Yulin HE, Xuan LU, Philippe FOURNIER-VIGER, Joshua Zhexue HUANG, 2024. A novel overlapping minimization SMOTE algorithm for imbalanced classification. *Frontiers of Information Technology & Electronic Engineering*, 25(9):1266-1281. https://doi.org/10.1631/FITEE.2300278

A novel overlapping minimization SMOTE algorithm for imbalanced classification

Key words: Imbalanced classification; Synthetic minority oversampling technique (SMOTE); Majority-class sample point; Minority-class sample point; Generalization capability; Overlapping minimization

Corresponding author: Yulin HE
E-mail: yulinhe@gml.ac.cn
ORCID: https://orcid.org/0000-0002-3415-0686

Motivation

- Data imbalance is a common issue in machine learning and data mining. Researchers try to rebalance the dataset by sampling techniques. However, existing methods are suffering from the following problems, generating samples in the overlapping region and aggravating intra-class imbalance, making the classifiers more difficult to identify the classification boundaries.

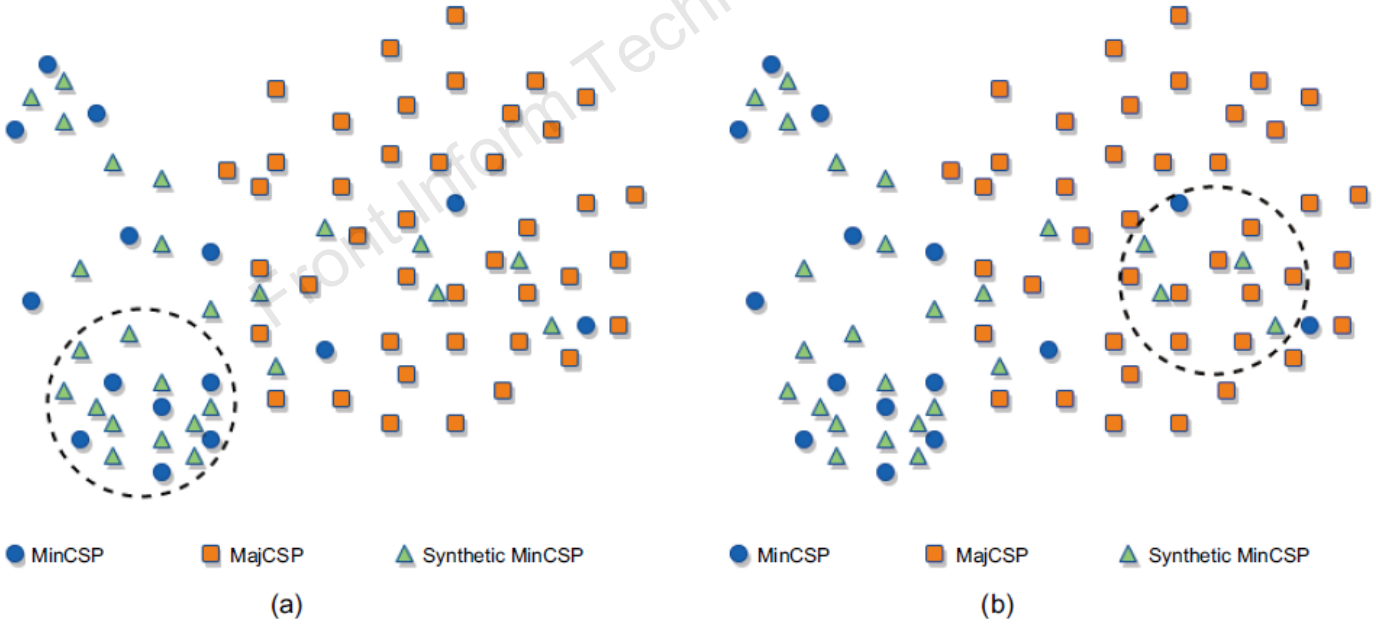


Fig. 1 Negative effects of existing synthetic minority oversampling techniques: (a) aggravating intra-class imbalance; (b) generating samples in overlapping regions

Main idea

- To address the overlapping minimization synthetic minority oversampling technique (OM-SMOTE) algorithm, which includes a method for mapping data into a more separable hidden space, named overlapping alleviation transformation (OAT).
- To further make sure that the synthetic samples would not fall in the overlapping region, we propose the retreating interpolation (RI) mechanism. To be specific, we design four kinds of interpolation rules based on four different location relationships of two selected samples.

Method

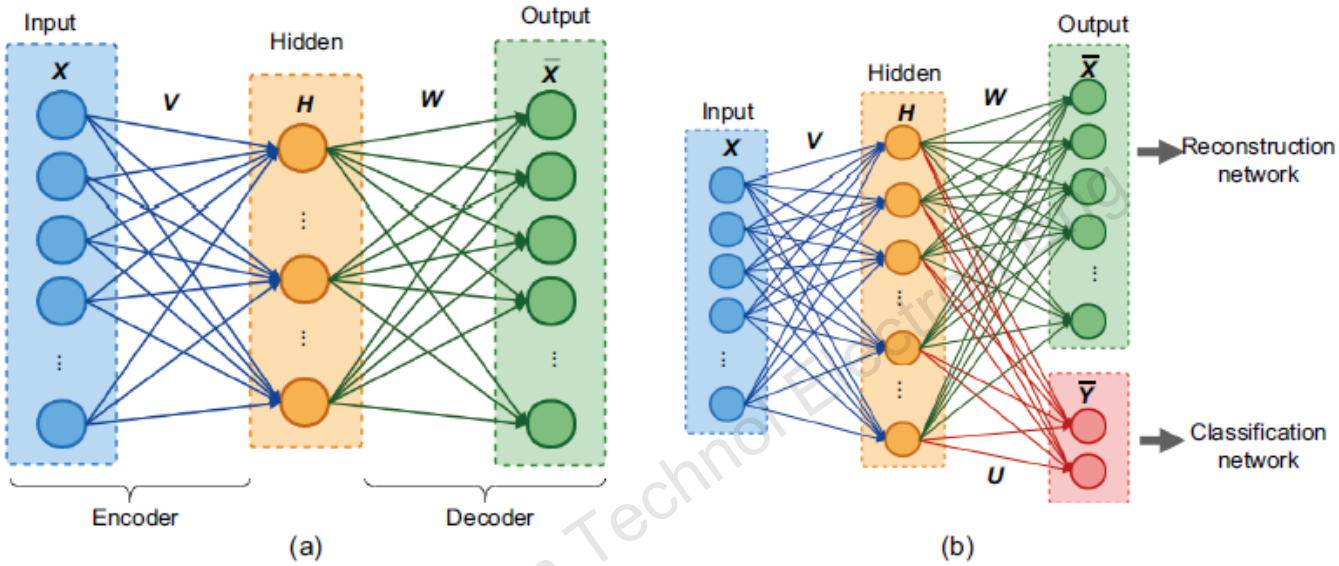


Fig. 2 Network structures corresponding to general auto-encoder and overlapping alleviation transformation methods: (a) general auto-encoder structure; (b) auto-encoder structure in an overlapping alleviation transformation method

Reconstruction network makes sure that the output is as similar as the input.
Classification network aims to guide the mapping process, which makes samples more separable in the hidden space.

Learning process can be describe as

$$\begin{cases} \boldsymbol{H}_{\mathcal{N}\times\mathcal{L}} = s(\boldsymbol{X}_{\mathcal{N}\times\mathcal{D}}\boldsymbol{V}_{\mathcal{D}\times\mathcal{L}}), \\ \overline{\boldsymbol{X}}_{\mathcal{N}\times\mathcal{D}} = s(\boldsymbol{H}_{\mathcal{N}\times\mathcal{L}}\boldsymbol{W}_{\mathcal{L}\times\mathcal{D}}), \\ \overline{\boldsymbol{Y}}_{\mathcal{N}\times 2} = s(\boldsymbol{H}_{\mathcal{N}\times\mathcal{L}}\boldsymbol{U}_{\mathcal{L}\times 2}), \end{cases}$$

Method

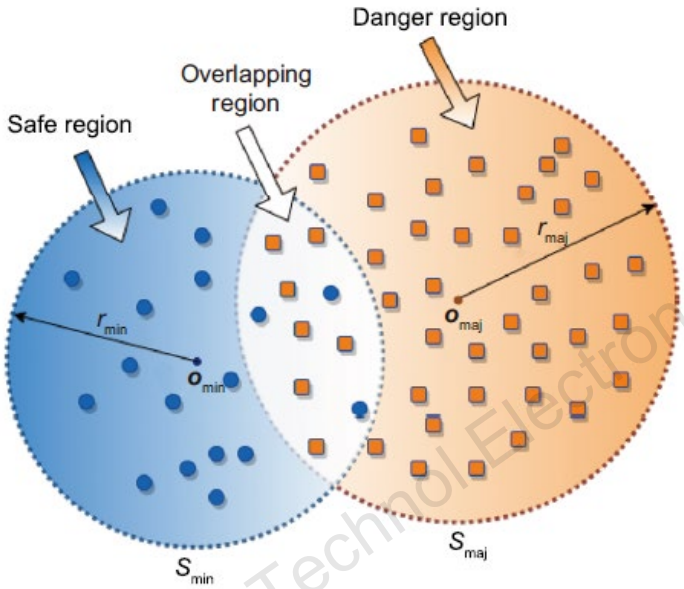


Fig. 3 Hyperspheres corresponding to MajCSPs and MinCSPs

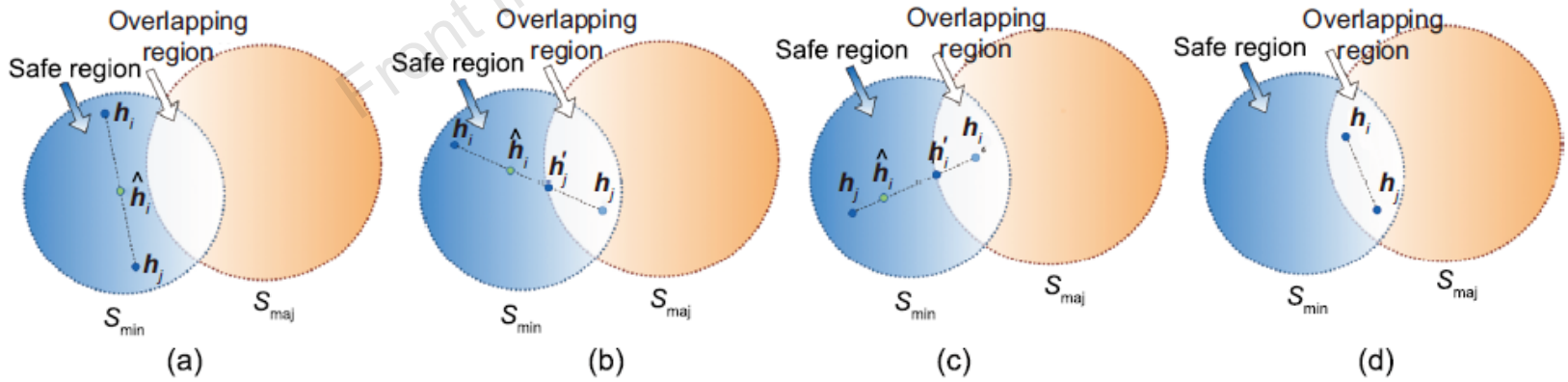


Fig. 4 Imputation rules of the retreating interpolation method: (a) direct interpolation; (b) auxiliary sample retreating interpolation; (c) seed sample retreating interpolation; (d) no interpolation

Results

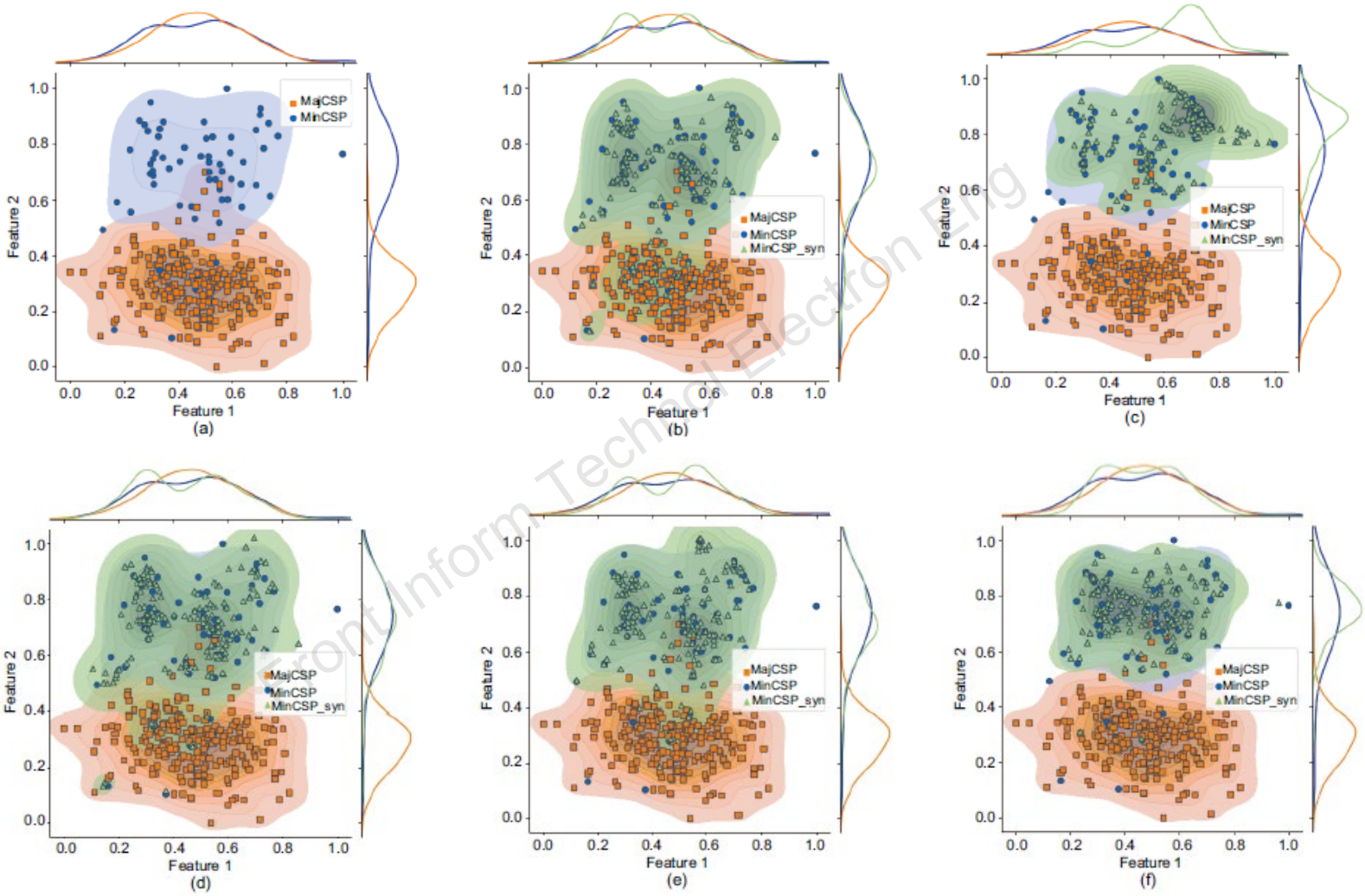
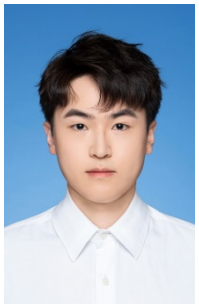


Fig. 10 Original dataset and oversampled datasets by state-of-the-art SMOTE-based algorithms with one-dimensional kernel density estimates displayed along the feature: (a) original dataset; (b) oversampled dataset by SMOTE; (c) oversampled dataset by k-means SMOTE; (d) oversampled dataset by G-SMOTE; (e) oversampled dataset by LoRAS; (f) oversampled dataset by OM-SMOTE

Conclusions

In this paper, we proposed a novel OM-SMOTE algorithm for imbalanced classification problems. OM-SMOTE is applied in two steps. First, an OAT method transforms the original data into a more separable space, and then an RI mechanism is used to avoid generating new data in the overlapping region.

Yulin HE is currently a researcher associated with Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ). His main research interests include big data approximate computing technologies, multi-sample statistical analysis theories and methods, and data mining/machine learning algorithms.



Xuan LU is currently pursuing his PhD at Shanghai Jiao Tong University. He obtained his BS and MS degrees from Shenzhen University in 2021 and 2024, respectively. His research interests include machine learning and artificial intelligence.

Philippe FOURNIER-VIGER is a distinguished professor at the College of Computer Science and Software Engineering at Shenzhen University, China. He obtained a title of national talent from the National Natural Science Foundation of China. He has published more than 300 research papers related to data mining, big data, and intelligent systems and applications, which have received more than 10 000 citations.

Joshua Zhexue HUANG received his PhD degree from The Royal Institute of Technology, Stockholm, Sweden, in 1993. He is currently a distinguished professor with the College of Computer Science and Software Engineering, Shenzhen University, China. He is also the director of Big Data Institute, China, and the deputy director of the National Engineering Laboratory for Big Data System Computing Technology.