Exploring GAN Variants for Balancing Imbalanced Dataset

1.Problem Statement

Class imbalance is an issue that often gets to deal with in real-world machine learning and especially in the medical domain. In the roperty of the above case, the classification models get biased towards the majority class resulting in poor detection of the minority class. This problem is very critical in medical-related procedures where the minority class sometimes stands for the severe or high-risk cases. By using conventional oversampling techniques, the problem of overfitting may occur, as these methods produce existing minority samples without adding new information. One of the limitations that the traditional methods have, Generative Adversarial Networks (GANs) have been proposed as an effective way to produce realistic synthetic data. As a result, the project targets leading the research in the area of GAN-based data augmentation for class imbalance as its main focus. The project uses Vanilla GAN and two GAN variants to synthesize samples for the minority class. After that, the performance of the resulting samples on classification is measured and compared.

2.Dataset Description and Imbalance Analysis

The study relied on the Heart Failure Clinical Records Dataset. It comprises clinical information taken from patients diagnosed with heart failure and has been a staple dataset for survival prediction tasks.

2.1 Dataset Characteristics

• Number of samples: 299

• Number of features: 12 clinical attributes

• Target variable: DEATH_EVENT

0: Patient survived (majority class)

1: Patient died (minority class)

The dataset consists of only numerical features, which makes it appropriate for the generation done via GANs with fully connected neural networks.

2.2 Class Imbalance Analysis

An exploratory data analysis indicated a prominent class imbalance. The number of people surviving was notably larger than that of people dying. This discrepancy encourages the application of data augmentation methods to enhance the representation of the minority class. A bar chart representation was employed to depict the distribution of classes, thereby affirming that the dataset is oriented towards the majority class. As a result of this imbalance, it is believed that the classification models which are mapped onto the original dataset will not be able to recognize death events effectively.

To solve the imbalance issue, three GAN models were performed and at the same time their training was limited only to the minority class instances (DEATH_EVENT = 1). The rationale behind training GANs only with the minority class is to make sure that the resultant samples will succinctly elevate the representation of the minority class.

3.1 Vanilla GAN

The Vanilla GAN consists of two neural networks:

- **Generator**: Produces synthetic samples from random noise.
- **Discriminator**: Distinguishes between real and synthetic samples.

The entirety of the neural networks was set up using dense layers solely. The discriminator loss function was binary cross-entropy, whereas the generator gained the ability to create samples that the discriminator deems real. To make the training process stable, the feature values were adjusted to fall within the range of [-1, 1].

Once the training was done, the generator was put into operation to generate synthetic samples of the minority class that were then mixed up with the original dataset to make a balanced one.

3.2 Wasserstein GAN (WGAN)

The Wasserstein GAN was implemented as an advanced variant to improve training stability. Instead of a discriminator, WGAN uses a **critic** that estimates the Wasserstein distance between real and generated data distributions.

Key characteristics of WGAN include:

- Removal of the sigmoid activation in the critic output
- Use of Wasserstein loss
- Weight clipping to enforce the Lipschitz constraint.

The WGAN was trained exclusively on minority class samples, and the trained generator was used to produce synthetic data for dataset balancing.

3.3 Least Squares GAN (LSGAN)

The Least Squares GAN was implemented as a second GAN variant. LSGAN replaces the binary cross-entropy loss with a least squares loss function, which reduces gradient vanishing and improves sample quality.

The generator and discriminator were implemented using dense layers, similar to the Vanilla GAN, but optimized using mean squared error loss. Synthetic minority class samples generated by LSGAN were added to the original dataset to form a balanced dataset variant.

4. Classifier Setup and Evaluation

4.1 Classifier Architecture

A Multilayer Perceptron (MLP) classifier was used to evaluate the impact of GAN-based data augmentation. The classifier architecture was kept identical across all experiments to ensure fair comparison.

The MLP consists of:

- Two hidden layers with ReLU activation
- One output layer with sigmoid activation
- Binary cross-entropy loss
- Adam optimizer

4.2 Evaluation Scenarios

The classifier was trained and evaluated under four scenarios:

1. Original imbalanced dataset
2. Dataset balanced using Vanilla GAN
3. Dataset balanced using WGAN
4. Dataset balanced using LSGAN

For each scenario, the dataset was split into training and testing sets using stratified sampling. Feature standardization was applied prior to training.

4.3 Evaluation Metrics

Classifier performance was evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

These metrics provide a comprehensive evaluation, particularly recall and F1-score, which are critical for assessing minority class detection.

5. Results and Comparisons

The results of the experiments indicate the application of GAN-based data augmentation as the cause of improvement in the classification performance that is quite evident.

•The classifier trained on the initial dataset that was largely imbalanced managed to reach a high accuracy but an extremely low recall for the minority class.

•The process of balancing the dataset using Vanilla GAN led to significant improvements in both recall and F1-score.

•WGAN was the one that showed the best overall performance across all metrics, especially recall which is a clear sign of high accuracy in the detection of the minority class.

•LSGAN was able to perform better than the case of the imbalanced dataset but still it could not reach WGAN's performance level.

The performance comparisons were shown in the form of tables and bar plots which clearly demonstrated the improvements for all the evaluation metrics.

6. Observations and Conclusions

This research indicates that using GANs for data augmentation is a good way to solve the problem of class imbalance in medical datasets. Training GANs on samples from only the minority class leads to the production of real-like synthetic data that will, in turn, improve classifier performance.

Out of the GANs that were tested, the WGAN model was the most stable in terms of training and also yielded the best classification results, especially in terms of recall and F1-score. This outcome underlines the need for careful selection of the GAN model for generation of tabular data.

To sum up, the use of GANs for augmentation not only represents the minority class significantly better but also results in classification that is more balanced and therefore more reliable. It is worth mentioning that the future work may involve the use of conditional GANs or larger datasets with the aim of further improving performance.