Addressing Class Imbalance in Credit Fraud Dataset using TabDDPM and TVAE: A Synthetic Data Generation Approach

Abstract

Class imbalance in credit fraud datasets poses challenges for accurate modelling and prediction. This report proposes a solution that utilizes the TabDDPM (Tabular Data with Diffusion Models) and TVAE (Triple Variational Autoencoder) models to generate synthetic data for the minority class (class 1) in the imbalanced dataset. The synthetic data generation process is performed using 10-fold cross-validation, followed by evaluating the performance of a Random Forest classifier on the balanced dataset. The evaluation metrics, including F1-Score, Kappa-Score, Average Precision Score, G-Mean, and Accuracy, are used to assess the classifier's performance in detecting credit fraud instances.

Introduction

Class imbalance is a common issue in credit fraud datasets, where the number of instances in the minority class is significantly lower than the majority class. This imbalance can lead to biased models and inaccurate predictions, as the classifier tends to favor the majority class. To address this problem, this report proposes the use of the TabDDPM and TVAE models to generate synthetic data for the minority class and create a balanced dataset for improved prediction accuracy.

Data Preparation

The dataset used in this experiment is the credit fraud dataset available on Kaggle. It consists of labelled instances with two classes: class 0 (non-fraud) and class 1 (fraud). The dataset contains transactions made by credit cards in September 2013 by European cardholders. The dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. The number of instances in class 1 is considerably lower, indicating a severe class imbalance. This report focuses on addressing this imbalance and improving the prediction performance for detecting credit fraud instances.

Synthetic Data Generation using TabDDPM and TVAE

The TabDDPM model employs diffusion models to generate synthetic data that closely resemble the original dataset. Similarly, the TVAE model uses the triple variational autoencoder approach to generate synthetic data. Both models can be utilized to address the class imbalance by creating synthetic samples specifically for the minority class instances. The experiment involves performing 10-fold cross-validation, ensuring that synthetic data is only generated for class 1 instances present in the training folds. This approach aims to create a balanced dataset with equal representation from both classes.

Evaluation Metrics

Once the dataset is balanced by generating sufficient synthetic data for class 1, the performance of a Random Forest classifier is evaluated. Several evaluation metrics are considered to assess the classifier's effectiveness in detecting credit fraud instances. These metrics include F1-Score, which measures the balance between precision and recall, Kappa-Score, which assesses the agreement between predicted and actual classes, Average Precision Score, which quantifies the classifier's ability to rank instances correctly, G-Mean,

which evaluates the classifier's performance on both classes, and Accuracy, which measures overall correctness.

Experimental Results and Analysis

After conducting the 10-fold cross-validation and evaluating the Random Forest classifier on the balanced dataset, the results obtained from the evaluation metrics are analyzed. The F1-Score provides insights into the classifier's ability to balance precision and recall, while the Kappa-Score indicates the agreement between predicted and actual classes. The Average Precision Score reflects the classifier's ranking capability, and the G-Mean evaluates its performance on both classes. Additionally, the Accuracy metric measures the overall correctness of the predictions. The analysis of these metrics provides valuable insights into the performance and effectiveness of the classifier in detecting credit fraud instances.

Conclusion

This report presented a comprehensive approach to address class imbalance in the credit fraud dataset using the TabDDPM and TVAE models for synthetic data generation and the Random Forest classifier for prediction. By generating synthetic data for the minority class, the dataset was balanced, leading to improved prediction accuracy. The evaluation metrics provided a robust assessment of the classifier's performance in detecting credit fraud instances. This approach can assist in developing more accurate and reliable fraud detection systems in the financial domain.

Future Directions

Future research can explore the utilization of other advanced synthetic data generation techniques and alternative classifiers to further enhance the performance of credit fraud detection. Additionally, investigating feature engineering methods and ensemble techniques

may lead to even better predictive models. Continuous exploration and refinement of such techniques will contribute to more effective fraud detection and prevention strategies in real-world scenarios.

References

- Credit card fraud dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- TabDDPM: https://arxiv.org/abs/2209.15421
- TVAE: https://arxiv.org/abs/1802.04403
- SDV Project: https://docs.sdv.dev/sdv/
- Colab Notebooks for TabDDPM: https://colab.research.google.com/drive/1Ug5rXd3WEI6ipTKFpD0ZDVQNNX1jv0ON?usp=drive_link
- Colab Notebook for TVAE: https://colab.research.google.com/drive/1hONiMzdPD8ve1SEqZitS9tzFJGpC0njr?usp=drive_link