IMPROVING CREDIT CARD FRAUD DETECTION USING GENERATIVE ADVERSARIAL NETWORKS

Hao Ning[1] Jun Ying[1] and Amir H. Jafari[1]

ABSTRACT

In this study, we implement different Generative Adversarial Networks (GAN) models as an oversampling strategy to generate more artificial fraud data to improve the classification of credit card fraud. We use data distribution to compare the quality of generated data since we are dealing with tabular data. The experimental results suggest the best models are Wasserstein GAN-Gradient Penalty (WGAN-GP) and Boundary Equilibrium GAN (BEGAN). Our results show that GAN works well with tabular data and has great potential in risk management applications.

Keywords *Imbalanced datasets · Oversampling · Generative Adversarial Networks · Risk Management*

[1] Data Science Program, The George Washington University, Washington, DC, USA
Hao Ning
hning@gwmail.gwu.edu
Jun Ying
juny@gwmail.gwu.edu
Amir H. Jafari
ajafari@gwu.edu

1. INTRODUCTION

As credit card transactions have become the mainstream consumption pattern, the number of credit card frauds has increased dramatically. Credit card fraud detection systems are essential for banks and financial institutions to minimize their losses nowadays. However, credit card transactions are usually extremely imbalanced with little portion of fraud.

In this study, we implement GAN [1] to generate artificial fraud data as an oversampling strategy to help with the detection of credit card fraudulent transactions.

Unlike using GAN working with images and evaluating the generated images by eyes, we are working with tabular data. We use boxplot of the data distribution to evaluate the generated data to compare the spectrum of the data and how well they overlap with the original data.

2. BACKGROUND

GAN is composed of the generator and the discriminator, which are against each other in a zero-sum game. The generator generates images while the discriminator evaluates them. The generator improves its performance to confuse the evaluation of discriminator, while the discriminator improves its performance to distinguish generated images and real images. This is a minimax problem, which means if the generator improved, the performance of the discriminator would definitely get worse. Original GAN still has many problems, such as model collapse, unstable training, etc. With the development of GAN, more and more types of GAN have been proposed, gradually solving the shortcomings of GAN. Therefore, this study used four more recent GANs to improve the performance of vanilla GAN. Wasserstein GAN (WGAN) [2] and BEGAN [3] can stabilize the training and avoid model collapse; WGAN-GP is an optimized version of WGAN which uses gradient penalty instead of weight clipping in WGAN [4].

Credit card fraud detection has attracted much attention from academia. Researchers proposed many credit card fraud detection techniques such as neural networks, decision tree, outlier detection and convolutional neural network (CNN) [5, 6]. There are also some researchers proposed using GAN for fraud detection [7]. Hung Ba's work compares several different GANs [8]. However, there is no clear way to evaluate the quality of generated data for tabular dataset. Therefore, in this work, we provide detailed evaluation of GAN, WGAN, WGAN-GP and BEGAN using data distribution and statistical scores.

3. METHOD

We use credit card fraud detection data from ULB Machine Learning Group, which includes 284,807 transactions made by European cardholders within two days through credit cards in September 2013. There are only 492 frauds in all transactions, therefore the dataset is highly imbalanced with the positive class (frauds) accounts for just 0.172%. Due to the confidentiality and privacy issues, the original data has been transformed to numerical value by Principal component analysis (PCA) and we do not know their actual meaning. The dataset has 31 features, 28 of them are PCA components named from V1 to V28 while the other three are named as "Time", "Amount" and "Class". All of the data is numeric and there are no missing values. For feature "Time", we transform it from seconds to hours, and for feature "Amount", we perform log+1 transformation to make it close to normal distribution.

First, we split the dataset into 80% train (with 394 fraud) and 20% test sets (with 98 fraud), class labels are stratified. Then we obtained the baseline model by using random oversampling (ROS), ridsearch and 5 fold cross validation to find the best parameter for XGBoost classifier.

We use the best parameters to make predictions on test data and evaluate with statistical metrics including accuracy, precision, recall, F1 and RUC AOC score. The parameter of the XGBoost classifier from the base model will be used for comparing different GAN sampling approaches. We use different GAN generators to generate 1000 new 'fraud data' and add them to the training dataset, make predictions then evaluate the performance. For WGAN, WGAN_GP we train discriminators twice more than generators.

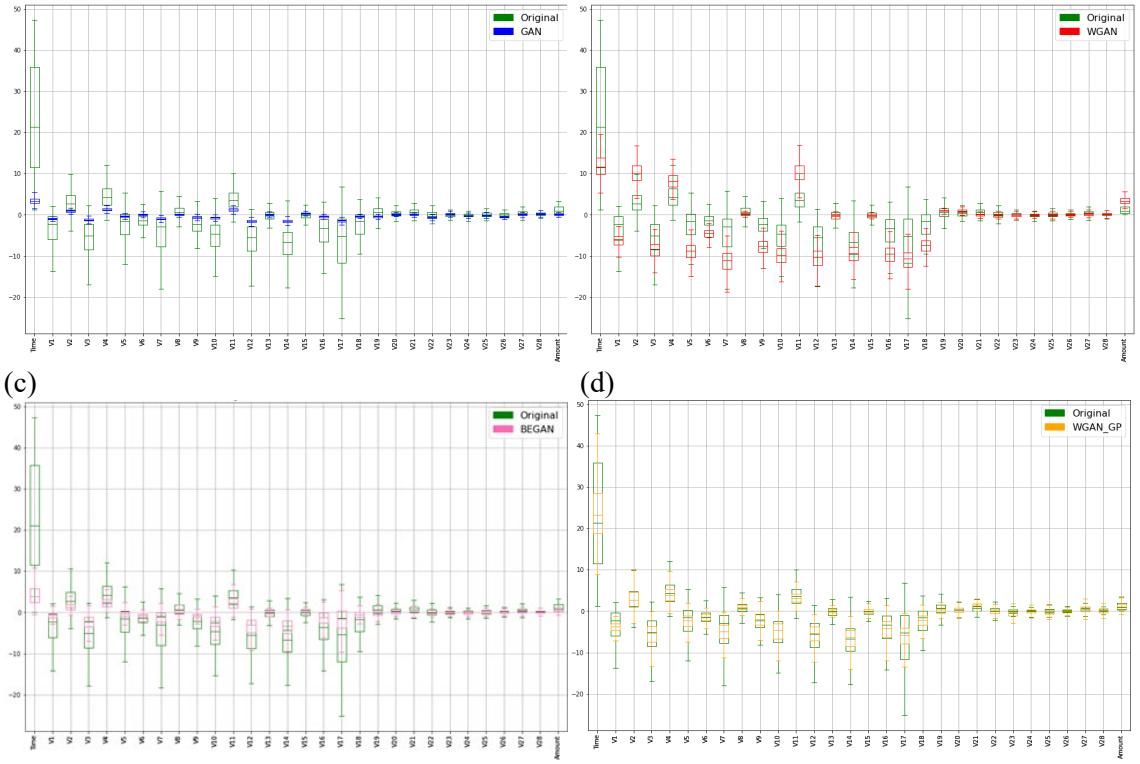
4. EXPERIMENTAL RESULTS

Since we are working with tabular data, we made some modifications on vanilla GAN, which works with image data in range of (0, 1). In our case, "tahn" is removed in order to generate appropriate data. We also removed the batch normalization layer since we found the training results are very bad.

4.1. Data Distribution

We use boxplot of data distribution to evaluate generated data with respect to the original fraud data in the train set. Vanilla GAN (Fig 1 (a)) usually suffers mode collapse and gradient vanishing problems, thus low spectrum of generated data is observed. For other improved GAN algorithms, WGAN (Fig 1 (b)) showed a wider range compared to GAN, however, generated data seems not to overlap very well with original data; BEGAN (Fig 1 (c)) also showed decent data spectrum with better overlap compared to WGAN; WGAN-GP (Fig 1 (d)) generated data has the best range and reliability.

(a) (b)



(c) (d)

Figure 1 Boxplot of original fraud data vs. Different GANs generated data

4.2. Statistical Metrics

We use accuracy, precision, recall, F1 and RUC AOC score for further evaluation. However, precision, recall and F1 are the most important indicators since we are dealing with imbalanced data, accuracy is not a fair indicator in this scenario. The statistical score of different models are shown in table 1. GAN only slightly improved the model performance which makes sense since GAN only generates a narrow range of data. For other improved GANs, mode collapse problem is resolved and the training is more stable, therefore, we observed noticeable improvement, especially WGAN-GP with the best F1 and AUC score, while BEGAN has the highest precision but low recall, this is due to the range of the generated data is not as wide as WGAN-GP. The statistical score echoes the distribution of generated data we got from the boxplot.

Table 1. Comparison of statistical score of different models.

	Base	***GAN***	***WGAN***	***WGAN-GP***	***BEGAN***
Accuracy	0.999491	0.999491	0.999544	0.999596	0.999596
Precision	0.855670	0.864583	0.867347	0.894737	0.903226
Recall	0.846939	0.846939	0.867347	0.867347	0.857143
F1 score	0.851282	0.855670	0.867347	0.880829	0.879581
ROC AUC score	0.923346	0.923355	0.933559	0.933586	0.928492

5. CONCLUSIONS

In this paper, we first use the ROS method to get a base classification model, then compare the data distribution and the statistical score of all the different GAN models on fraud detection. We showed that GANs work well with tabular data. Vanilla GAN slightly improved fraud detection since it only generated low spectrum data, while for other improved GAN models, they are able to generate a wider range of data and overlap well with the original fraud data. Especially WGAN-GP and BEGAN with top F1 scores. Using GAN as an oversampling strategy has great potential in credit card fraud detection and extremely imbalanced dataset.

For real world problems, numbers are not the only thing to look at. In fraud detection (similarly in sick patient detection, risk detection etc), the cost of false negative (FN) is usually higher than false positive (FP) [9]. We don't want to label/predict a fraudulent transaction (TP) as non-fraudulent (FN). Because in the first case the mistake in classification will be identified in further investigations. To control the fraud detection cost, the system should take into account both the cost of fraudulent behaviour that is detected and the cost of preventing it.

REFERENCES

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. & Bengio, Y., (2014) "Generative adversarial nets", *Advances in Neural Information Processing Systems*, pp 2672-2680.

[2] Arjovsky, M., Chintala, S., & Bottou, L., (2017) "Wasserstein GAN", arXiv:1701.07875 [stat.ML]

[3] Berthelot, D., Schumm, T., & Metz, L., (2017) "Began: Boundary equilibrium generative adversarial networks", arXiv:1703.10717 [cs.LG]

[4] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C., (2017) "Improved Training of Wasserstein GANs", arXiv:1704.00028 [cs.LG]

[5] Chaudhary, K., Yadav, J. & Mallick, B., (2012) "A review of fraud detection techniques: Credit card", *International Journal of Computer Applications*, Vol. 45, No.1, pp 39-44

[6] Fu, K., Cheng, D., Tu, Y., Zhang, L., A. C., (2016) "Credit Card Fraud Detection Using Convolutional Neural Networks", *Neural Information Processing*, Vol 9949, pp 483-490

[7] Chen, J., Shen, Y. & Ali, R., (2018) "Credit Card Fraud Detection Using Sparse Autoencoder and Generative Adversarial Network", *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp 1054-1059

[8] Ba, H. (2019) "Improving Detection of Credit Card Fraudulent Transactions using Generative Adversarial Networks", arXiv:1907.03355 [cs.LG]

[9] Zojaji, Z., Atani, R. E., & Monadjemi, A. H. (2016) "A survey of credit card fraud detection techniques: data and technique oriented perspective", arXiv:1611.06439 [cs.CR]