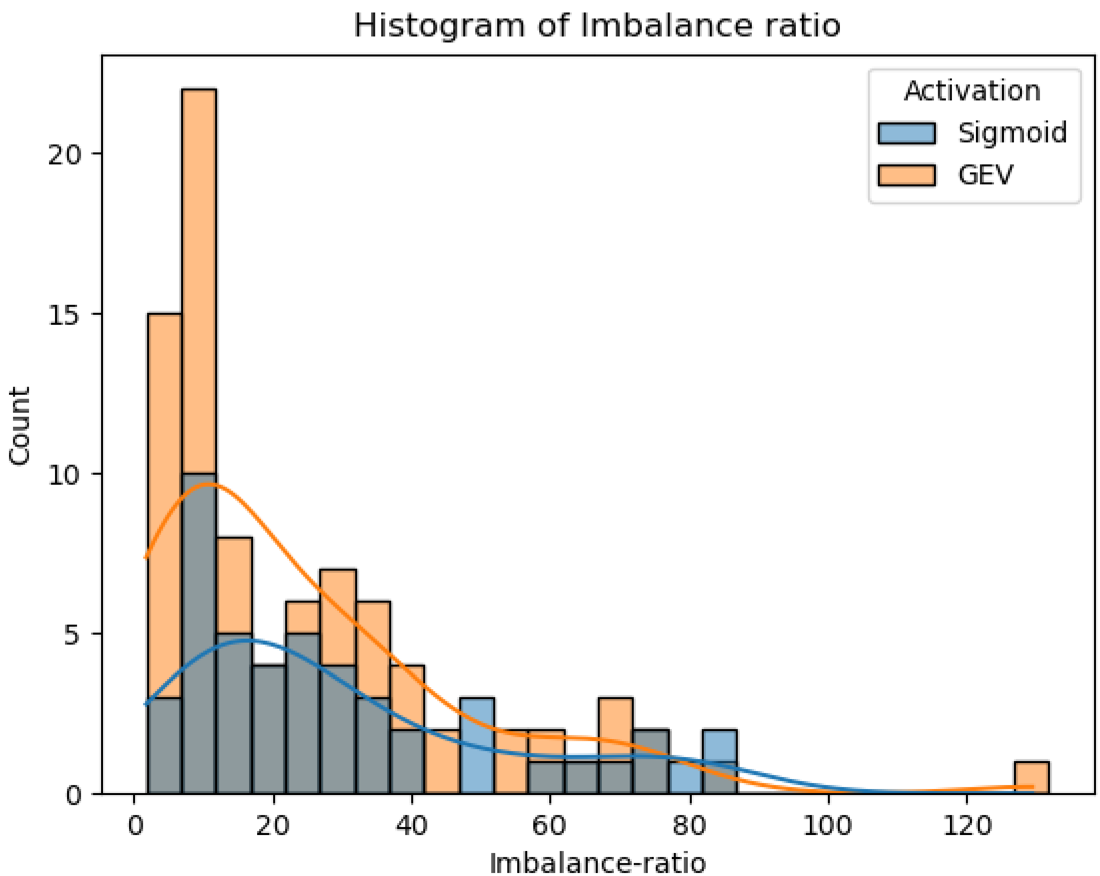
imbalance ratio >= 50

	F1-score	Geometric Mean	AUC	Balanced Accuracy	Brier Inaccuracy	Total
(1)	**9**	2	9	2	12	34
(2)	4	12	9	13	**14**	52
(3)	2	12	9	13	**14**	50
(4)	7	13	**12**	13	10	55
(5)	8	**17**	**12**	**18**	6	**61**

– evaluation results of (5) is good

03 Testing

Test result



This graph shows which of the Sigmoid and GEV(proposed method) work better for the same imbalance ratio

04 Conclusion and Limitations

Conclusion

To address the class imbalance problem, we suggested a method for better classification

Combining the GEV activation function, Cost-Sensitive Learning (α-balanced Focal Loss, Threshold moving), and Over-sampling (SMOTE)

As a result, the proposed combination method classified imbalance data better than the existing single usage method.

04 Conclusion and Limitations

Limitations

- The proposed method takes much time than the sigmoid method.

- According to the previous paper (SMOTE), the result are better when oversampling and undersampling are applied together.

Finally, the proposed methods in this study were applied
only to binary classification,
but it can be extended to multinomial classification in the future

GEV activation function for binary classification of class imbalance data

Thank you

Hyebin Park
hyebinpark000@gmail.com

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No.2020R1I1A3069260) and BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).