Alike and Unlike: Resolving Class Imbalance Problem in Financial Credit Risk Assessment

Yang Liu[1,2], Xiang Ao*[1,2], Qiwei Zhong[3], Jinghua Feng[3], Jiayu Tang[3], Qing He[1,2,4]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Alibaba Group, Hangzhou, China
[4]Henan Institutes of Advanced Technology, Zhengzhou University, Zhengzhou 450052, China
{liuyang17z,aoxiang,heqing}@ict.ac.cn,{yunwei.zqw,jinghua.fengjh,jiayu.tangjy}@alibaba-inc.com

ABSTRACT

Financial credit risk assessment serves as the impetus to evaluate the credit admission or potential business failure of customers in order to make early actions prior to the actual financial crisis. It aims to predict the probability that a customer may belong to a high-risk group, which is usually formulated as a binary classification problem. However, due to the lack of high-risk samples, the prevailing models suffer from the severe class-imbalance problem. Oversampling those high-risk users could alleviate this problem but the effect of noise examples is also amplified. In this paper, we propose a novel adversarial data augmentation method to solve the class imbalance problem in financial credit risk assessment. We train a generator for synthetic sample generation with a discriminator to identify real or fake instances. Besides, an auxiliary risk discriminator is trained cooperatively with the generator to assess the credit risk. Experimental results on three real-world datasets demonstrate the effectiveness of the proposed framework.

CCS CONCEPTS

• **Computing methodologies → Neural networks**.

KEYWORDS

data augmentation, generative model, class imbalance problem

ACM Reference Format:
Yang Liu, Xiang Ao, Qiwei Zhong, Jinghua Feng, Jiayu Tang, Qing He. 2020. Alike and Unlike: Resolving Class Imbalance Problem in Financial Credit Risk Assessment. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412111

1 INTRODUCTION

The assessment of financial credit risk[2] plays an essential role on both economics and society. It aims to predict the possibility of high-risk resulting from a corporation or consumer's inability to fulfil its contractual obligations. However, in actual business scenarios, most users belong to low-risk group as they could pay off their loans on time. Only a very small fraction of users are deemed as high-risk which belongs to the minority class but is our target to be identified. Therefore, the task of financial credit risk assessment is typically a class imbalance problem where the high-risk and low-risk samples are severely skewed.

The class imbalance problem has attracted increasing attention due to the prevalence of imbalanced data in real-world applications. For image classification[5], data often exhibit long-tailed class distribution, i.e., most data belong to a few majority classes. For name entity recognition[11], most tokens are backgrounds with tagging class O and the number of tokens tagged O is 8 times as many as those with entity labels in the widely-used OntoNotes5.0[12] dataset. Re-sampling methods[1, 4] could alleviate the class imbalance problem and can be further divided into two strategies, i.e. undersampling the majority class and oversampling the minority class. Both strategies are adoptable but still have some drawbacks. Undersampling methods reduce the amount of training data thus may limit the generalization ability of the model and suffer from overfitting problem. Oversampling methods could be further categorized into random oversampling and generative oversampling. Random oversampling methods repeat the samples from the minority class randomly and do not add informative content, which may amplify the effect of noise samples. Generative oversampling methods generate new synthetic samples to augment the minority class without loss of generalization.

There are two challenges to adopt generative oversampling methods for financial credit risk assessment. Firstly, the synthetic representations output by generative models should resemble high-risk user representations meanwhile be distinguishing from low-risk user representations. It is hard to train a generative model to meet these two requirements simultaneously. Secondly, the high-risk samples to be generated is the minority class, of which the number is deficient to train a model substantially well. In financial areas, high-risk users account for less than 1% of all the users and could not span the whole user space.

Taking both challenges into consideration, we propose a novel **A**dversarial **D**ata **A**ugmentation method with **A**uxiliary discriminato**R** (**ADAAR**) for financial credit risk assessment in this paper.

*Corresponding Author.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412111

To overcome the first challenge, we design an adversarial data augmentation framework consisting of a generator, a discriminator, and an auxiliary discriminator for synthetic sample generation. The discriminator identifies fake synthetic samples from real high-risk samples and the generator is trained to fool the discriminator thus they are adversarially learned so that synthetic samples and high-risk samples are "alike". Besides, an auxiliary risk discriminator is designed to assess the risk of users such that synthetic samples are "unlike" the low-risk ones. The generator is optimized cooperatively with the auxiliary risk discriminator to reduce the assessment error. As for the second challenge, we argue that the majority class could help with the generation of the minority class since they belong to the same user space. Thus we train an autoencoder with all user representations such that the decoder is able to map the latent space to the user space and used to initialize the generator.

Experiments on three real-world datasets show that ADAAR outperforms all the baseline methods by augmentation with high-risk alike and low-risk unlike user representations.

We give a detailed description of ADAAR in Section 2 and the experimental results are shown in Section 3. Subsequently we review the related work in Section 4 and conclude this paper in Section 5.

2 METHOD

In this section, we focus on how to generate new synthetic user representations for the class imbalance problem in financial credit risk assessment. Given a set of financial users represented by $\mathcal{X}_{\text{ori}} = \mathcal{X}_{\text{high}} \cup \mathcal{X}_{\text{low}}$, the low-risk users denoted by $\mathcal{X}_{\text{low}}$ are the majority class while the high-risk users $\mathcal{X}_{\text{high}}$ are the minority class. Our task is to generate a set of synthetic high-risk users $\mathcal{X}_{\text{syn}}$ so that the performance of the risk assessment model trained on $\mathcal{X}_{\text{ori}} \cup \mathcal{X}_{\text{syn}}$ could be improved from the model trained on $\mathcal{X}_{\text{ori}}$.

2.1 Overview

The overall framework is shown as Figure 1. First, we train an autoencoder to reconstruct all the user representations. Then the generator G is initialized with the parameters of the decoder De and the two discriminators D_1, D_2 are initialized with the encoder En. The discriminator D_1 is trained with the generator G under the adversarial loss such that $\mathcal{X}_{\text{syn}}$ and $\mathcal{X}_{\text{high}}$ are alike. The risk discriminator D_2, together with the generator G, is optimized with the binary cross-entropy loss so that $\mathcal{X}_{\text{syn}}$ and $\mathcal{X}_{\text{low}}$ are unlike.

2.2 Pre-training

We train an autoencoder at first to extract useful information from high-dimensional raw user representations. Basically, an autoencoder is composed of an encoder En and a decoder De. The input of the autoencoder is the original user representations, denoted as $\mathcal{X}_{\text{ori}}$, and the output is the reconstructed user representations $\mathcal{X}_{\text{rec}}$. For $x \in \mathcal{X}_{\text{ori}}$, $\hat{x} = \text{De}(\text{En}(x))$ and then $\hat{x} \in \mathcal{X}_{\text{rec}}$. L_2 reconstruction loss as shown in (1) is used to train the autoencoder.

$$\mathcal{L}_{ae} = \sum_{x \in \mathcal{X}_{\text{ori}}} \|x - \hat{x}\|_2^2 \tag{1}$$

2.3 Data Generation

The generator starts from the latent space $\mathcal{Z}$ and the generated samples could be denoted as $G(z), z \in \mathcal{Z}$. The discriminator D_1 is designed to distinguish synthetic representations $\mathcal{X}_{\text{syn}} = \{G(z), z \in \mathcal{Z}\}$ from the real high-risk user representation $\mathcal{X}_{\text{high}}$. Meanwhile, the generator wants to fool D_1 by generating more realistic high-risk user samples. Suppose we label $\mathcal{X}_{\text{syn}}$ as high-risk samples, the discriminator D_2 is trained to identify high-risk samples (both real and synthetic) from low-risk samples. Therefore, the generator would be optimized to generate samples that could reduce the risk assessment error. We define the oversampling rate $K = \frac{|\mathcal{X}_{\text{high}} \cup \mathcal{X}_{\text{syn}}|}{|\mathcal{X}_{\text{high}}|}$ thus the output size of generator is $|\mathcal{X}_{\text{syn}}| = (K-1) \cdot |\mathcal{X}_{\text{high}}|$.

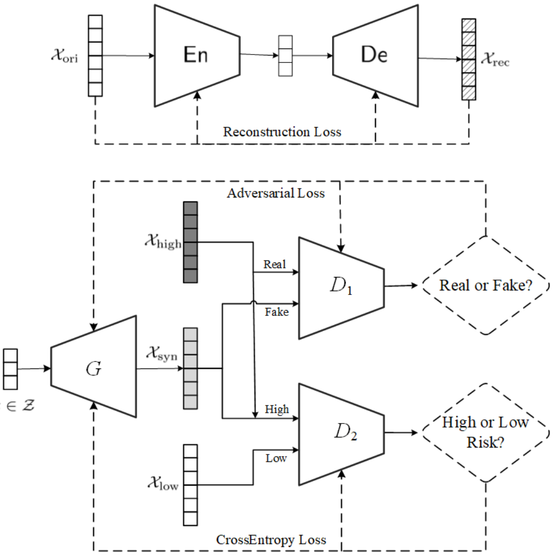

Figure 1: The framework for ADAAR. En **and** De **are pre-trained to initialize D_1, D_2 and G. D_1 is trained adversarially with G so that $\mathcal{X}_{\text{syn}}$ and $\mathcal{X}_{\text{high}}$ are alike. D_2 is trained cooperatively with G so that $\mathcal{X}_{\text{syn}}$ and $\mathcal{X}_{\text{low}}$ are unlike.**

2.4 Training Procedure

The adversarial loss $\mathcal{L}_{adv}$ is designed for the discriminator D_1 and the generator G while the cross-entropy loss $\mathcal{L}_{ce}$ is adopted for the discriminator D_2 and the generator G.

$$\mathcal{L}_{adv} = -\sum_{x \in \mathcal{X}_{\text{high}}} \log(D_1(x)) - \sum_{z \in \mathcal{Z}} [1 - \log(D_1(G(z)))] \tag{2}$$

$$\mathcal{L}_{ce} = -\sum_{x \in \mathcal{X}_{\text{high}} \cup \mathcal{X}_{\text{syn}}} \log(D_2(x)) - \sum_{x \in \mathcal{X}_{\text{low}}} \log(1 - D_2(x)) \tag{3}$$

The parameters of G, D_1, D_2 are optimized with Adam[6] optimizer alternately. D_1 is firstly optimized to minimize the adversarial loss $\mathcal{L}_{adv}$ and G is subsequently trained to maximize it. Next, both G and D_2 are trained to minimize the cross-entropy loss $\mathcal{L}_{ce}$.

3 EXPERIMENT

3.1 Dataset

We collect three real-world datasets from an online credit payment service provided by Alibaba Group and the users are arranged chronologically. M7(ranging from 2018/07/01 to 2018/07/31),

Table 1: Dataset Statistics

Dataset	#Users	#Major	#Minor	Rate
M7	334,695	330,785	3,910	1.18%
M9	404,491	400,778	3,713	0.93%
M11	524,935	520,369	4,566	0.88%

M9(ranging from 2018/09/01 to 2018/09/30) and M11(ranging from 2018/11/01 to 2018/11/30) are users from one month and we collect 908 attributes for each user. According to the general definition in financial area, we define high-risk users as those who default within one month and low-risk users as others. It is noteworthy that the time interval between training and testing set should not be less than one month since we need the data in the next month when defining the label. Therefore, we design two groups of experiments, namely M7/M9 and M9/M11. M7/M9 is trained on M7 and tested on M9 while M9/M11 is trained on M9 and tested on M11. The statistical information of these three datasets is exhibited in Table 1. The ratio of the high-risk users is around 1% in these datasets.

3.2 Compared Methods

NS(No Sampling) indicates no sampling strategy is adopted. **ROS** (Random OverSampling) repeats the minority class for several times randomly. **SMOTE**[1] generates new synthetic minority samples by performing linear interpolation operations between existing minority samples and their nearest neighbors within the same class. **ADASYN**[4] is a novel adaptive synthetic sampling approach and uses a weighted distribution for different minority class examples according to their level of difficulty in learning. **BAGAN**[10] is balancing GAN for data augmentation, of which the discriminator has a single output that returns either a problem-specific class label or the label fake. **GLGAN**[13] considers both global and local information of the given data in the synthetic minority sample generation process. **ADAAR** is our method.

3.3 Experimental Settings

The major parameters of ADAAR include the latent vector dimension d, learning rate η, oversampling rate K, the batch size m, the maximum epoch $MaxEpoch$, the architecture of En, De, D_1, D_2, and G. For the three datasets above, we set $d = 100$, $\eta = 0.0002$, $K = 20$, $m = 256$, $MaxEpoch = 10$. En and De are implemented by an MLP and the dimension of each layer is $[908, 256, 100, 256, 908]$. D_1 and D_2 have four layers with the same dimension $[908, 256, 100, 2]$. Both De and G has three layers with dimension $[100, 256, 908]$. For each layer, LeakyRELU with slope 0.2 is adopted as activation function. For the baselines, we finely tuned the corresponding parameters in order to perform a fair comparison. An MLP classifier with two hidden layer sized $[32, 16]$ is trained as the base risk model to evaluate the performance of each method for the class imbalance problem. The ADAAR framework is implemented in Pytorch.

3.4 Evaluation Metrics

We use two widely adopted metrics to measure the performance of our data augmentation strategy on financial credit risk assessment, namely **AUC** and **R@P$_\mathbf{N}$**.

The first metric **AUC** is the area under the ROC Curve. The second metric **R@P$_\mathbf{N}$** means the **R**ecall when the **P**recision equals N. In our dataset, we set N=10%. Since the minority rate in credit payment services is low in general (about 1% in our dataset), this metric that lifts 10 times in our dataset (10% vs 1%) indicates the ability to detect top-ranked positive samples and balance the impact on the real-world business system. The higher **AUC** and **R@P$_\mathbf{N}$** indicate the higher performance of the approaches.

3.5 Results and Analysis

We report the the average and standard deviation of 5 runs in Table 2 and ADAAR achieves at most 1.8% improvement in AUC and 3.8% improvement in R@P$_{0.1}$, compared with the strongest baseline BAGAN. SMOTE and ADASYN perform worse than NS and ROS in R@P$_{0.1}$ therefore it is not reasonable to apply interpolation directly on user representations for new samples synthesis. GLGAN performs interpolation on the latent representations and gets better results in R@P$_{0.1}$ than SMOTE or ADASYN but is still lower than BAGAN and our ADAAR. ADAAR improves from BAGAN because the designed risk discriminator makes synthetic samples unlike low-risk users. Besides, GAN-based models have smaller variance than other baselines and ADAAR is the most stable from our results.

3.6 Ablation Study

We conduct the ablation test by removing the component of ADAAR to prove its effectiveness. Three key components are identified in ADAAR, namely the autoencoder(*AE*), the discriminator D_1 and the risk discriminator D_2. From the results in Table 2, we could see that both AUC and R@P$_{0.1}$ would decline by removing any part of ADAAR. ADAAR w/o *AE* gets the worst performance because it is cruicial for the generator to learn the user representation space to generate realistic samples. Two discriminators are also necessary to make synthetic samples alike high-risk and unlike low-risk users.

3.7 Parameter Sensitivity

For the class imbalance problem, we need to decide the number of synthetic samples for the minority class to achieve the best performance. Therefore we evaluate the sensitivity of ADAAR and two strong baselines, BAGAN and ROS, with respect to the oversampling rate K on M7/M9. As shown in Figure 2, as K increases from 10 to 50, the performances of ADAAR and BAGAN change much more slightly than that of ROS, which indicates that GAN-based methods are more robust to sampling rate. ROS performs poorly when K is large because repeating the minority class may magnify the effect of noise samples. Besides, ADAAR outperforms BAGAN and ROS under all these settings of K with much smaller variance.

4 RELATED WORK

The class imbalance problem can be tackled by data re-sampling or cost-sensitive learning. Re-sampling methods adjust the number of examples by oversampling the minority class or undersampling the majority class. Oversampling methods include interpolation-based methods like SMOTE[1] and its variants[4], GAN-based generative models like BAGAN[10] and GLGAN[13]. Cost-sensitive learning methods influence the loss function by assigning relatively higher costs to examples from minor classes, including Focal Loss[8], Dice Loss[7], Class-Balanced Loss[3], etc. Our work is different from them since we design a novel adversarial data

Table 2: Performance of compared methods on M7/M9 and M9/M11.

	Dataset	M7/M9		M9/M11	
	Method	AUC	R@P$_{0.1}$	AUC	R@P$_{0.1}$
Baselines	NS	0.8698 ± 0.0029	0.3626 ± 0.0223	0.8366 ± 0.0041	0.2399 ± 0.0099
	ROS	0.8742 ± 0.0031	0.3909 ± 0.0275	0.8468 ± 0.0076	0.2478 ± 0.0253
	SMOTE	0.8717 ± 0.0079	0.3606 ± 0.0404	0.8410 ± 0.0059	0.1933 ± 0.0226
	ADASYN	0.8751 ± 0.0019	0.3582 ± 0.0252	0.8389 ± 0.0071	0.2072 ± 0.0170
	BAGAN	0.8740 ± 0.0012	0.3991 ± 0.0104	0.8410 ± 0.0046	0.2523 ± 0.0081
	GLGAN	0.8737 ± 0.0016	0.3849 ± 0.0128	0.8341 ± 0.0043	0.2455 ± 0.0109
Ours	ADAAR	**0.8780 ± 0.0009**	**0.4170 ± 0.0065**	**0.8592 ± 0.0008**	**0.2910 ± 0.0063**
Ablation Test	ADAAR w/o *AE*	0.8736 ± 0.0021	0.3871 ± 0.0126	0.8322 ± 0.0026	0.2384 ± 0.0138
	ADAAR w/o D_1	0.8748 ± 0.0019	0.3946 ± 0.0097	0.8380 ± 0.0049	0.2644 ± 0.0328
	ADAAR w/o D_2	0.8757 ± 0.0015	0.3928 ± 0.0059	0.8549 ± 0.0076	0.2689 ± 0.0253

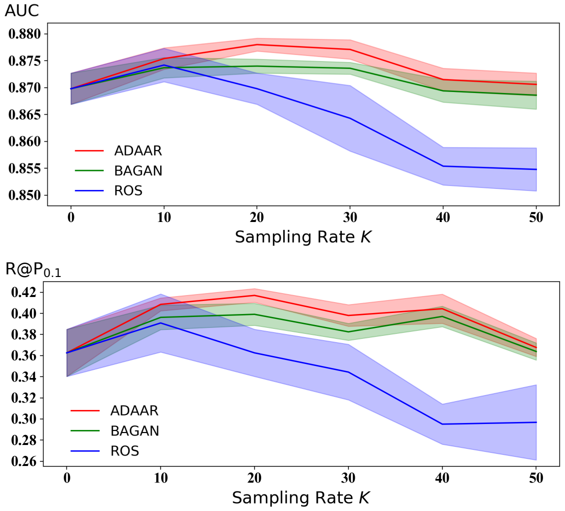

Figure 2: AUC and R@P$_{0.1}$ on M7/M9 with different oversampling rate K.

augmentation framework to solve the class imbalance problem in financial credit risk assessment.

Financial credit risk assessment recently are achieved by machine learning [9, 14] or graph mining methods [15]. Both the two technical routes suffer from the class imbalance problem and most of the approaches adopt the random oversampling method to resolve. Different from that, we solve the problem by a novel adversarial data augmentation framework.

5 CONCLUSION

In this paper, we study the class imbalance problem of financial credit risk assessment and propose ADAAR, an adversarial data augmentation framework. The synthetic samples output by ADAAR resembles real high-risk users since we design an autoencoder to learn the user space and the generator has to fool the discriminator which identifies fake samples. Meanwhile, synthetic samples could be distinguished from low-risk users because we have an auxiliary discriminator to assess the risk. Experimental results demonstrate that ADAAR outperforms other data augmentation methods on three real-world datasets.

ACKNOWLEDGMENT

This research work is supported by Alibaba Group through Alibaba Innovative Research Program. The work is also supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002104, the National Natural Science Foundation of China under Grant No. 61976204, U1811461, the Project of Youth Innovation Promotion Association CAS. We thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *JAIR* (2002).

[2] Ning Chen, Bernardete Ribeiro, and An Chen. 2016. Financial credit risk assessment: a recent review. *Artificial Intelligence Review* (2016).

[3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *CVPR*.

[4] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*.

[5] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for imbalanced classification. In *CVPR*.

[6] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[7] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice Loss for Data-imbalanced NLP Tasks. In *ACL*.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *CVPR*.

[9] Can Liu, Qiwei Zhong, Xiang Ao, Sun Li, Wangli Lin, Jinghua Feng, Qing He, and Jiayu Tang. 2020. Fraud Transactions Detection via Behavior Tree with Local Intention Calibration. In *KDD*.

[10] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. 2018. Bagan: Data augmentation with balancing gan. In *ICML Workshop*.

[11] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* (2007).

[12] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *ACL*.

[13] Wentao Wang, Suhang Wang, Wenqi Fan, Zitao Liu, and Jiliang Tang. 2020. Global-and-Local Aware Data Generation for the Class Imbalance Problem. In *SDM*.

[14] Ya-Lin Zhang, Jun Zhou, Wenhao Zheng, Ji Feng, Longfei Li, Ziqi Liu, Ming Li, Zhiqiang Zhang, Chaochao Chen, Xiaolong Li, et al. 2019. Distributed deep forest and its application to automatic detection of cash-out fraud. *TIST* (2019).

[15] Qiwei Zhong, Yang Liu, Xiang Ao, Binbin Hu, Jinghua Feng, Jiayu Tang, and Qing He. 2020. Financial Defaulter Detection on Online Credit Payment via Multi-view Attributed Heterogeneous Information Network. In *WWW*.