E-SMOTE: Embedding-based Oversampling with Group Awareness for Fraud Detection

Fariza Urakhova[1], Tserenpurev Chuluunsaikhan[2], Jeong-Hun Kim[3,*], Aziz Nasridinov[1,*]

[1]Department of Computer Science, Chungbuk National University, Cheongju, South Korea
[2]Bigdata Research Institute, Chungbuk National University, Cheongju, South Korea
[3]Department of Computer Science and Information Engineering, Kunsan National University, Gunsan, South Korea
{fariza, teo}@chungbuk.ac.kr, kimjh@kunsan.ac.kr, aziz@chungbuk.ac.kr

Abstract

Fraud detection is one of the crucial areas in today's digital age. However, severe class imbalance often challenges it, leading to biased models struggling to detect fraud effectively. This paper proposes Embedding-based SMOTE, a subgroup-aware oversampling method based on representation learning with an Autoencoder and Contrastive Learning. To preserve subgroup structure, our method applies SMOTE within each minority subgroup (safe, borderline, rare, outlier), using increase coefficients specific to each group. Experiments show that E-SMOTE achieves competitive and balanced results across classifiers, even when more difficult samples are introduced.

Introduction

Fraud detection is critical in modern financial systems. While machine learning and deep learning are widely applied to solve fraud issues, class imbalance remains a key challenge. Fraud cases are rare, often leading models to favor the majority class and miss fraudulent transactions.

To solve the class imbalance problem, many oversampling methods have been proposed. However, most of them do not consider the internal structure of the minority class. For example, SMOTE (Chawla et al. 2002) generates synthetic samples between random minority instances, mainly in dense regions, leading to an intra-class imbalance problem. Recent methods such as AWTDO (Wang et al. 2022) and DDSC-SMOTE (Li et al. 2024), overcome this by clustering the minority class and generating synthetic data within each sub-cluster. While this improves data diversity, both algorithms have two limitations: (1) the number and shape of sub-clusters is arbitrary and may not reflect the actual semantic structure of minority class; (2) outlier samples are often removed at an early stage, which may lead to the loss of rare but important samples.

To address these limitations, this paper proposes E-SMOTE, a subgroup-aware oversampling method that operates in a structured embedding space. Unlike other methods, our proposed method characterizes each minority instance based on its local neighborhood in the embedding space. Each instance is assigned to one of four predefined subgroups: safe, borderline, rare, or outlier. We use an Autoencoder combined with Contrastive Learning to obtain more meaningful subgroup separation. This helps organize the minority class in a better-structured way. Most importantly, E-SMOTE retains all subgroup types, including difficult cases such as rare and outliers. By applying proportional oversampling, it maintains the original distribution among subgroups. This ensures better coverage of complex minority instances, which are often overlooked by existing methods.

Proposed Method

Figure 1 presents the overview of the proposed method. The proposed method consists of four steps as follows:

Mapping into embedding space. We first train an Autoencoder to map input data into an embedding space and apply Contrastive Learning on minority class embeddings using dropout and noise-based augmentations. This improves the embedding space by clustering similar minority samples and separating those different.

Clustering minority class samples. Following Napierala et al. (2015), we group minority samples into safe, borderline, rare, and outliers, based on the number of minority-class neighbors among the 5 nearest neighbors. Safe samples have only minority neighbors, borderline have mixed neighbors, rare have one minority neighbor, and outliers have none.

*The corresponding authors.

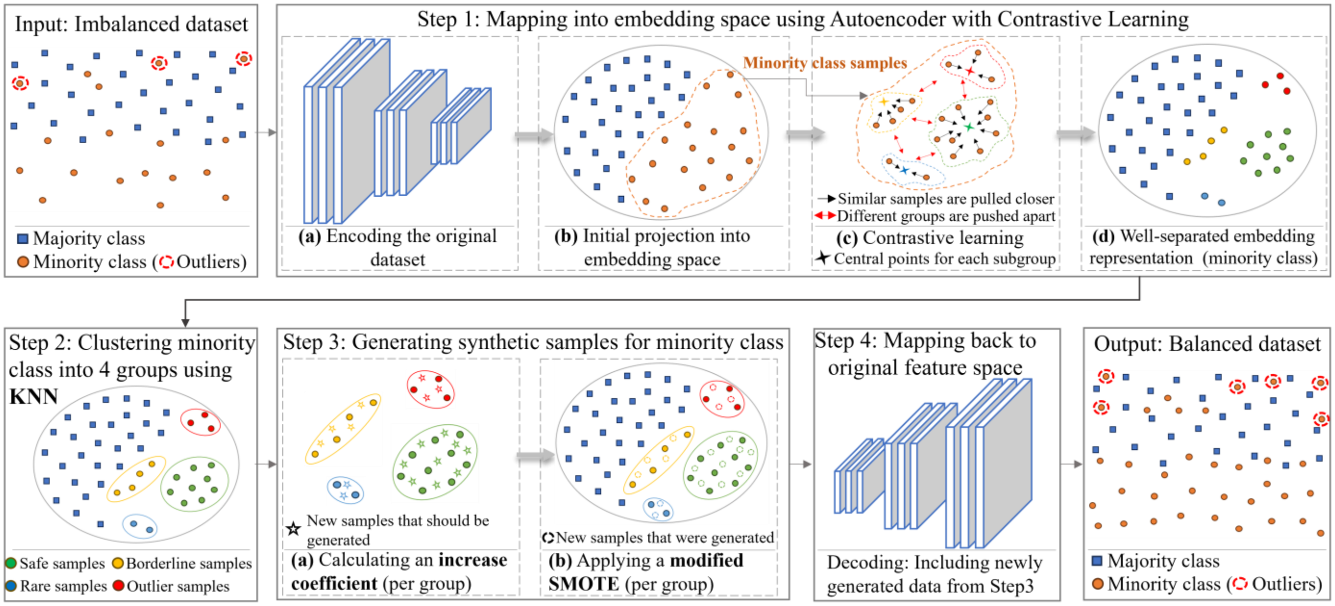

Figure 1: The proposed method consists of four main steps: (1) embedding via Autoencoder with Contrastive Learning, (2) subgroup clustering, (3) group-aware oversampling, and (4) projecting the balanced dataset back to the original feature space.

Generating synthetic samples. We perform group-aware oversampling through the following two steps:
(a) To preserve the internal distribution of the minority class, we assign each subgroup an increase coefficient based on its proportion in the original data. This increase coefficient determines how many synthetic samples should be generated for each subgroup, ensuring balanced representation.
(b) We then apply SMOTE separately within each minority subgroup in the embedding space, generating new samples using only intra-group neighbors.

Mapping back to the original feature space. The decoder reconstructs original and synthetic embeddings into the original feature space, producing a balanced dataset. This final dataset is then used to train classifiers for detecting fraud.

Performing Evaluation

Results. We evaluate our method on the public Credit Card Fraud dataset from Kaggle, with a 1:578 class imbalance ratio. The data is split 80:20, and oversampling is applied only to the training set to ensure a realistic test scenario. Table 1 compares the performance of three classifiers (i.e., Random Forest, XGBoost, and Multi-Layer Perceptron) using F1-score and G-mean, standard metrics for imbalanced learning. Rather than optimizing evaluation metrics alone, our proposed method prioritizes the generation of diverse and realistic minority samples, including rare and complex fraud cases. While this may slightly impact certain scores, it leads to a more balanced and practically useful model for fraud detection tasks.

	F1-score			G-mean		
	RF	XGB	MLP	RF	XGB	MLP
SMOTE	84.1	80.9	76.3	91.5	94.2	91.5
B-SMOTE	84.2	85	82.7	88.6	91.5	90.9
ADASYN	82.7	76.8	76.3	89.8	90.9	91.5
E-SMOTE	86.2	84.4	83.3	90.9	90.9	91.5

Table 1: Performance comparison (F1-score/G-mean).

Conclusion and Future Work

This study introduces E-SMOTE, a subgroup-aware oversampling algorithm designed to address the class imbalance problem in fraud detection. E-SMOTE generates more realistic and informative synthetic samples by operating in a structured embedding space and preserving subgroup diversity. Unlike conventional oversampling methods, our proposed method preserves rare and complex fraud cases and increases their presence in training data. Ignoring or underrepresenting these cases may result in models that perform well on average but fail to detect the most critical threats in real-world scenarios. Future work will focus on evaluating E-SMOTE on larger and more diverse datasets and adapting it to other domains where detecting rare, high-risk anomalies is crucial.

References

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *The Journal of Artificial Intelligence Research*, 16: 321–357. https://doi.org/10.1613/jair.953.

Li, H.; Jin, Y.; Zhao, Y.; and Wang, Y. 2024. Data Oversampling Algorithm Based on Data Distribution and Spectral Clustering. *The Journal of Supercomputing*, 80: 17760–17789. https://doi.org/10.1007/s11227-024-06132-7.

Napierala, K.; and Stefanowski, J. 2016. Types of Minority Class Examples and Their Influence on Learning Classifiers from Imbalanced Data. *Journal of Intelligent Information Systems*, 46(3): 563–597. https://doi.org/10.1007/s10844-015-0368-1.

Wang, X.; Gong, J.; Song, Y.; and Hu, J. 2022. Adaptively Weighted Three-Way Decision Oversampling: A Cluster Imbalanced-Ratio Based Approach. *Applied Intelligence*, 53: 312–335. https://doi.org/10.1007/s10489-022-03394-7.