



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사학위논문

SMS: 불균형 이진 분류를 위한 인공 데이터 샘플링 기법

**SMS: A Deep Synthetic Minority Sampler for
Imbalanced Binary Classification**

2021년 2월

서울대학교 대학원
컴퓨터공학부
이재원

SMS: 불균형 이진 분류를 위한 인공 데이터 샘플링 기법

SMS: A Deep Synthetic Minority Sampler for
Imbalanced Binary Classification

지도교수 강 유

이 논문을 공학석사 학위논문으로 제출함

2020년 10월

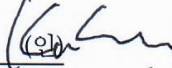
서울대학교 대학원

컴퓨터공학부


이재원

이재원의 석사 학위논문을 인준함

2020년 12월

위 원 장 김 선 

부 위 원 장 강 유 

위 원 김 건 희 

Abstract

SMS: A Deep Synthetic Minority Sampler for Imbalanced Binary Classification

Jae-Won LEE

Department of Computer Science & Engineering

Graduate School

Seoul National University

Given an imbalanced dataset, how can we create high fidelity synthetic minority instances for training robust and unbiased classifiers? Data imbalance is common in mission-critical fields where costs associated with procuring minority instances are prohibitively expensive. Training classifiers on imbalanced datasets result in unreliable predictions and low performance. Oversampling techniques are employed to restore balance to the dataset, allowing the classifier to learn a more accurate representation of the true data distribution. Thus, generating a set of synthetic samples that are i) realistic, ii) containing varying degrees of class confidence, and iii) diverse is essential. Existing methods create samples that do not satisfy all the desired properties.

We propose Synthetic Minority Sampler (SMS), an oversampling framework designed for highly imbalanced datasets. SMS employs two generators to create a balanced ratio of normal and borderline samples that teach classifiers a robust and

unbiased class representation. SMS accounts for the scarce minority instances via a class-conditional diversity loss to ensure that generated minority samples are diverse. Additionally, SMS stabilizes the training process by introducing a weighted random sampler to balance the class proportion of mini-batches, and data augmentation to prevent the discriminator from overfitting. Experimental results show that models trained on an imbalanced dataset augmented with synthetic data sampled from SMS outperform competitors in the binary classification task, achieving up to 10.06% higher F1-score than the competitors.

Keywords : Generative Adversarial Network, Data Imbalance, Oversampling

Student Number : 2019-25223

Contents

I. Introduction	1
II. Related Works	4
III. Proposed Method	6
3.1 Overview	6
3.2 Normal and Borderline Sample Generation	8
3.3 Class-conditional Diversity Loss	10
3.4 Generators, Discriminator, and Classifier Training	11
3.5 Stabilizing Training	12
IV. Experiments	14
4.1 Experimental Settings	14
4.2 Performance	19
4.3 Synthetic Image Quality and Diversity	22
4.4 Ablation Study	23
V. Conclusion	26
References	27
Abstract in Korean	29

List of Figures

Figure 1.	SMS enables achieving the highest predictive performance for imbalanced classification.	2
Figure 2.	Architecture of SMS. The first two layers are shared between G_B and G_N to learn common features. The remaining layers exploit the shared knowledge to generate normal and borderline samples.	7
Figure 3.	G_B aims to generate samples that are difficult to classify. The conditional discriminator D ensures that the generated samples do not diverge far from the learned data distribution modeled by the conditional probability $D(x y)$. G_N spawns points in regions that are easy to classify, ensuring that the sparse, safe regions are not neglected.	10
Figure 4.	Visualization of synthetic samples generated by ACGAN, CDCGAN, GAMO, and SMS (from top to bottom). (a) SMS creates realistic and diverse images despite the lack of minority instances. (b) SMS generates diverse samples containing the key features of the digit 4 while retaining large variations of noisy features for training a robust classifier.	19

Figure 5. Classifiers trained using real and synthetic minority samples generated by SMS achieves the best performance on the severely imbalanced SVHN dataset. The performance gaps between the classifiers trained on samples generated by SMS and competitors are the largest when the ratio is 200:1. 20

Figure 6. Examples of minority class instances (recyclable objects) in the WASTE classification dataset. Despite belonging to the same class, each object has different key salient features, making it difficult to identify common features defining the minority class. 22

List of Tables

Table 1. Dataset summary.	15
Table 2. The precision, recall and F1-score of classifiers on the test dataset $data_{test}$ after training on augmented dataset $data_{synthetic}$. SMS shows the best performance across all datasets. Bold text indicates the best results.	18
Table 3. Classification performance of evaluation model on CIFAR-10 and SVHN when trained on augmented dataset generated by SMS and its competitors. Each method is trained on a dataset with an imbalance ratio of 400:1.	21
Table 4. FID score results. Lower values indicate better image quality and diversity. The best scores are highlighted in bold. SMS_{-W} , SMS_{-D} are variants of SMS without the weight sharing and diversity loss components.	23
Table 5. Ablation studies. SMS outperforms both 1) SMS_{-W} a variant of SMS without weight sharing, and 2) SMS_{-D} , a variant of SMS without the class-conditional diversity loss term.	24

Chapter 1

Introduction

Given an imbalanced dataset, how can we create a set of realistic and diverse synthetic minority instances for training robust classifiers that generalize well? Imbalanced datasets are common in mission-critical fields such as cancer or foreign object detection, where highly prohibitive cost is associated with misclassifying the minority class instances. Thus, overcoming the data imbalance problem is crucial from both the business and research standpoint.

Learning from imbalanced data is challenging for the following reasons: (i) severe lack of minority instances makes learning accurate decision boundaries difficult, and (ii) models are incentivized to minimize costs by focusing on majority class instances. An appropriate solution is to oversample the minority class to augment the imbalanced dataset. Existing works attempt to alleviate data imbalance via oversampling using simple heuristics or generative models such as Generative Adversarial Networks (GANs). To effectively oversample and train a robust classifier, we must overcome the following challenges: (i) create realistic samples that contain varying degrees of class confidence, empowering a classifier to learn an accurate and general decision boundary, (ii) create diverse samples by avoiding mode collapse, and (iii) stabilize training on imbalanced data. Realistic borderline samples with low class confidence scores are valuable because they teach the classifier an unbiased decision boundary. With highly imbalanced datasets, normal samples with high class confidence scores close to real minority points help classifiers learn the class distribution.

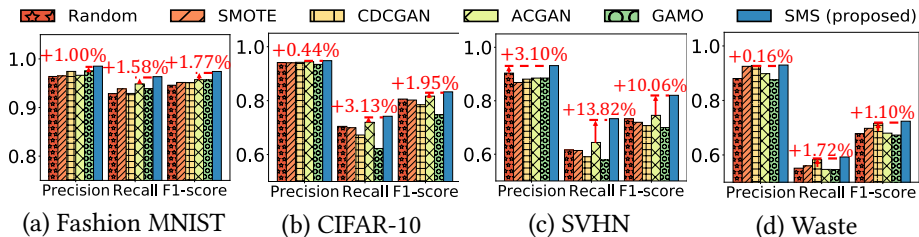


Figure 1: SMS enables achieving the highest predictive performance for imbalanced classification.

Mode collapse occurs when the generator maps multiple points in the noise space to the same point in the data space, thus stifling diversity. Discriminators are also prone to ignoring or overfitting on the scarce minority samples, resulting in unbalanced training. However, existing methods [1, 2, 3, 4] fail to generate points that contain varying degrees of class confidence. Additionally, GAN-based methods [2, 3, 4] also suffer from mode collapse and stability issues when trained on imbalanced datasets.

To address the aforementioned issues, we propose SMS (a Deep Synthetic Minority Sampler for Imbalanced Binary Classification), a novel GAN-based oversampling framework for highly imbalanced datasets. By jointly learning borderline and normal features, SMS generates realistic and diverse samples with varying degrees of class confidence that guides classifiers to learn a robust and unbiased decision boundary through its novel weight sharing dual generator model. Additionally, SMS tackles the second challenge by adding a class-conditional diversity loss term that alleviates mode collapse on minority instances. We stabilize GAN training by leveraging a weighted random sampler and data augmentation strategies to add variations and balance the ratio of majority and minority instances in each mini-batch. Due to its careful design, SMS generates high-quality synthetic samples for training robust classifiers. We summarize our main contributions as follows:

- **Novel generation framework for imbalanced data.** We generate samples of both high and low degrees of class confidence to train a robust and unbiased classifier (see Sections 3.2 and 3.4).
- **Robustness to severe imbalance.** Compared to existing methods, our model is robust to severe data imbalance. Our class-conditional diversity loss (see Section 3.3) selectively targets the minority class instances, alleviating mode collapse on the minority class. Coupled with the training stabilization technique (see Section 3.5), our model is capable of learning an accurate representation of the minority class and its decision boundary.
- **Experiment.** Experiments (see Section 4) show that SMS generates the best synthetic samples for training a classifier with imbalanced data (see Figure 1).

We review related works in Section 2, introduce our proposed method SMS in Section 3, evaluate SMS and competitors in Section 4, and conclude in Section 5.

Chapter 2

Related Works

Traditional data sampling techniques utilize simple heuristics to generate minority samples [1, 5]. Chawla et al. [1] proposed SMOTE, an over-sampling technique that interpolates new samples on the space between the target sample and a randomly selected nearest neighbor. Although SMOTE works well on simple multivariate datasets, its inability to appropriately measure distance between images ultimately results in samples that do not accurately represent the target class.

Real-world datasets contain borderline samples. These borderline points offer valuable information for differentiating between majority and minority instances. Creating only borderline samples results in a more brittle class distribution, due to neglecting sparse safe regions where minority classes reside, and the possibility of generating noisy samples. Last et al. [6] address this issue by proposing a method for generating samples in safe and crucial regions of the data space, thus avoiding the generation of noise and reinforcing the crucial sparse regions with synthetic points, enabling models to learn a robust decision boundary.

Recent advancements in generative models [7, 8] have propelled research in the field of synthetic image generation. Mirza et al. [9] proposed Conditional GAN, which conditions a generator on class labels $G(z|y)$ to generate samples that represent the given class, providing a degree of control over the distribution of the generated sample. Odena et al. [3] proposed a conditional GAN framework that leverages an auxiliary classifier to generate high-fidelity samples. Although ACGAN works well on

balanced data, its classifier-driven training regime is not suited for training on imbalanced data.

Recent works [4, 10] have employed deep learning techniques to oversample image datasets. DOS [10] generates multiple overloaded instances from each training sample by pairing with different targets sampled in the linear subspace of the original data to incrementally shift the targets to the class mean. Thus, DOS is limited to sampling uniformly across the minority class distribution, resulting in a lack of borderline samples. GAMO [4] introduces a novel three-way adversarial training between the generator, discriminator, and classifier to learn class decision boundaries. However, because the convex hull and data distribution modeled by the conditional discriminator is learned in the latent space, these learned representations may not properly reflect the true data distribution, resulting in noisy or duplicate samples when projected onto the data space.

Chapter 3

Proposed Method

We propose SMS, a GAN-based oversampling framework for highly imbalanced binary classification datasets. We provide a brief overview of our method in Section 3.1. Then, we describe the normal and borderline sample generation in Section 3.2, and introduce the class-conditional diversity loss in Section 3.3. We discuss the generator, discriminator, and classifier training process in Section 3.4, and examine the weighted random sampling and data augmentation strategy in Section 3.5.

3.1 Overview

We design SMS to create synthetic points that emulate the distribution of real-world datasets by generating a balanced proportion of normal and borderline samples. Figure 2 shows the overall architecture of SMS. We address the following challenges present in existing GAN-based oversampling frameworks.

1. **Generating samples for training a robust and unbiased classifier.** Augmented datasets consisting only of points residing near existing minority instances teach classifiers a biased representation. How can we generate samples with key features required for training a robust and unbiased classifier?
2. **Creating diverse samples.** When GANs collapse, the generator maps multiple points in the noise space to the same image, which adversely impacts diversity. How can we mitigate mode collapse?

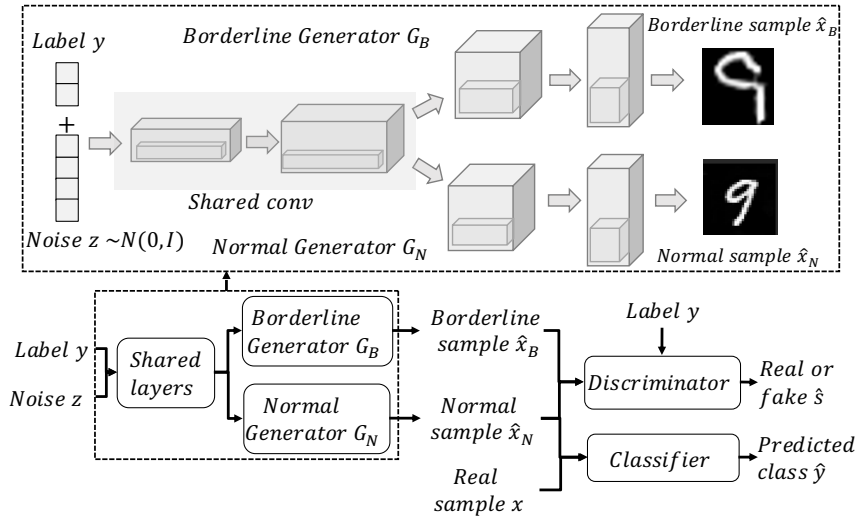


Figure 2: Architecture of SMS. The first two layers are shared between G_B and G_N to learn common features. The remaining layers exploit the shared knowledge to generate normal and borderline samples.

3. **Stabilize training on imbalanced data.** The discriminator is likely to ignore minority instances when training on highly imbalanced datasets. How can we stabilize the training process?

We address the aforementioned challenges with the following ideas:

1. **Normal and borderline sample generation.** SMS employs a normal and borderline generator to generate realistic samples with varying degrees of class confidence. Together, these samples help train a robust classifier with substantial generalization capabilities (details in Sections 3.2 and 3.4).
2. **Class-conditional diversity loss.** The class-conditional diversity loss encourages the generators to map different points in the noise space to different outputs. By placing emphasis on the minority instances, we prevent mode collapse

for the minority class, enabling our model to generate a diverse range of minority samples (details in Sections 3.3 and 3.4).

3. **Stabilizing training.** We ensure that each mini-batch has approximately an equal proportion of majority and minority instances. We then apply data augmentation to the balanced mini-batch, preventing the discriminator from overfitting (details in Section 3.5).

3.2 Normal and Borderline Sample Generation

The main challenge for designing generators is to produce synthetic samples with varying degrees of class confidence to train a robust and unbiased classifier. We tackle the challenge by employing two generators, normal generator G_N and borderline generator G_B .

Normal generator. Because of the lack of real minority instances, normal synthetic minority instances are required to fortify sparse regions of the data distribution where minority instances reside. This helps the classifier learn a robust decision boundary. C denotes the classifier network that models the conditional probability $p(y|x)$ as $C(x) = P(y = 1|x)$, where $y = 1$ represents the minority class label. G_N minimizes the following loss function:

$$\begin{aligned}
 L_{G_N} = & \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G_N(z|y)))] \\
 & - \mathbb{E}_{z \sim p_z(z)} [y \log C(G_N(z|y)) + (1 - y) \log(1 - C(G_N(z|y)))]
 \end{aligned}
 \tag{3.1}$$

The loss function in Equation 3.1 drives G_N to generate samples that are easy to classify by fooling the discriminator D and minimizing the error of the classifier C . G_N is rewarded for creating realistic samples near regions where real minority

instances reside.

Borderline generator. We generate borderline samples because they offer valuable information for differentiating between majority and minority instances that are not present in the original dataset. This helps the classifier learn an unbiased class representation. G_B minimizes the following loss function:

$$\begin{aligned}
 L_{G_B} = & \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G_B(z|y)))] \\
 & + \mathbb{E}_{z \sim p_z(z)} [C(G_B(z|y)) \log C(G_B(z|y)) + (1 - C(G_B(z|y))) \log(1 - C(G_B(z|y)))]
 \end{aligned}
 \tag{3.2}$$

In contrast to the normal generator G_N , the borderline generator G_B is rewarded for generating points near the decision boundary by maximizing the entropy of predictions generated from the classifier; that is, it aims to make the predictions close to 50% as shown in Equation 3.2. The discriminator loss term regularizes G_B so that generated samples are bound to the estimated real data distribution modeled by the conditional discriminator $D(x|y)$ as shown in Figure 3.

Weight sharing. SMS employs two generators: G_B and G_N , responsible for generating borderline and normal samples, respectively. G_B and G_N share weights in the first two layers to exploit the fact that both normal and borderline instances share similar features. As a result, knowledge of common features is shared between G_B and G_N , allowing each model to learn a more accurate representation of normal and borderline samples. Additionally, the number of parameters in the model is reduced thanks to the weight sharing.

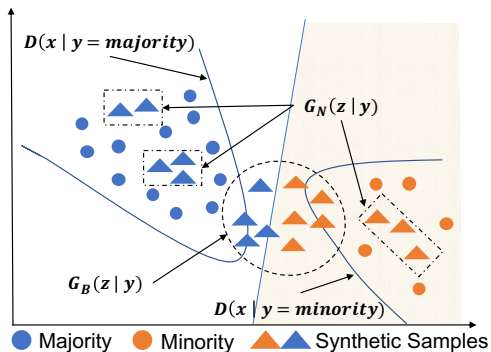


Figure 3: G_B aims to generate samples that are difficult to classify. The conditional discriminator D ensures that the generated samples do not diverge far from the learned data distribution modeled by the conditional probability $D(x|y)$. G_N spawns points in regions that are easy to classify, ensuring that the sparse, safe regions are not neglected.

3.3 Class-conditional Diversity Loss

Mode collapse is an issue, because it severely limits the diversity of generated samples. Due to the severe lack of minority instances, a reliable strategy is necessary to steer the generators away from mode collapse. If such measures are not taken, the generator is likely to collapse away from low density regions in the data space. We propose a class-conditional diversity loss to create samples that vary from one another in proportion to the pairwise distance in the noise space z . In Equation 3.3, let β be a set of minority class examples $\{y, z\}$ in a mini-batch. $\{y_i, z_i\}$ and $\{y_j, z_j\}$ refer to two labels and noise vector pairs used to generate images. G refers to the generating function that maps labels and noise vectors to images. $dist$ is the L1 distance between a pair of generated images. The class-conditioning incentivizes the generator to mitigate mode collapse on the minority class, thus ensuring sample diversity.

The diversity loss is defined as follows:

$$L_{div} = \exp \left(- \sum_{(y_i, z_i) \in \beta} \sum_{(y_j, z_j) \in \beta} \|z_i - z_j\|_2^2 \cdot \text{dist}(G(y_i, z_i), G(y_j, z_j)) \right) \quad (3.3)$$

The loss function increases the pairwise distance between sampled points in each batch in proportion to the pairwise Euclidean distance $\|z_i - z_j\|_2^2$ between the noise vectors z_i, z_j . This loss function adapts the diversity loss [11] by penalizing the generator based on the class-conditional pairwise distance between sampled points in each batch to aptly account for minority instances.

3.4 Generators, Discriminator, and Classifier Training

Due to weight sharing, SMS optimizes both generators jointly. When training the generators, we minimize the following loss function L_G :

$$L_G = L_{G_N} + L_{G_B} + \lambda L_{div} \quad (3.4)$$

The discriminator ensures that generated samples are realistic and contains key properties of the target class by conditioning on labels. During adversarial training, the discriminator D is tasked with correctly identifying real and fake samples by maximizing Equation 3.5:

$$L_D = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G_N(z|y)))] \\ + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G_B(z|y)))] \quad (3.5)$$

The classifier C is tasked with classifying real and generated samples. It acts as a critic, pushing the generators to become better at their respective tasks via constant feedback. We minimize the cross entropy loss, where $y = 1$ represents minority labels and $C(x) = P(y = 1|x)$ as defined in Section 3.2.

$$\begin{aligned}
L_C = & -\mathbb{E}_{x \sim p_{data}(x)} [y \log C(x) + (1 - y) \log(1 - C(x))] \\
& -\mathbb{E}_{z \sim p_z(z)} [y \log C(G_N(z|y)) + (1 - y) \log(1 - C(G_N(z|y)))] \\
& -\mathbb{E}_{z \sim p_z(z)} [y \log C(G_B(z|y)) + (1 - y) \log(1 - C(G_B(z|y)))]
\end{aligned} \tag{3.6}$$

3.5 Stabilizing Training

We propose a combination of weighted random sampling and differentiable data augmentation to stabilize training. We first leverage a weighted random sampler to balance each mini-batch. The weighted random sampler adds weights to each sample in proportion to the class imbalance. For example, if there are 1,000 majority and 10 minority instances, the weights assigned to each majority and minority instances will be 1, and 100, respectively. In other words, the probability of sampling a majority instance on the first draw is 50% since the sum of weights assigned to majority and minority instances add up to 1,000 each, with a grand total of 2,000. Afterwards, we apply differentiable data augmentation proposed by Zhao et al. [12]. Although the discriminator receives a balanced ratio of majority and minority samples, the discriminator inevitably sees the same images repeatedly, eventuating in overfitting. To address this issue, we apply differentiable data augmentation comprised of image translations, cutout, and color changes to prevent the discriminator from overfitting to the limited pool of minority instances.

We describe the training procedure of SMS. For each mini-batch, we create a

balanced mini-batch using the weighted random sampler. Afterwards, we generate both normal and borderline samples and apply the augmentation techniques on both the real and generated samples. We update the discriminator by maximizing the loss function L_D in Equation 3.5. Then, we update the classifier by minimizing the loss function L_C in Equation 3.6. Lastly, we update the generators by minimizing the loss function L_G in Equation 3.4. This process repeats until termination.

Chapter 4

Experiments

We run experiments to answer the following questions:

- **Q1. Performance (Section 4.2).** How well does a classifier perform when trained on a dataset augmented by SMS? How effective is SMS when trained on datasets with rare objects?
- **Q2. Quality and diversity (Section 4.3).** How realistic and diverse are the synthetic minority samples generated by SMS?
- **Q3. Ablation study (Section 4.4).** Does the weight sharing strategy help improve accuracy? What impact does the diversity loss have on performance?

4.1 Experimental Settings

Dataset. For each dataset, we select two classes to be the majority and minority class. The FASHION MNIST [13] dataset is made up of gray-scaled images that do not follow the natural image distribution. We designate the trouser and coat as the majority and minority class, respectively. The CIFAR-10 [14] dataset contains natural images of various objects and animals. We designate the car and deer as the majority and minority class, respectively. The SVHN [15] dataset features natural images of street view house numbers. We designate the digits one and four as the majority and minority class, respectively. The WASTE classification dataset is a binary class dataset featuring natural images of organic and recyclable objects. Because the original im-

Table 1: Dataset summary.

Dataset	Dimensions	Majority Class		Minority Class	
		Training	Test	Training	Test
FASHION MNIST ¹	$1 \times 28 \times 28$	6,000	1,000	30	200
CIFAR-10 ²	$3 \times 32 \times 32$	5,000	1,000	25	200
SVHN ³	$3 \times 32 \times 32$	5,000	1,000	25	200
WASTE ⁴	$3 \times 32 \times 32$	5,000	1,000	25	200

ages vary in size, we resize all images to $3 \times 32 \times 32$ before feeding the training data to each model. We designate the organic objects as the majority class, and the recyclable objects as the minority class. We enhance the difficulty of the task by enforcing an imbalance ratio of 200:1 on all datasets. For example, in the FASHION MNIST dataset, 30 samples in the minority class are selected and combined with 6,000 majority instances to create an imbalanced dataset. Unless otherwise specified, all experiments are conducted on the datasets as specified in Table 1.

Competitor. We compare the performance of SMS to the following competitors.

- **Random.** This method randomly selects existing minority instances for over-sampling.
- **SMOTE [1].** Until the dataset is balanced, for each minority point, SMOTE performs linear interpolation between the point and its randomly selected k-nearest minority class neighbors.
- **CDCGAN [2].** CDCGAN utilizes convolutional strides and transposed convolutions for downsampling and upsampling, respectively.
- **ACGAN [3].** ACGAN leverages an auxiliary classifier in the adversarial train-

¹<https://github.com/zalando-research/fashion-mnist>

²<https://www.cs.toronto.edu/~kriz/cifar.html>

³<http://ufldl.stanford.edu/housenumbers>

⁴<https://www.kaggle.com/techsash/waste-classification-data>

ing process to generate high-quality samples for each class. During the adversarial training process, the discriminator and the generator work together to create samples that the auxiliary classifier can classify with ease.

- **GAMO** [4]. GAMO is a minority oversampling framework that learns an intermediate representation of the minority class via a three-way adversarial training between the classifier, discriminator, and generator. After training, it learns to project the intermediate representations sampled by the generator back onto the data space to create synthetic samples. We use the parameter settings specified in the GAMO paper to train the models.

Model training. All methods are trained on a Dell PowerEdge T630 INTEL Zeon E5-2630 2.2GHz server with Geforce GTX 1080Ti GPUs. We train each of our methods for 1200 epochs with the exception of GAMO, which is trained according to the specifications provided in the supplementary materials of the GAMO paper.

Hyperparameters. With the exception of GAMO, we apply weighted random sampling and data augmentation described in Section 3.5 to each GAN-based method to prevent immediate mode collapse. We discovered that applying the aforementioned technique to GAMO deteriorates performance. We train SMS by setting the weight of the class-conditional diversity loss term λ to 1. The mini-batch size is set to 16 for all datasets. The Adam optimizer [16] is used to train all models. We set the learning rate of the discriminator, generator, and classifier in SMS to 0.0002, 0.0001, and 0.0002. We apply weight decay of 10^{-5} to the classifier to avoid overfitting.

Evaluation. We train each method on the imbalanced dataset $data_{real}$. After training, we oversample minority instances to create $data_{synthetic}$. The size of $data_{synthetic}$ is equal to the difference between the number of majority and minority instances in $data_{real}$. We evaluate the effectiveness of the synthetic minority samples

by training a classifier on $data_{real} + data_{synthetic}$. Because the severe data imbalance disqualifies accuracy as a suitable evaluation metric, we use the precision, recall, and F1-score to evaluate the performance of the classifier. The evaluation model is a CNN with two convolutional layers followed by max pooling and two dense layers. Evaluation is performed on test data $data_{test}$ according to the specifications in Table 1.

Table 2: The precision, recall and F1-score of classifiers on the test dataset $data_{test}$ after training on augmented dataset $data_{synthetic}$. SMS shows the best performance across all datasets. Bold text indicates the best results.

Datasets	Metrics	Random	SMOTE	CDCGAN	ACGAN	GAMO	SMS (proposed)
Fashion MNIST	Precision	0.9639	0.9651	0.9743	0.9663	0.9753	0.9851
	Recall	0.9285	0.9385	0.9290	0.9485	0.9390	0.9635
	F1-score	0.9459	0.9516	0.9511	0.9573	0.9568	0.9742
SVHN	Precision	0.9035	0.8684	0.8809	0.8848	0.8850	0.9315
	Recall	0.6160	0.6140	0.5905	0.6440	0.5785	0.7330
	F1-score	0.7325	0.7194	0.7070	0.7454	0.6997	0.8204
CIFAR-10	Precision	0.9412	0.9402	0.9421	0.9440	0.9344	0.9482
	Recall	0.7045	0.6995	0.6725	0.7195	0.6225	0.7420
	F1-score	0.8058	0.8022	0.7848	0.8166	0.7472	0.8325
WASTE	Precision	0.8805	0.9252	0.9284	0.8983	0.8756	0.9299
	Recall	0.5515	0.5600	0.5825	0.5470	0.5465	0.5925
	F1-score	0.6977	0.9034	0.7159	0.6800	0.6730	0.7238

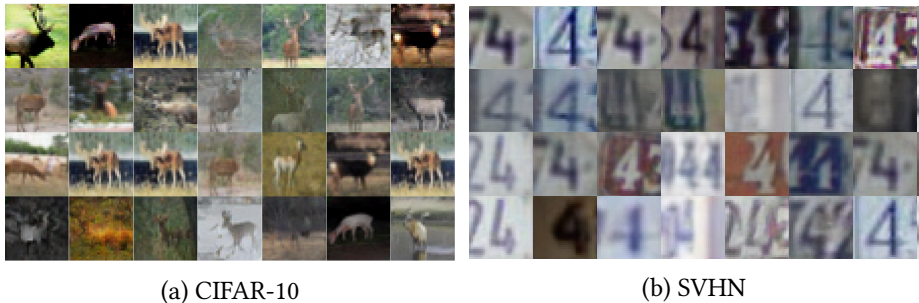


Figure 4: Visualization of synthetic samples generated by ACGAN, CDCGAN, GAMO, and SMS (from top to bottom). (a) SMS creates realistic and diverse images despite the lack of minority instances. (b) SMS generates diverse samples containing the key features of the digit 4 while retaining large variations of noisy features for training a robust classifier.

4.2 Performance

Standard evaluation. We evaluate the performance of classifiers trained with real and synthetic minority samples generated by SMS and competitors. Table 2 and Figure 1 shows the precision, recall, and F1-score of classifiers trained on datasets augmented by SMS and its competitors. SMS helps the classifier achieve the highest precision, recall, and F1-score compared to other classifiers trained using real and synthetic minority samples generated by competitors for all datasets. The F1-scores of the classifier trained with the dataset augmented by SMS are 1.77%, 10.06%, 1.95%, and 1.10% higher than that of the second-best performing method on FASHION MNIST, SVHN, CIFAR-10, and the WASTE classification dataset, respectively. Moreover, there are large performance gaps between SMS and its competitors on SVHN in Figure 1c. We observe in Figure 4b that SMS generates samples with both the key features of the minority class and a diverse range of noise present in the original SVHN dataset, which helps classifiers filter out the noise and focus on the important features for

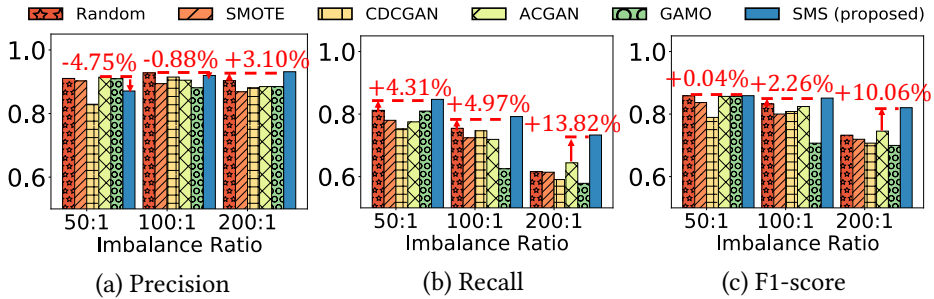


Figure 5: Classifiers trained using real and synthetic minority samples generated by SMS achieves the best performance on the severely imbalanced SVHN dataset. The performance gaps between the classifiers trained on samples generated by SMS and competitors are the largest when the ratio is 200:1.

identifying minority instances. We note that SVHN contains noisy features that may distract networks from the key salient features such as the digit 2 next to the digit 4 on the bottom left hand image of Figure 4b. By generating samples with both high and low degrees of class confidence, SMS creates samples that help train a robust and unbiased classifier compared to its competitors.

Next, we evaluate the performance of classifiers trained using real and synthetic minority samples generated by SMS and competitors under the imbalanced ratios 50:1, 100:1, and 200:1 on SVHN. Figure 5 shows that the classifier trained using real and synthetic minority samples generated by SMS achieves the highest F1-score for all three imbalance ratios. The performance gaps are the largest when the ratio is 200:1; SMS helps the classifier achieve up to 3.10%, 13.82%, and 10.06% higher precision, recall, and F1-score compared to the second-best performing method. These results indicate that SMS helps train a robust classifier even on severely imbalanced datasets.

Evaluation with rare objects. To further assess the ability of SMS to generate synthetic samples on highly imbalanced datasets containing rare objects, we conduct

Table 3: Classification performance of evaluation model on CIFAR-10 and SVHN when trained on augmented dataset generated by SMS and its competitors. Each method is trained on a dataset with an imbalance ratio of 400:1.

Method	SVHN			CIFAR-10		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Random	0.9198	0.5225	0.6664	0.9160	0.5895	0.7173
SMOTE	0.8684	0.5195	0.6501	0.9220	0.6120	0.7357
DC-GAN	0.9205	0.5275	0.6707	0.9322	0.6075	0.7356
AC-GAN	0.9198	0.5225	0.6664	0.9296	0.5900	0.7218
GAMO	0.9177	0.5075	0.6536	0.9274	0.5750	0.7099
SMS	0.9212	0.5425	0.6833	0.9394	0.6550	0.7718

additional experiments by reducing the number of majority and minority instances in SVHN and CIFAR-10 to 4800, and 12, respectively. This creates two new datasets with an imbalance ratio of 400:1. Table 3 shows that SMS achieves the highest F1-score even on the severely imbalanced datasets. These results indicate that SMS generates better quality samples for learning an unbiased, general class representation compared to its competitors even on severely imbalanced datasets where real minority instances are extremely scarce.

We additionally evaluate the performance of SMS on the WASTE classification dataset. We note that the minority class instances in the WASTE classification dataset come in many different forms as shown in Figure 6. Thus, the key features of the minority class cannot be easily captured due to the large variations between samples of the same class. This makes the task more challenging in comparison to the other three datasets, where each class instance possesses common salient features. For example, in SVHN, instances of the digit 1 have common salient features that clearly distinguishes the instance from an instance belonging to the class digit 4. We note that the recall and F1-score of the classifier trained with the dataset augmented by

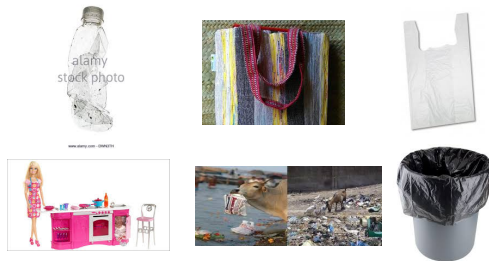


Figure 6: Examples of minority class instances (recyclable objects) in the WASTE classification dataset. Despite belonging to the same class, each object has different key salient features, making it difficult to identify common features defining the minority class.

SMS are 1.72%, and 1.10% higher than that of the second-best performing method on the WASTE classification dataset as shown in Figure 1d. This demonstrates that SMS performs better than its competitors even on more challenging datasets containing rare objects.

4.3 Synthetic Image Quality and Diversity

We compute the Fréchet Inception Distance [17] scores for each method to evaluate the quality and diversity of the generated samples. The experimental results in Table 4 show that SMS has the lowest FID-score, indicating that SMS generates the most diverse and highest-quality images. To further examine the diversity and quality of generated images, we analyze the images generated by each framework. We observe on the last row of Figure 4b that SMS generates the most diverse range of samples including borderline features (see third image from the right on the last row). Note that other competitors show large numbers of duplicates, which is evident from observing images generated from SVHN and CIFAR-10 as shown in Figure 4a. We omit the FID-scores for the WASTE classification dataset due to the resizing require-

Table 4: FID score results. Lower values indicate better image quality and diversity. The best scores are highlighted in bold. SMS_{-W} , SMS_{-D} are variants of SMS without the weight sharing and diversity loss components.

Method	FASHION MNIST	CIFAR-10	SVHN
Random	32.3	37.3	30.2
SMOTE	39.4	34.9	28.3
DC-GAN	14.9	31.0	15.9
AC-GAN	12.7	20.8	17.2
GAMO	28.4	30.3	15.2
SMS_{-W}	17.0	20.5	18.5
SMS_{-D}	15.9	22.6	20.3
SMS	12.4	17.6	14.7

ments, which impacts the accuracy of the FID-score. The input images must have the same size and dimensions during the training and testing phase. Because the original dataset contains images of varying size and dimensions, the images are resized to ensure that the dataset contains images of fixed size and dimension. Because the generated samples are compared against the resized images, the calculated FID-score will not accurately reflect the quality and diversity of the generated samples.

4.4 Ablation Study

Effect of Weight Sharing on Performance. SMS leverages weight sharing to learn common features between normal and borderline samples. To verify the performance enhancing properties of weight sharing, we compare SMS to SMS_{-W} , which does not use weight sharing.

Rows of SMS_{-W} and SMS of Table 5 show the improvements of SMS via weight sharing. We note that the classifiers trained using real and synthetic minority samples generated by SMS outperforms that of SMS_{-W} with an average precision, recall, and

Table 5: Ablation studies. SMS outperforms both 1) SMS_{-W} a variant of SMS without weight sharing, and 2) SMS_{-D}, a variant of SMS without the class-conditional diversity loss term.

Method	Metric	FASHION MNIST	SVHN	CIFAR-10	WASTE
SMS	Precision	0.9851	0.9315	0.9482	0.9299
	Recall	0.9635	0.7330	0.7420	0.5925
	F1-score	0.9742	0.8204	0.8325	0.7238
SMS _{-W}	Precision	0.9743	0.9282	0.9376	0.9120
	Recall	0.9290	0.6985	0.7385	0.5770
	F1-score	0.9511	0.7971	0.8262	0.7068
SMS _{-D}	Precision	0.9276	0.9315	0.9330	0.8916
	Recall	0.9485	0.6305	0.6940	0.5665
	F1-score	0.9379	0.7417	0.7959	0.6928

F1-score improvement of 1.14%, 2.95%, and 2.13% across all four datasets. The consistent results demonstrate that weight sharing does help the generators perform better by jointly learning the common features between borderline and normal instances. To further validate the performance difference, we analyze the FID-score of generated samples from each method. We note that the FID-score decreases by an average of 20.53%, which indicates that the generated samples are more diverse and realistic with weight sharing. This shows that weight sharing helps improve the quality of the generated samples.

Effect of Class-conditional Diversity Loss on Performance. We verify the effects of the class-conditional diversity loss term on the overall performance of SMS.

Rows of SMS_{-D} and SMS of Table 5 show the accuracy improvement due to the diversity loss term. Precision, recall, and F1-score of the classifier trained using real and synthetic minority samples generated by SMS increase by 2.42%, 7.34%, and 5.89% on average across all four datasets. This demonstrates that the diversity loss

term has improved the quality of the dataset by encouraging the generators to create diverse samples. We validate the diversity of the generated samples by observing the FID scores.

Rows of SMS_{-D} and SMS in Table 4 show the FID score improvement due to the diversity loss term. We note that the FID score decreases by an average of 23.77%. The results show that the diversity loss term improves both the quality and diversity of the generated samples.

Chapter 5

Conclusion

We propose SMS, an oversampling framework for sampling high-fidelity minority instances. We observe that existing deep-learning based methods overfit or in worst-cases, immediately collapse when trained on highly imbalanced data. Moreover, we note that existing methods either create borderline or safe samples but not both. Based on observations, we introduced an oversampling framework that jointly learns the common and borderline features to generate a more well-rounded synthetic dataset. Extensive experiments show that SMS generates the most diverse, classifier-friendly synthetic dataset. Future works include extending the method to generalize well to multi-class datasets.

References

- [1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, 2002.
- [2] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *ICLR*, 2016.
- [3] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” in *ICML*, 2017.
- [4] S. S. Mullick, S. Datta, and S. Das, “Generative adversarial minority oversampling,” in *ICCV*, 2019.
- [5] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: adaptive synthetic sampling approach for imbalanced learning,” in *IJCNN*, 2008.
- [6] F. Last, G. Douzas, and F. Bação, “Oversampling for imbalanced learning based on k-means and SMOTE,” *CoRR*, 2017.
- [7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” in *NeurIPS*, 2014.
- [8] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *ICLR*, 2014.
- [9] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014.
- [10] S. Ando and C. Huang, “Deep over-sampling framework for classifying imbalanced data,” in *ECML PKDD 2017*, 2017.
- [11] J. Yoo, M. Cho, T. Kim, and U. Kang, “Knowledge extraction with no observable data,” in *NeurIPS*, 2019.
- [12] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han, “Differentiable augmentation for data-efficient gan training,” in *NeurIPS*, 2020.

- [13] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” 2017.
- [14] A. Krizhevsky, V. Nair, and G. Hinton, “Cifar-10 (canadian institute for advanced research),”
- [15] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.
- [16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
- [17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter

요 약

클래스가 불균형한 데이터가 주어졌을 때 어떻게 소수 클래스에 대한 데이터를 인공적으로 증대하여 클래스 분류 성능을 높일 수 있을까? 데이터 불균형 문제는 고장 진단 및 질병 분류와 같이 한쪽의 클래스 수가 다른 한쪽의 수보다 극단적으로 적을 때 발생하는 문제를 일컫는다. 이러한 불균형한 데이터를 통해 학습된 모델은 잘못된 예측 결과 좋지 못한 분류 성능을 보인다. 이를 해결하기 위해 일반적으로 소수 클래스에 대해 인공적으로 샘플을 증대하여 각 클래스의 샘플의 수를 동일하게 하는 방식을 사용한다. 인공적으로 증대가 된 샘플은 사실적이고 기존의 샘플과 동일하지 않아야 하며 다양한 성질을 포함하여야 하는데 선행 연구들은 이러한 요소를 충족하지 못하고 있다.

해당 논문에서는 불균형한 데이터셋에서 높은 품질의 인공 데이터를 오버샘플링 (oversampling) 하는 프레임워크인 Synthetic Minority Sampler (SMS)를 제안한다. SMS 는 두 개의 생성기를 사용하여 구분이 명확한 샘플과 명확하지 않은 샘플을 적절한 비율로 생성하고 이를 통해 분류기를 더욱 견고하고 일반화된 방향으로 학습시킨다. SMS 는 해당 논문에서 고안된 손실 함수 (class-conditional diversity loss) 를 사용하여 인공적으로 생성된 소수 클래스 샘플의 다양성을 보장한다. 또한 미니 배치의 클래스 비율을 적절하게 배분하는 임의 샘플러와 구분기 (discriminator)의 오버피팅 방지를 위한 데이터 증강 기법을 사용하여 SMS 의 학습을 안정화한다. 실험 결과에서는 SMS 를 통해 생성된 인공 데이터를 기존의 데이터셋에 추가하여 학습한 모델이 이진 분류 (binary classification) 문제에서 탁월한 성능을 보였으며, 경쟁 메소드보다 10.06% 높은 F1 스코어를 기록하였다.

주요어: 생산적 적대 신경망, 불균형한 데이터, 오버샘플링

학번 : 2019-25223