Information Systems Review
Vol. 24, No. 4 November 2022 http://dx.doi.org/10.14329/isr.2022.24.4.097

불균형 정형 데이터를 위한 SMOTE와 변형 CycleGAN 기반 하이브리드 오버샘플링 기법

A Hybrid Oversampling Technique for Imbalanced Structured Data based on SMOTE and Adapted CycleGAN

노 정 담 (Jung-Dam Noh) Afreeca TV VOD 데이터 팀 주니어
최 병 구 (Byounggu Choi) 국민대학교 경영대학 AI빅데이터융합경영학과 교수, 교신저자

요 약

이미지와 같은 비정형 데이터의 불균형 클래스 문제 해결에 있어 생산적 적대 신경망(generative adversarial network)에 기반한 오버샘플링 기법의 우수성이 알려짐에 따라 다양한 연구들이 이를 정형 데이터의 불균형 문제 해결에도 적용하기 시작하였다. 그러나 이러한 연구들은 데이터의 형태를 비정형 데이터 구조로 변경함으로써 정형 데이터의 특징을 정확하게 반영하지 못한다는 점이 문제로 지적되고 있다. 본 연구에서는 이를 해결하기 위해 순환 생산적 적대 신경망(cycle GAN)을 정형 데이터의 구조에 맞게 재구성하고 이를 SMOTE(synthetic minority oversampling technique) 기법과 결합한 하이브리드 오버샘플링 기법을 제안하였다. 특히 기존 연구와 달리 생산적 적대 신경망을 구성함에 있어 1차원 합성곱 신경망(1D-convolutional neural network)을 사용함으로써 기존 연구의 한계를 극복하고자 하였다. 본 연구에서 제안한 기법의 성능 비교를 위해 불균형 정형 데이터를 기반으로 오버샘플링을 진행하고 그 결과를 SMOTE, ADASYN(adaptive synthetic sampling) 등과 같은 기존 기법과 비교하였다. 비교 결과 차원이 많을수록, 불균형 정도가 심할수록 제안된 모형이 우수한 성능을 보이는 것으로 나타났다. 본 연구는 기존 연구와 달리 정형 데이터의 구조를 유지하면서 소수 클래스의 특징을 반영한 오버샘플링을 통해 분류의 성능을 향상시켰다는 점에서 의의가 있다.

키워드 : 불균형 데이터, 오버샘플링, 순환 생산적 적대 신경망, SMOTE 기법, 정형 데이터

Ⅰ. 서 론

정보기술의 급격한 발전을 기반으로 디지털 시대가 도래함에 따라 데이터의 양이 기하급수적으로 증가하고 있다. 2020년 말 전 세계 데이터의 총량은 64.2 제타바이트(zettabytes)에 이르렀으며 2025년까지 180 제타바이트를 넘어설 것으로 예측되고 있다(Statista, 2021). 스마트폰의 사용 증가와 코로나-19로 인한 페이스북, 구글, 아마존 등과 같은 디지털 미디어 플랫폼의 사용 증가로 인해 이러한 추세는 2022년에도 계속되어 매일 2,500조(quintillion) 바이트의 데이터가 생성될 것으로 예측되고 있다(Wise, 2022). 이처럼 데이터가 급격하게 증가함에 따라 이를 기반으로 복잡한 현실 문

제를 해결하기 위한 다양한 머신러닝(machine learning) 기법들이 개발되어 왔으며 많은 분야에서 뛰어난 성능 개선을 보이고 있다. 예를 들면, 바둑 경기에 있어 인공지능이 세계 챔피언을 이기는 성과를 보이고 있으며, 이미지 인식에 있어 전문가를 뛰어넘는 수준을 보이고 있으며, 자연어 처리에 있어 유의미한 성능 개선을 보이고 있다(Krizhevsky *et al.*, 2017; Silver *et al.*, 2018).

머신러닝 성능 개선을 위해서는 문제 해결을 위한 알고리즘 뿐만 아니라 데이터의 품질 역시 매우 중요하다(Sambasivan *et al.*, 2021). IBM (2020)에 따르면 미국 시장에서만 나쁜 품질의 데이터로 인해 발생하는 비용이 연간 3조 달러 이상일 것으로 추산하고 있다. 나아가 전 세계 금융 기관을 대상으로 한 조사에 따르면 응답자의 43%가 인공지능 도입의 가장 큰 장애 요인으로 나쁜 데이터 품질을 지목하고 있다(Refinitive, 2019). 그러나 이러한 데이터 품질의 중요성에도 불구하고 대부분의 기존 연구들은 알고리즘 개선에만 초점을 두고 있을 뿐 데이터 품질에는 상대적으로 관심이 부족했던 것도 사실이다(Chen *et al.*, 2021).

데이터 품질은 데이터 중복(redundancy), 노이즈(noise), 결측값(missing values), 불일치(inconsistency) 등 다양한 요인과 관련이 있다(Aydilek *et al.*, 2013; Bosu and MacDonell, 2013; Chandola *et al.*, 2009). 이 가운데 불균형 데이터(imbalanced data)는 데이터의 품질을 저해하는 가장 일반적이며 중요한 요인 가운데 하나로 언급되고 있다(Krawczyk, 2016). 불균형 데이터는 문제 해결을 위해 데이터를 서로 다른 클래스(classes)로 분류할 때, 클래스를 구성하는 특정 사건이나 개념이 매우 드물게 발생하거나 특정 클래스에 대한 데이터 수집의 제약으로 인해 소수의 특정 클래스에 지나치게 많은 데이터가 포함되는 경우 발생하게 된다(Fernández *et al.*, 2017; Thejas *et al.*, 2022). 이러한 불균형 데이터는 의료 진단, 텍스트 분류, 침입 탐지, 클릭 사기 탐지 등 다양한 응용 분야에서 빈번하게 발생하고 있다(Liu *et al.*, 2009; Tek *et al.*, 2010). 불균형 데이터에 기반한 예측은 필연적으로 다수의 데이터를 포함하고 있는 클래스에 편향된(biased) 결과를 보이게 되어 예측의 성능을 저해하게 된다. 따라서 이러한 불균형 데이터를 해결하는 것은 머신러닝 성능 향상을 위한 매우 중요한 연구과제이다.

이에 따라 불균형 데이터 문제를 해결하기 위한 다양한 방법이 데이터와 알고리즘 수준에서 이루어져 왔다(Yap *et al.*, 2013). 특히 데이터 수준에서 해결 방법인 데이터 샘플링 기법은 사용되는 클래스 분류기(classifier)와 독립적이기 때문에 가장 널리 활용되고 있다(Fernández *et al.*, 2017). 데이터 샘플링 기법은 다수 클래스 데이터와 소수 클래스 데이터의 개수를 조정하여 불균형 정도를 해소하는 방법으로(Khoshgoftaar *et al.*, 2015), 두 클래스 중 어떤 클래스의 데이터 개수를 조정 하느냐에 따라 언더샘플링(undersampling)과 오버샘플링 (oversampling)으로 구분된다. 언더샘플링은 소수 클래스의 데이터 수에 맞게 다수 클래스의 데이터를 제거하는 기법이며 랜덤(random) 언더샘플링, 토멕링크(Tomek's link), ENN(edited nearest neighbors) 등이 있다(Tomek, 1976; Wilson, 1972). 오버샘플링은 다수 클래스의 데이터 수에 맞게 소수 클래스 데이터를 인공적으로 만들어 내는 기법을 말하며 랜덤 오버샘플링, SMOTE(synthetic minority oversampling technique), ADASYN(adaptive synthetic sampling) 등이 있다(Chawla *et al.*, 2002; He *et al.*, 2008).

의료 진단이나 사기 탐지와 같은 현실 문제에 있어 다수 클래스를 잘못 예측하여 발생하는 비용보다 소수 클래스를 잘못 예측하여 발생하는 비용이 대부분 더 크다. 이러한 이유로 인해 많은 연구자는 언더샘플링에 비해 오버샘플링이 더 좋은 성과를 보인다고 주장하고 있으며(Fernández *et al.*, 2017; Mohammed *et al.*, 2020), 지금까지 다양한 방식의 오버샘플링 기법들이 제안되어 왔다(Cao and Wang, 2011; Zhou *et al.*, 2013). 이러한 기법들 가운데 SMOTE 기법은 가장 널리 활용되어 왔으며

현실 세계의 불균형 데이터 문제를 어느 정도 해결하였다는 평가를 받고 있다(Chawla *et al.*, 2002). 그러나 SMOTE와 같은 통계 기반 오버샘플링 방식은 기존 데이터를 기반으로 새로운 데이터를 생성하기 때문에 생성된 데이터의 다양성이 부족하고 이로 인해 분류 모델의 과적합(overfitting) 문제를 야기한다는 단점이 지적되고 있다(Soltanzadeh and Hashemzadeh, 2020).

최근 딥러닝(deep learning) 기법이 발전함에 따라 기존 통계 기반 오버샘플링 기법들의 단점을 보완하기 위하여 딥러닝에 기반을 둔 variational auto encoding(VAE), glow model, deep sampling model 등과 같은 다양한 오버샘플링 기법들이 등장하고 있다(Kingma and Dhariwal, 2018). 예를 들면, Fangyu *et al.*(2021)은 딥러닝 기반 VAE를 변형하여 불균형 데이터의 문제를 해결하고자 시도하였다. 이러한 딥러닝 기반 오버샘플링 기법 가운데 생산적 적대 신경망(GAN: generative adversarial networks)에 기반한 오버샘플링 기법이 가장 좋은 성능을 보이는 것으로 알려져 있다(Douzas and Bacao, 2018). GAN은 데이터의 분포를 학습하여 새로운 데이터를 학습된 분포에 맞게 생성하는 기법으로 이미지 데이터와 같은 비정형 데이터를 다룰 때 주로 사용되어 왔으며 deep convolutional GAN(DCGAN)(Radford *et al.*, 2016), Wasserstein GAN(WGAN)(Arjovsky *et al.*, 2017), conditional GAN(CGAN)(Mirza and Osindero, 2014), CycleGAN (Zhu *et al.*, 2017) 등과 같이 GAN을 개선한 다양한 모델이 제안되고 있다.

GAN이 높은 성능을 보이자 최근 들어 몇몇 연구자들이 정형 데이터의 불균형 문제 해결에도 이를 활용하기 시작하였다(Quintana and Miller, 2019). 예를 들면, Ba(2019)는 GAN 기반 오버샘플링 기법을 활용하여 신용카드 사기 거래 분류를 시도하였으며 Engelmann and Lessmann(2021)은 GAN의 확장 모델인 CWGAN(conditional Wasserstein GAN)을 기반으로 오버샘플링을 수행하고 이를 통해 대출 신청자에 대한 부도 위험 가능성을 예측하고자 하였다. 그러나 정형 데이터의 불균형 문제 해결을 위해 GAN을 활용한 연구들은 데이터의 형태를 비정형 데이터의 구조인 2차원(가로×세로) 구조로 변형시켜 오버샘플링을 진행하였다는 점에서 문제가 있다. 이는 GAN이 이미지 데이터와 같은 비정형 데이터의 분석을 위해 제안되었다는 점에서 기인한다. 그러나 데이터의 구조를 비정형 데이터 구조로 변경함으로써 정형 데이터의 특징이 소실되어 현실 세계를 정확하게 반영할 수 없기 때문에 이를 해결할 필요가 있다. 예를 들면, 이미지 데이터에서 픽셀 값은 최소-최대 변환(min-max transform)을 사용함으로써 -1부터 1까지의 값을 갖는 가우시안 분포(Gaussian distribution)를 따른다. 그러나 정형 데이터의 연속 변수는 가우시안이 아닌 긴 꼬리를 갖는 분포를 갖기 때문에 생성된 값은 0을 중심으로 하지 않는다(Xu, 2020). 정형 데이터를 비정형 데이터의 구조인 2차원으로 변형할 경우 이러한 정형 데이터의 특징이 손실될 수 있다. 또한 정형 데이터의 경우 연속형과 이산형 데이터가 혼합된 경우가 대부분이기 때문에 데이터 생성 시 이를 반영하여 연속형과 이산형이 혼합된 데이터를 생성할 수 있어야 하지만 데이터의 구조를 변형하면 이러한 데이터의 혼합 생성이 어려워진다(Xu, 2020).

이러한 기존 연구의 한계점을 극복하기 위해 본 연구에서는 GAN 기법 가운데 하나인 순환 생산적 적대 신경망(CycleGAN)[1]을 이용하여 소수 클래스 데이터의 특징을 추출하고 이를 SMOTE 기반 오버샘플링 통해 생성된 데이터와 합성하고자 한다. 본 연구에서 제안한 기법은 오버샘플링 과정에서 정형 데이터의 특징을 유지하게 함으로써 현실 세계의 특징을 보다 정확하게 반영한 데이터를 생성할 수 있다. 나아가 2차원 구조를 갖는 기존 GAN 기법들과 다르게 1차원 합성곱 신경망

1) 순환 생산적 적대 신경망은 데이터의 특성을 추출하여 학습시킬 수 있기 때문에 오버샘플링 시 소수 클래스 데이터의 특징을 반영함으로써 데이터의 분포 뿐 아니라 특성을 반영한 데이터를 생성할 수 있다(Zhu *et al.*, 2017).

(1D-convolution network)을 사용함으로써 정형 데이터의 형태에 맞는 새로운 모형을 제안하고자 한다. 이를 통해 정형 데이터 분석에 있어 소수 클래스 데이터의 특징을 보다 명확하게 추출할 수 있게 함으로써 기존 연구의 한계를 극복하고자 한다. 나아가 제안된 모형을 신용카드 사기 거래 탐지 분야에 적용함으로써 이의 유용성을 검증하고자 한다. 본 연구는 기존 연구와 달리 정형 데이터의 구조를 유지하면서 소수 클래스의 특징을 반영한 오버샘플링을 통해 분류의 성능을 향상시켰다는 점에서 기존 연구와 차별점이 있다.

본 연구는 다음과 같이 구성된다. 다음 장에서는 오버샘플링 관련 기존 연구를 요약하고 본 연구에서 제안하는 모형의 기반이 되는 CycleGAN을 소개한다. 제Ⅲ장에서는 기존 오버샘플링 방식의 문제점 해결을 위해 본 연구에서 제안하는 모형의 구조를 설명한다. 제Ⅳ장에서는 제안된 모형의 성능을 검증하기 위한 실험 데이터, 실험 설정, 실험 결과를 기술한다. 마지막으로 본 연구의 시사점과 한계점을 서술하고 향후 연구과제를 제안한다.

Ⅱ. 선행 연구

2.1 오버샘플링 기법

오버샘플링은 소수 클래스의 데이터를 복제하거나 생성함으로써 다수 클래스의 데이터와 균형을 이룸으로써 불균형 데이터 문제를 해소하고자 하는 방법을 의미한다(Mohammed *et al.*, 2020). 오버샘플링을 위한 가장 단순하지만 강건한(robust) 기법은 랜덤 오버샘플링이다(Ling and Li, 1998). 랜덤 오버샘플링은 무작위로 소수 클래스의 데이터를 선택하고 이를 반복적으로 사용하여 새로운 데이터를 생성하는 방법으로 가중치를 증가시키는 효과가 있다. 그러나 이 기법으로 생성된 데이터는 기존 데이터를 복제한 것이기 때문에 필연적으로 과적합 문제를 야기시킨다(Chawla *et al.*, 2004). 이러한 문제를 해결하기 위해 Chawla *et al.*(2002)은 SMOTE라는 기법을 제안하였다. SMOTE 기법은 소수 클래스로부터 랜덤하게 선택된 데이터를 중심으로 k개의 최근접 이웃을 찾고 이들 간의 선형 연결 구조 사이에 새로운 데이터를 합성하는 방식을 말한다. 이 기법은 여러 응용 분야에서 적용되었으며 데이터 불균형을 해결하는 데 우수한 성과를 보여왔다. 그러나 소수 클래스를 기반으로 새로운 데이터를 생성하기 때문에 생성된 데이터가 소수 클래스의 특성만을 반영하고 데이터 노이즈에 취약하다는 점이 그 한계로 지적되어 왔다(He and Garcia, 2009).

이러한 한계를 극복하기 위해 100여 개가 넘는 SMOTE의 다양한 변형 기법들이 제안되어 왔다. 예를 들면, Han *et al.*(2005)은 소수 클래스 데이터와 다수 클래스 데이터 간의 경계를 기준으로 SMOTE를 통해 새로운 데이터를 생성하는 방식인 Borderline-SMOTE를 제안하였다. He *et al.*(2008)은 소수 클래스 데이터를 생성함에 있어 소수 클래스 데이터 주변의 다수 클래스 데이터 밀도에 따라 가중치를 부여하는 기법인 ADASYN을 제안하였다. 이외에도 Cao and Wang(2011)은 데이터의 밀도와 분포 정보를 기반으로 소수 클래스 데이터를 생성하는 SMOBD(synthetic minority over-sampling based on samples density) 기법을, Zhou *et al.*(2013)은 데이터 내부 및 주변에 로컬 선형 파티션(local linear partitions)을 생성하고 각 파티션 별로 SMOTE를 적용하여 데이터를 생성하는 assembled SMOTE 기법을 제안하는 등을 제안하였다. 그러나 이러한 SMOTE 기반 오버샘플링 기법은 데이터를 완전히 새로 만드는 방식이 아니라 기존 데이터를 기반으로 새로운 데이터를 생성하기 때문에 생성된 데이터의 다양성이 부족하고 분류 모델에서 과적합 문제를 야기한다는 단점을 완전히 해소하지 못하고 있다(Cao and Wang, 2011). 나아가 소수 클래스 데이터의 밀도가 낮은 경우 다수 클래스에 속하는 데이터를 소수 클래스로 생성하는 등의 노이즈를 완벽하게 제거할 수 없어 분류 모델의 정확도가 낮다는 점이 문제로 지적되고 있다(Islam *et al.*, 2022).

〈표 1〉 GAN과 이의 변형 모형의 장단점

저자	GAN 유형	장점	단점
Goodfellow *et al.*(2014)	GAN (generative adversarial network)	• 빠른 데이터 생성 속도 • 고르지 않은 (sharp) 확률분포 처리	• 기울기 소실 • 모드 붕괴 • 생성된 이미지 해상도 낮음
Mirza and Osindero(2014)	CGAN (conditional GAN)	• 원하는 클래스를 포함한 데이터 생성 • 현실적인 이미지 생성	• 높은 연산 비용 • 적절한 조건 생성의 어려움
Radford *et al.*(2016)	DCGAN (deep convolutional GAN)	• 안정적인 학습 • 개선된 이미지 질	• 제한된 종류의 데이터 생성 • 하이퍼파라미터 선택에 민감 • 생성기와 판별기의 불균형으로 인한 과적합
Arjovsky *et al.*(2017)	WGAN (Wasserstein GAN)	• 기울기 소실 문제 해결 • 개선된 이미지 질 • 모드 붕괴 문제 해결	• 가중치에 따라 기울기 소실 문제 유발 • 복잡한 모델 기능 제한 • 수렴이 용이하지 않음
Zhu *et al.*(2017)	CycleGAN (cycle GAN)	• 대립하는 클래스 없이 데이터 생성 • 현실적인 이미지 생성 • 상대적으로 적은 학습 데이터	• 기하학적 변화를 갖는 데이터 생성에 취약 • 생성된 데이터의 변화가 작음

규모, 다양성, 생성 속도로 특징지어지는 빅데이터는 불균형 데이터의 문제점을 더욱 극명하게 드러내고 있다(Leevy *et al.*, 2018). 이에 따라 많은 연구가 빅데이터 처리를 위해 제안된 다양한 딥러닝 기법을 활용하여 불균형 데이터 문제를 해결하고자 하였다(Johnson and Khoshgoftaar, 2019). 특히 GAN에 기반한 오버샘플링 연구는 많은 연구자로부터 불균형 데이터 문제 해결을 위한 중요한 시도로 주목받고 있다. GAN은 실제 데이터와 유사한 가상 데이터를 생성하려는 생성기(generator)와 실제 데이터와 생성된 데이터를 구별하려는 판별기(discriminator) 간의 적대적 게임을 기반으로 제안되었다(Goodfellow *et al.*, 2014). 이러한 적대 학습(adversarial learning)을 기반으로 개발된 GAN은 원본데이터와 유사한 가상 데이터를 만들 수 있기 때문에 오버샘플링 연구에 쉽게 적용할 수 있었을 뿐 아니라 그 성능 면에서도 매우 우수한 것으로 판명되었다(Douzas and Bacao, 2018). 그러나 GAN은 생성기와 판별기의 불균등한 성능, 불안정한 수렴, 출력 데이터의 제어 불가능, 대립하는 클래스 데이터를 구하기 어려운 점 등과 같은 문제점이 한계로 지적되어 왔다(Nazari and Branco, 2021). 이러한 문제들을 해결하기 위해 GAN을 개선한 다양한 기법들이 제안되어 왔다.[2] GAN과 이의 변형 모형의 장단점을 간략하게 요약하면 다음 <표 1>과 같다.

Mirza and Osindero(2014)는 GAN의 출력 데이터 제어 불가능을 개선하기 위해 GAN의 생성기와 판별기에 추가정보 (y)를 부여하여 조건부 생성 모델을 만드는 기법인 CGAN을 제안하였다. Radford *et al.*(2016)은 생성기와 판별기의 불균등한 성능으로 인한 문제를 해결하기 위해 GAN에 합성곱 신경망(convolutional neural networks)을 결합하여 DCGAN을 제안하였다. Arjovsky *et al.*(2017)은 GAN의 불안정한 수렴으로 인한 저조한 학습 원인

2) 데이터 생성에 있어 GAN의 우수성이 확인됨에 따라 이를 기반으로 하는 100여 개에 가까운 다양한 변형 모형이 제안되었다. 각 변형 모형에 대한 비교는 Gui *et al.*(2021), Saxena and Cao(2022), Wang *et al.*(2022) 등의 연구에 요약되어 있다.

을 쿨백-라이블러 발산(Kullback-Leibler divergence)의 한계로 지적하며 손실 함수를 와서스테인 거리(Wasserstein distance)로 대체한 WGAN을 제시하였다. Zhu *et al.*(2017)은 대립하는 클래스 데이터가 필수적으로 요구되는 GAN의 단점을 극복하기 위해 대립하지 않는 데이터를 활용하여 데이터를 생성하는 기법인 CycleGAN을 제안하였다.

2.2 생산적 적대 신경망(GAN)을 활용한 정형 데이터 생성 연구

GAN이 비정형 데이터의 불균형 문제 해결에 높은 성능을 보임에 따라 정형 데이터의 불균형 문제 해결에도 이를 활용하고자 하는 시도가 이루어지고 있다. 이러한 연구들은 크게 GAN 기반 연구와 GAN 기반 기법과 다른 기법을 혼합한 하이브리드 연구로 구분할 수 있다. GAN 기반 연구들은 오버샘플링을 위해 GAN이나 이를 개선한 모형을 사용한 연구들이다. 예를 들면, Ba(2019)는 다양한 GAN 기반 오버샘플링 기법을 활용하여 신용카드 사기 거래 분류를 시도하였다. 분석결과 데이터 불균형 문제 해결에 있어 WGAN을 활용한 오버샘플링이 가장 우수한 성능을 보인 반면 CGAN을 활용한 경우 다른 모델과 유의한 차이가 없는 것으로 나타났다. Fiore *et al.*(2019)은 GAN을 활용하여 오버샘플링을 수행한 후 신용카드 사기 거래 탐지를 분석하였다. 연구 결과 GAN이 SMOTE 기법에 비해 F-1 값 측면에서 우수한 성능을 보이는 것으로 나타났다. Gangwar and Ravi(2019)는 GAN 기반 오버샘플링 기법을 활용하여 신용카드 사기 거래 분류를 시도하였다. 분석결과 GAN과 WGAN이 전통적인 오버샘플링 기법에 비해 정밀도, F-1 값, 거짓 양성(false positive) 비율 면에서 더 우수한 성능을 보이는 것으로 나타났다. Quintana and Miller(2019)는 TGAN(tabular GAN)을 기반으로 오버샘플링을 수행하고 이를 기반으로 스마트 빌딩의 주관적 온열 쾌적감(subjective thermal comport)을 예측하고자 하였다. 이 연구는 연속형 변수와 이산형 변수를 모두 고려함으로써 보다 현실적인 정형 데이터의 불균형 문제를 해결하고자 하였다. 분석결과 TGAN을 통해 생성된 데이터가 기존 데이터의 특성을 잘 반영하는 것으로 나타났다. Dlamini and Fahim(2021)은 GAN의 확장 모형인 CGAN을 활용하여 오버샘플링을 수행하고 이를 기반으로 네트워크 침입 탐지를 포착하고자 하였다. 분석결과 SMOTE나 ADASYN과 같은 전통적인 방법론에 비해 CGAN이 F-1 값, 정밀도, 재현율(recall) 등에서 보다 우수한 성능을 보이는 것으로 나타났다. Engelmann and Lessmann(2021)은 GAN의 확장 모델인 CWGAN을 기반으로 오버샘플링을 수행하고 이를 통해 대출 신청자에 대한 부도 위험 가능성을 예측하고자 하였다. 분석결과 SMOTE를 포함한 6개의 오버샘플링 모델에 비해 CWGAN이 더 우수한 성능을 보이는 것으로 나타났다. Wang and Yao(2022)는 Unrolled GAN 기반 오버샘플링 기법을 활용하여 신용카드 사기 거래 탐지를 시도하였다. 분석결과 전통적인 오버샘플링 기법인 SMOTE나 ADASYN에 비해 Unrolled GAN이 우수한 성능을 보이는 것으로 나타났다.

GAN 기반 기법과 다른 기법을 혼합한 하이브리드 연구들은 GAN 기반 기법과 다른 기법을 혼합함으로써 모형의 성능을 개선하고자 하는 연구들이다. 예를 들면, Yang *et al.*(2020)은 지도 적대 변이형 오토인코더(supervised adversarial variational autoencoder)와 WGAN을 결합하여 오버샘플링을 시행하고 이를 통해 네트워크 침입 탐지를 시도하였다. 분석결과 제안된 모형이 기존 SMOTE나 ADASYN 기반 모형에 비해 정확성, 탐지 비율, F-1 값, 거짓 양성 비율 면에서 우수한 성능이 있음을 파악하였다. Zhou *et al.*(2020)은 오토인코더와 GAN을 결합하여 오버샘플링을 시행하고 이를 통해 베어링 불량품 탐지를 시도하였다. 분석결과 오토인코더와 GAN을 결합한 오버샘플링 기법이 SMOTE나 ADASYN과 같은 전통적인 오버샘플링 기법에 비해 정확성 측면에서 우수한 성능을 보이

는 것으로 나타났다. Kate *et al.*(2022)은 GAN 기반 오버샘플링과 단일 클래스 SVM(one-class support vector machine) 기반 언더샘플링 방식을 혼합하여 은행의 고객 이탈 예측, 보험 회사의 사기 거래 예측, 대출 신청자의 부도 예측 등을 시도하였다. 분석결과 기존 언더샘플링 방식에 비해 AUC(area under curve) 값 측면에서 제안된 모형이 우수한 성능을 보이는 것을 확인하였다. Sharma *et al.*(2022)은 SMOTE와 GAN을 결합하여 SMOTified-GAN을 제안하고 이를 통해 신용카드 사기 거래 탐지를 시도하였다. 분석결과 SMOTified-GAN 모형이 F-1 값과 정밀도 측면에서 SMOTE나 GAN에 비해 우수한 성능을 보이는 것으로 나타났다. Zhu *et al.*(2022)은 GAN 기반 오버샘플링과 이웃 기반 가중 언더샘플링(neighborhood-based weighted undersampling) 방식을 혼합하여 고객 이탈 예측 연구를 수행하였다. 분석결과 제안된 오버샘플링 기법이 전통적인 오버샘플링 기법인 SMOTE나 ADASYN 뿐 아니라 WGAN 등에 비해서도 G 값, AUC 값, 정확성 측면에서 우수한 성능을 보이는 것으로 나타났다.

GAN을 활용한 오버샘플링 연구들은 SMOTE나 ADASYN과 같은 전통적인 방법의 한계점을 극복하고 다양한 기준에서 보다 우수한 성능을 보임으로써 정형 데이터의 불균형 문제 해결에도 일정 정도 기여한 것이 사실이다. 그러나 이러한 기존 연구들은 오버샘플링을 위한 모형 구조에 데이터의 형태를 끼워 맞추었기 때문에 정형 데이터의 특징을 정확하게 추출하지 못한다는 단점이 있다. 즉, 이미지 데이터와 같은 비정형 데이터의 분석을 위해 제안된 GAN의 모형에 데이터를 맞추기 위해 1차원 구조인 정형 데이터를 2차원(가로×세로)인 비정형 데이터의 구조로 변형시켜 오버샘플링을 진행하였다. 데이터의 구조를 비정형 데이터에 맞게 변형하면 GAN을 효율적으로 활용할 수 있는 장점이 있는 반면 정형 데이터의 특징이 소실되어 데이터의 특징을 정확하게 반영하지 못함으로써 생성된 데이터가 현실을 정확하게 반영하지 못한다는 단점이 있다.

2.3 순환 생산적 적대 신경망(CycleGAN)

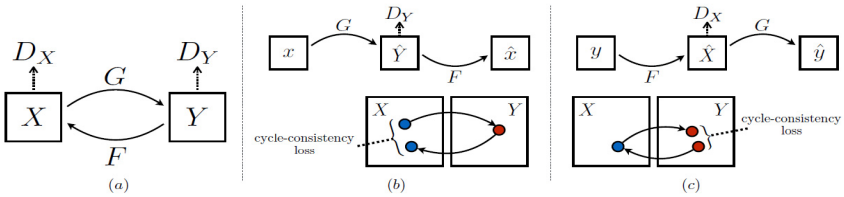
CycleGAN은 이미지 변환을 위해 GAN에서 파생된 모델이다. GAN은 비지도학습을 기반으로 실제 데이터의 확률분포를 모델링 하여 실제 데이터와 같은 확률분포를 갖는 새로운 가상의 데이터를 생성하는 생성모델의 한 유형으로 가상 데이터를 생성하는 생성기와 생성된 데이터와 실제 데이터를 구분하는 분류기로 구성되어 있다(Goofellow *et al.*, 2014). 생성기의 목적은 데이터의 분포를 학습 후 실제와 같은 가상의 데이터를 생성하는 것이 목적이며 판별기는 생성기가 생성한 데이터와 실제 데이터를 구분하는 것이 목적이다. 이를 수식으로 나타내면 다음 식 (1)과 같다.

$$\min_G \max_D V(G,D) = E_{x \sim Pdata(x)}[\log D(x)] + E_{z \sim P_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

이때, G: 생성모형, D: 판별모형,
$P_{data}(x)$: 실제 데이터 분포, $P_z(z)$: 생성된 가상 데이터

식 (1)에서 알 수 있듯이 생성기는 손실함수 V(G, D)를 최소화하기 위해 완벽에 가까운 가상 데이터를 생성하도록 학습을 진행하려 하고, 판별기는 실제 데이터와 가상의 데이터를 완벽하게 구별하기 위해 학습을 진행하려 한다. 이처럼 GAN은 생성기와 판별기가 서로 경쟁적으로 발전하는 구조라 할 수 있다.

그러나 GAN은 대립하는 클래스 데이터를 요구한다는 단점이 존재한다. Zhu *et al.*(2017)은 이러한 단점을 극복하기 위해 대립하지 않는 데이터를 활용하여 데이터를 생성하는 기법인 CycleGAN을 제안하였다. 특히 CycleGAN은 데이터의 특성을 추출하여 학습한 후 이를 나머지 데이터에 대립하는 형식으로 데이터를 생성한다는 점에서 GAN과 다른 특징을 갖고 있다. 이에 따라 모형의 구조 역시 기존 GAN과 달리 생성기와 판별기가 하나



〈그림 1〉 CycleGAN 구조

씩 더 추가되어 있다. 이를 표현하면 다음 <그림 1>과 같다.

<그림 1>의 (a)는 X, Y는 변환하고자 하는 각 도메인의 실제 데이터를 의미하고 G와 F는 생성기를, DX와 DY는 판별기를 의미한다. 이때 생성기 G(또는 F)의 목적은 X(또는 Y) 도메인의 실제 데이터의 분포를 학습한 후 Y(또는 X) 도메인의 가상 데이터를 생성하는 것이 목적이며 판별기 DY(또는 DX)는 생성기 G(또는 F)가 생성한 데이터와 실제 데이터를 구분하는 것이 목적이다. 이를 식으로 표현하면 각각 다음 식 (2), (3)과 같다.

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim P_{data}(y)}[\log D_Y(y)] + E_{x \sim P_{data}(x)}[\log(1 - D_Y(G(x)))] \quad (2)$$

$$L_{GAN}F, D_X, Y, X) = E_{x \sim P_{data}(x)}[\log D_X(x)] + E_{y \sim P_{data}(y)}[\log(1 - D_X(F(y)))] \quad (3)$$

<그림 1>의 (b)에서 X 도메인의 샘플 x는 생성기 G를 통해 $\hat{Y}$로 변환되고 DY에 의해 판별된다. $\hat{Y}$는 다시 생성기 F를 통해 $\hat{x}$로 변화되어 최초의 x와 손실함수를 계산한다. <그림 1>의 (c)는 Y 도메인의 샘플 y는 생성기 F를 통해 $\hat{X}$로 변환되고 DY에 의해 판별된다. $\hat{X}$는 다시 생성기 G를 통해 $\hat{y}$로 변화되어 최초의 y와 손실함수를 계산한다. 이러한 구조를 순환 일관성 손실(cycle-consistency loss)이라 하고 다음 식 (4)와 같이 나타낼 수 있다.

$$L_{cyc}(G,F) = E_{x \sim P_{data}(x)}[\|F(G(x)) - x\|_1] + E_{y \sim P_{data}(y)}[\|G(F(y)) - y\|_1] \quad (4)$$

따라서 CycleGAN의 전체 손실함수는 GAN 손실과 순환 일관성 손실을 합한 다음 식 (5)와 같다.

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F) \quad (5)$$

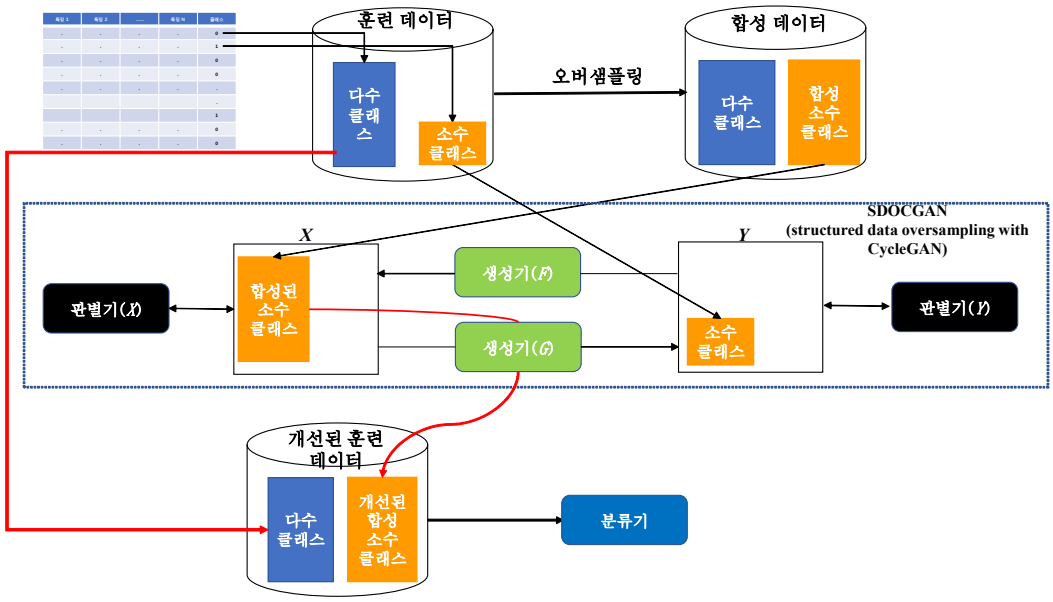
이때 λ는 순환 일관성 손실의 영향력을 결정하기 위해 도입되었다. 이를 바탕으로 CycleGAN의 학습 목표를 나타내면 다음 식 (6)과 같다.

$$G^*, F^* = arg\min_{G,F}\max_{D_X, D_Y}(G, F, D_X, D_Y) \quad (6)$$

Ⅲ. 연구 방법

3.1 연구 모형

CycleGAN의 순환 일관성 손실함수의 특성은 생성된 데이터가 원래의 데이터로 다시 돌아가도록 제약을 주는 것이다. 따라서 CycleGAN에 의해 생성된 데이터의 형태는 크게 변화되지 않는 특성이 있으며(Zhu *et al.*, 2017), 이미지 분야의 경우 이러한 특성은 단점이 된다. 그러나 기존 데이터의 특성을 크게 변형시키지 못한다는 CycleGAN의 특성은 오버샘플링의 경우 오히려 장점이 될 수 있다. 나아가 CycleGAN은 데이터의 특성을 추출하여 학습한 후 나머지 데이터에 대립하는 형식으로 데이터를 생성하기 때문에 오버샘플링 시 소수 클래스 데이터의 특성을 반영한 데이터를 생성할 수 있는 장점이 있다(최형욱, 2020). 따라서 본 연구에서는 SMOTE 기법으로 생성된 데이터를



〈그림 2〉 연구 절차

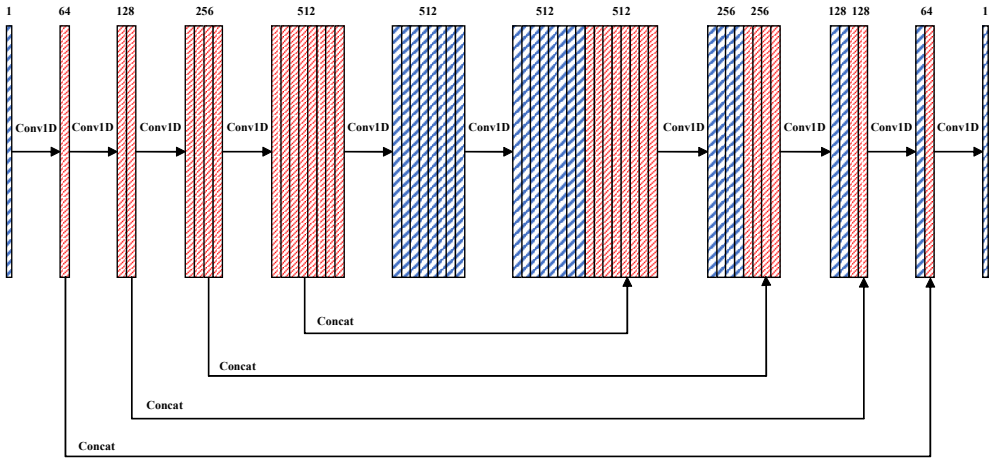
CycleGAN을 통해 학습시켜 소수 클래스 데이터의 특성을 반영한 개선된 가상의 데이터를 생성하였으며 이를 SDOCGAN(structured data oversamping with CycleGAN)이라 명명한다. 이렇게 생성된 데이터를 기반으로 분류 모형을 구축한 후 성능을 측정하였다. 본 연구의 절차를 요약하면 다음 <그림 2>와 같다.

본 연구의 절차를 자세히 설명하면 다음과 같다. 첫째, 불균형 정형 데이터를 전통적인 오버샘플링 기법인 SMOTE를 활용하여 합성 데이터를 생성한다. 둘째, 이렇게 합성된 소수 클래스 데이터와 기존 불균형 데이터의 소수 클래스 데이터를 이용하여 SDOCGAN 학습 데이터 셋을 구성하고 학습을 진행한다. 이때, 합성된 소수 클래스 데이터와 기존 불균형 데이터의 소수 클래스 데이터는 대립하는 구조가 된다. 셋째, SMOTE 기법으로 합성된 소수 클래스 데이터에 학습된 SDOCGAN의 생성기(G)를 이용하여 소수 클래스 데이터의 특성이 추가된 개선된 합성 데이터를 생성한다. 마지막으로 개선된 합성 데이터와 불균형 정형 데이터의 다수 클래스 데이터와 합하여 구축된 개선된 훈련 데이터를 이용하여 분류기를 통해 학습을 진행한 후 분류를 시행한다.

3.2 SDOCGAN의 구조

GAN을 활용하여 정형 데이터의 불균형 문제를 해결하고자 하는 기존 연구들은 정형 데이터를 2차원 이미지로 변환하여 오버샘플링을 시행함으로써 1차원 형태인 정형 데이터의 특징을 반영하는 데 한계가 존재한다. 따라서 본 연구에서는 정형 데이터의 형태를 그대로 유지하기 위하여 1차원 합성곱 신경망을 활용하여 데이터의 특징을 추출하고자 한다. 1차원 합성 곱 신경망은 정형 데이터의 특징을 더 잘 반영할 수 있을 뿐만 아니라 데이터 특징 추출에도 우수한 성능을 보이는 것으로 나타났다(Bai *et al.*, 2018). 이를 위해 기존 CycleGAN 내부의 생성기 구조를 2차원 합성곱 신경망에서 1차원 합성곱 신경망으로 변경하였다. 변경된 CycleGAN의 생성기 구조는 다음 <그림 3>과 같다.

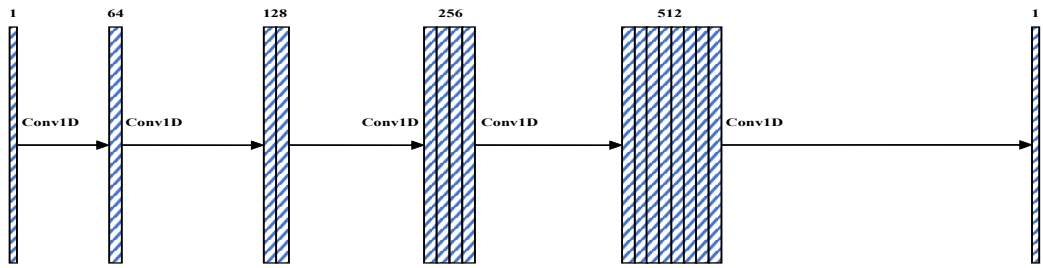
모든 1차원 합성곱 신경망의 커널 크기는 2, 스



〈그림 3〉 SDOCGAN의 생성기 (G) 구조

트라이드 크기는 1로 지정하였으며 같은 크기의 결과가 나오게 하기 위해 패딩을 사용하였다. 또한, 첫 번째와 마지막 층을 제외한 각 1차원 합성곱 신경망 층 이후에 인스턴스 정규화(instance normalization) 층을 추가하여 출력 값을 차원별로 정규화하였다. 차원을 늘릴 때는 Leaky-ReLu 활성함수를, 차원을 줄일 때는 ReLu 활성함수를 사용하였다. 모든 층에서 원활한 학습 시작을 위해 랜덤 노멀 이니셜라이저(random normal initializer)를 사용하여 가중치를 초기화하였다. 초기 가중치 값은 평균과 표준편차가 각각 0과 0.02인 정규분포에서 무작위로 추출하였다. CycleGAN의 Unet 구조와 같이 추출된 특징을 더 잘 전달하기 위해 앞부분과 뒷부분을 연결해주는 연결 층을 추가하였다. 마지막으로 모든 1차원 합성곱 신경망에 대한 아웃풋 차원의 수는 각각 64, 128, 256, 512, 512, 512, 256, 128, 64, 1로 지정하였다.

판별기 또한 2차원 합성곱 신경망을 1차원 합성곱 신경망으로 변경하였으며 이의 구조는 다음 <그림 4>와 같다. 생성기와 동일하게 모든 1차원 합성곱 신경망에서 커널의 크기는 2, 스트라이드는 1로 지정하였으며 패딩을 사용하였다. 초기화를 위해 평균 0, 표준편차 0.02인 랜덤 노멀 이니셜라이저를 사용하였다. 첫 번째 층을 제외한 모든 1차원 합성곱 신경망 뒤에는 인스턴스 정규화 층을 추가하였다. 마지막 층을 제외한 모든 부분에서 Leaky-ReLu 활성함수를 사용하였다. 마지막 바로 전 층에서는 최대한 많은 특징을 추출하기 위해 노이즈를 제거하였다. 1차원 합성곱 신경망의 아웃풋 차원 수는 64, 128, 256, 512, 1로 지정하였다.



〈그림 4〉 SDOCGAN의 판별기 (D) 구조

Ⅳ. 실험 및 연구 결과

4.1 데이터 셋

본 연구에서 제안된 SDOCGAN을 이용한 하이브리드 오버샘플링 기법의 성능을 파악하기 위하여 두 가지 실험을 진행하였다. 첫째, 파이썬의 사이킷런 패키지를 이용하여 인공적으로 불균형 데이터를 생성하여 이를 대상으로 제안된 기법의 성능을 평가하였다. 정형 데이터의 오버샘플링은 다양한 요인에 의해 영향을 받는다(Ba, 2019). 그 가운데 데이터의 차원과 불균형 정도는 오버샘플링의 성능에 지대한 영향을 미치는 것으로 나타났다. 차원의 숫자가 많을수록 오분류(misclassification)의 가능성이 증가하고, 불균형 정도가 심할수록 과적합의 위험이 증가하여 오버샘플링의 성능이 하락하게 된다(Deepa and Punithavalli, 2011). SDOCGAN의 적용에 있어 이들의 영향을 정확하게 파악하기 위하여 다음과 같은 두 가지 데이터 셋을 준비하였다. 먼저 차원의 영향을 파악하기 위해 다수 클래스 데이터와 소수 클래스 데이터의 비율을 99:1로 고정하고 차원의 수를 20, 25, 30, 35, 40으로 구분한 5종류의 데이터 셋을 생성하였다. 다음으로 불균형 정도의 영향을 파악하기 위해 차원의 수를 30으로 고정하고 불균형 정도를 99:1, 99.5:0.5, 99.7:0.3으로 구분한 3종류의 데이터 셋을 생성하였다. 학습을 위한 8개 데이터 셋의 데이터 양은 각각 25만개로 고정하였으며 모든 데이터 셋에서 70%는 학습용으로, 30%는 평가용으로 분리하여 사용하였다.

둘째, SDOCGAN이 실제 사례에 얼마나 잘 적용되는지를 평가하기 위하여 불균형 데이터 문제 해결을 위해 캐글(Kaggle)에서 사용되었던 신용카드 거래 데이터 셋을 사용하였다. 해당 데이터는 총 284,807건의 신용거래 데이터로 구성되어 있으며 이 가운데 0.172%인 492건만 사기 거래로 판별된 불균형 데이터이다. 전체 데이터 가운데 70%인 199,365건을 학습용으로 사용하였으며 30%인 85,442건을 평가용 데이터로 사용하여 SDOCGAN의 성능을 평가하였다.

4.2 실험 설정

SDOCGAN을 활용한 하이브리드 오버샘플링 기법의 성능 비교를 위해 SMOTE, ADASYN, Polynom_fit_SMOTE(Gazzah and Amara, 2008), SMOTE_IPF(SMOTE_iterative partitioning filter)(Sáez *et al.*, 2015)의 성능을 함께 확인하였다. SMOTE와 ADASYN은 오버샘플링을 위해 전통적으로 가장 많이 활용되었기 때문에 선택하였으며 Polynom_fit_SMOTE와 SMOTE_IPF는 오버샘플링 기법 가운데 가장 좋은 성능을 보이는 것으로 나타났기 때문이다(Kovács, 2019).

성능 비교를 위한 SDOCGAN의 학습 세부사항은 다음과 같다. 먼저 적대적 학습을 위해 순환 일관성 손실의 영향력인 λ(식 (5) 참조)는 10으로 설정하였다(Zhu *et al.*, 2017). 모든 옵티마이저는 학습률 0.0002, 베타1 0.5를 가진 Adam을 사용하였다. 분류기는 가장 대중적으로 사용되며 우수한 성능을 보이는 랜덤 포레스트를 사용하였으며(Fernández-Delgado *et al.*, 2014), 분류의 성능은 F-1 값으로 비교하였다. CycleGAN의 학습은 생성된 이미지를 보고 중단을 판단하기 때문에 정확한 중단 기준이 없다. 따라서 기존 연구에서 주로 활용되는 100 epoch로 학습을 진행하였다(Zhu *et al.*, 2017). 학습과 검증은 모두 파이썬 환경에 설치된 Tensorflow 2.0을 이용하여 진행하였다.

4.3 실험 결과

먼저 차원의 영향을 파악하기 위하여 다수 클래스 데이터와 소수 클래스 데이터의 비율을 99:1로 고정하고 차원의 수를 다양하게 변경하여 인공적으로 생성된 데이터를 기반으로 개별 오버샘플링 기법 간의 성능 비교를 실시하였다. 이를 요약하면 다음 <표 2>와 같다.

〈표 2〉 차원에 따른 오버샘플링 기법 간 성능 비교

차원	원본데이터	SMOTE	ADASYN	Polynom_fit_ SMOTE	SMOTE_IPF	본 연구 (SMOTE + SDOCGAN)
20	0.4361	0.4564	0.4593	0.4392	0.4575	**0.4719**
25	0.4460	0.4478	0.4590	**0.4747**	0.4421	0.4604
30	0.4306	0.4404	0.4274	**0.4591**	0.4417	0.4487
35	0.3800	0.4221	0.4152	0.3852	0.4222	**0.4460**
40	0.3711	0.4213	0.4139	0.3700	0.4099	**0.4393**

<표 2>에서 알 수 있듯이 25와 30차원의 경우 Polynom_fit_SMOTE가 본 연구에서 제안한 하이브리드 오버샘플링에 비해 더 높은 성능을 보이는 것으로 나타났으나 그 차이는 각각 0.0143과 0.0104로 크지 않았다. 특히 Polynom_fit_SMOTE의 경우 데이터의 차원이 높아짐에 따라 F-1 값의 점수가 큰 폭으로 하락하였을 뿐만 아니라 마지막 40차원의 경우 오히려 원본데이터의 F-1 값보다 낮은 값을 보이는 것으로 나타났다. 반면 본 연구에서 제안한 SMOTE와 SDOCGAN을 혼합한 기법의 경우 20, 35, 40차원에서 다른 기법들에 비해 높은 성능을 보였다. 특히 35와 40차원의 경우 다른 기법들과는 매우 큰 차이 있는 것으로 나타났을 뿐 아니라 차원의 변화에도 상대적으로 안정적인 변화를 보이는 것으로 나타났다. 이를 통해 전반적으로 본 연구에서 제안한 하이브리드 오버샘플링 기법이 가장 안정적이며 우수한 성능을 보인다는 것을 알 수 있다.

다음으로 데이터 간 불균형 정도의 영향을 파악하기 위하여 차원의 수를 30으로 고정하고 불균형 정도를 다양하게 변경하여 인공적으로 생성된 데이터를 기반으로 본 연구에서 제안한 기법과 기존 오버샘플링 기법 간 성능 비교를 실시하였다. 이를 요약하면 다음 <표 3>과 같다.

<표 3>에서 나타난 것과 같이 불균형 정도가 99:1일 경우 Polynom_fit_SMOTE 기법이 본 연구에서 제안한 하이브리드 오버샘플링 기법보다 약간 더 좋은 성능을 나타냈다. 그러나 불균형 정도가 99.5:0.5와 99.7:0.3으로 점점 심해짐에 따라 Polynom_fit_SMOTE는 원본데이터의 F-1 값보다 낮은 값을 보이는 것으로 나타났다. 반면 본 연구에서 제안한 하이브리드 오버샘플링 기법의 경우 가장 높은 F-1 값을 보였다. 즉, 본 연구에서 제안한 기법은 99:1일 경우 Polynom_fit_SMOTE 기법에 비해 약간 낮은 성능을 보인 반면 불균형 정도가 심해질수록 다른 모든 기법에 비해 우수한 성능을 보이는 것으로 나타났다.

다음으로 실제 신용카드 사기 거래 데이터를 기반으로 본 연구에서 제안한 기법과 기존 기법 간의 성능을 비교하였다. <표 4>에서 알 수 있듯

〈표 3〉 불균형 정도 따른 오버샘플링 기법 간 성능 비교

불균형 정도 (다수:소수)	원본데이터	SMOTE	ADASYN	Polynom_fit_ SMOTE	SMOTE_IPF	본 연구 (SMOTE + SDOCGAN)
99:1	0.4306	0.4404	0.4274	**0.4591**	0.4417	0.4480
99.5:0.5	0.2166	0.2825	0.2712	0.1638	0.2712	**0.2949**
99.7:0.3	0.0978	0.1576	0.1530	0.0402	0.1506	**0.2046**

〈표 4〉 신용카드 사기 거래 데이터에 대한 오버샘플링 기법 간 성능 비교

성능 지표	원본데이터	SMOTE	ADASYN	Polynom_fit_ SMOTE	SMOTE_IPF	본 연구 (SMOTE + SDOCGAN)
재현률(recall)	0.7536	0.8261	0.8261	0.7971	0.8188	**0.8478**
정밀도(precision)	**0.9204**	0.8028	0.8201	0.8271	0.8071	0.8540
F-1 값	0.8287	0.8143	0.8231	0.8118	0.8129	**0.8509**

이 다른 기법들에 비해 본 연구에서 제안한 기법이 우수한 성능을 보이는 것으로 나타났다. 특히 기존 기법들의 성능이 매우 낮은 것으로 나타났는데 이는 데이터의 불균형 정도가 99.828:0.172로 심했기 때문인 것으로 판단된다.

<표 4>에서 알 수 있듯이 원본데이터에 비해 모든 기법의 재현률 값이 향상되었음을 알 수 있다. 이는 소수 클래스 데이터의 적발이 중요한 오버샘플링의 경우 재현율이 다른 성능 지표에 비해 상대적으로 중요하기 때문에 오버샘플링 기법들이 이를 향상시키는 것에 초점을 두었기 때문으로 이해할 수 있다. 그러나 재현율 값을 향상시키기 위해서는 정밀도 값을 낮추어야 한다. 이로 인해 기존 오버샘플링 기법의 경우 두 지표의 조화 평균인 F-1 값이 원본데이터에 비해 낮아진 것으로 판단된다. 본 연구에서 제안한 하이브리드 오버샘플링 기법의 경우 다른 기법들에 비해 정밀도 점수의 하락 폭이 작고 재현율 점수의 상승이 높았기 때문에 F-1 값 역시 우수한 것으로 나타났다. 본 연구에서 제안한 기법의 경우 원본데이터와 비교하여 재현율 값은 0.0942, F-1 값은 0.0222 상승하였다.

본 연구에서 제안한 SDOCGAN이 SMOTE가 아닌 다른 오버샘플링 기법을 혼합하여 사용할 때도 성능을 향상시킬 수 있는지 파악하기 위하여 신용카드 사기 거래 데이터를 활용하여 추가적인 실험을 진행하였다. <표 5>에 나타나 있듯이 ADASYN과 Polynom_fit_SMOTE를 SDOCGAN과 혼합한 경우에도 단일 기법에 비해 F-1 값이 향상되었음을 알 수 있다. 이를 통해 본 연구에서 제안한 기법의 확장 가능성을 확인할 수 있다.

4.4 추가 실험

본 연구에서 제안한 기법의 성능을 보다 정확하게 확인하기 위해서는 GAN 기반의 다른 오버샘플링 기법과 비교할 필요가 있다. 그러나 동일한 실험 환경의 구축이 불가능하기 때문에 이를 직접적으로 확인할 수 없다. 따라서 동일한 신용카드 사기 거래 데이터를 기반으로 GAN을 적용하여 오버샘플링을 진행한 가장 최신 연구인 김예원 등(2020)의 결과와 비교함으로써 간접적으로 본 연구에서 제안한 기법의 성능을 확인하고자 하였다.

〈표 5〉 SDOCGAN과 다른 기법을 혼합한 오버샘플링 기법 간 성능 비교

성능 지표	ADASYN	ADASYN + SDOGAN	Polynom_fit_ SMOTE	Polynom_fit_ SMOTE + SDOGAN	SMOTE	본 연구 (SMOTE + SDOCGAN
재현률(recall)	0.8261	0.8333	0.7971	0.8261	0.8261	0.8478
정밀도(precision)	0.8201	0.8156	0.8271	0.8507	0.8028	0.8540
F-1 값	0.8231	**0.8244**	0.8118	**0.8382**	0.8143	**0.8509**

〈표 6〉 GAN기반 오버샘플링 기법을 활용한 연구 결과와 비교

		적용 전 AUC 값	적용 후 AUC 값	AUC 값 증가량
본 연구	SMOTE +SDOGAN	0.876759	0.923796	**0.047037**
김예원 등 (2020)	GAN	0.910690	0.913723	0.003033
	CGAN	0.910690	0.916118	0.005428
	WGAN	0.910690	0.926209	0.015519
	CWGAN	0.910690	0.916128	0.005438

김예원 등(2020)의 연구에서 AUC 값을 활용하여 성능을 평가하였기 때문에 본 연구에서 제안한 기법에 대한 AUC 값을 구하고 이를 통해 성능 비교를 수행하였다. 다만 같은 데이터를 이용했음에도 불구하고 학습용과 실험용 데이터의 구분 과정, 랜덤 포레스트 분류기를 사용하는 과정에서 동일한 환경을 구축하는 것이 불가능하였기 때문에 AUC 값 자체가 아니라 두 기법 간의 AUC 값의 증가량을 통해 성능을 비교하였으며 그 결과는 <표 6>에 나타나 있다.

<표 6>에서 알 수 있듯이 본 연구에서 제안한 기법을 활용한 오버샘플링의 경우 AUC 값이 0.047037(0.923796-0.876759)만큼 증가한 반면 김예원 등(2020)의 연구에서 가장 좋은 성능을 나타낸 WGAN의 경우 0.015519(0.926209-0.910690)만큼 증가하였다. 이를 통해 본 연구에서 제안한 기법이 GAN을 이용한 다른 기법에 비해 우수한 성능을 보인다고 추론할 수 있다. 물론 AUC 값의 증가량으로 성능을 비교하는 것이 좋은 방법은 아니다. 그러나 상이한 실험 환경으로 인해 절대적인 비교 기준이 없다는 점과 AUC 값이 각각 0.923796과 0.926209로 크게 차이가 나지 않는 점을 고려해 볼 때 증가량을 통한 간접적인 성능 비교가 일정 정도 의미가 있다고 할 수 있다.

Ⅴ. 결 론

비정형 데이터의 불균형 문제 해결에 있어 GAN 기반 오버샘플링 기법의 우수성이 알려짐에 따라 다양한 관련 연구가 진행되어 왔다(Douzas and Bacao, 2018). 특히 최근에는 이를 정형 데이터의 불균형 문제 해결에도 적용하기 시작하였다(Quintana and Miller, 2019). 그러나 이러한 연구들은 데이터의 형태를 비정형 데이터 구조로 변경함으로써 정형 데이터의 특징을 정확하게 반영할 수 없다는 한계점이 존재한다. 본 연구에서는 이를 해결하기 위해 CycleGAN을 정형 데이터의 구조에 맞게 재구성하고 이를 기존의 SMOTE 기법과 결합한 하이브리드 오버샘플링 기법을 제안하였다. 특히 기존 연구와 달리 1차원 합성곱 신경망을 사용함으로써 기존 연구의 한계를 극복하고자 하였다. 해당 기법의 성능을 확인하기 위해 인공적으로 생성한 8종류의 데이터 셋과 신용카드 사기 거래 데이터 셋의 분석에 적용하여 구한 F-1 값을 기존 오버샘플링 기법들과 비교하였다. 분석결과 차원이 많을수록, 불균형 정도가 심할수록 제안된 모형이 우수한 성능을 보이는 것으로 나타났다. 또한, 동일한 데이터를 활용한 김예원 등(2020)의 연구와 비교를 통해 본 연구에서 제안한 기법의 우수성을 확인하였다.

5.1 연구의 시사점

본 연구는 다음과 같은 이론 및 실무적 의의가 있다. 첫째, 정형 데이터의 특성을 반영함으로써 GAN 기반 오버샘플링 기법의 성능을 개선하였다. GAN은 이미지 데이터의 생성을 위해 고안되어(Goodfellow *et al.*, 2014), 이미지 생성뿐만 아니라 이미지 변형(Zhu *et al.*, 2016), 이미지 복원(Pathak *et al.*, 2016) 등에서 우수한 성능을 보이는

것으로 나타났다. 이미지 데이터 분석을 위해 제안된 특성으로 인해 GAN을 정형 데이터의 오버샘플링에 활용한 연구들은 1차원 구조인 정형 데이터를 이미지와 같은 2차원 구조로 변경하여 오버샘플링을 진행하였다(Dlamini and Fahim, 2021). 이러한 데이터의 구조 변형은 GAN을 효율적으로 활용한다는 점에서는 장점이 있으나 정형 데이터의 특징을 정확하게 반영하지 못한다는 단점이 존재한다. 본 연구에서는 1차원 합성곱 신경망을 활용하여 정형 데이터의 형태를 그대로 유지함으로써 정형 데이터의 특징을 정확하게 반영하고자 하였다. 나아가 데이터 특징 추출에도 우수한 성능을 보이는 것으로 알려진 1차원 합성곱 신경망을 활용함으로써(Bai *et al.*, 2018), GAN 기반 오버샘플링 기법의 성능 개선에 기여하였다.

둘째, 소수 클래스 데이터의 특성을 반영한 데이터를 생성함으로써 오버샘플링 성능을 개선하였다. 본 연구에서는 이미지 데이터 생성에 활용되어 온 CycleGAN을 활용하여 정형 데이터에 대한 오버샘플링을 진행하였다. CycleGAN은 생성된 데이터가 원래의 데이터로 다시 돌아가도록 제약을 하기 때문에 데이터 형태를 크게 변화시키지 못한다는 단점이 존재한다(Zhu *et al.*, 2017). 그러나 이러한 단점은 기존 데이터의 특성을 변형시키지 못하기 때문에 확률분포에 따라 생성된 데이터가 기존 데이터의 특성을 잘 보전한다는 점에서 오버샘플링의 경우에는 오히려 장점이 될 수 있다. 또한, CycleGAN은 데이터의 특성을 추출하여 학습한 후 나머지 데이터에 대립하는 형식으로 데이터를 생성하기 때문에 오버샘플링 시 소수 클래스 데이터의 특성을 반영한 데이터를 생성할 수 있다(최형욱, 2020). 이처럼 본 연구는 CycleGAN의 활용을 통해 오버샘플링 성능의 개선을 위한 새로운 방법을 제안하였다는 점에서 의의가 있다.

셋째, SMOTE와 SDOCGAN을 혼합한 하이브리드 기법을 제안함으로써 오버샘플링 성능을 개선하였다. SMOTE는 생성된 데이터의 다양성이 부족하고 과적합의 문제를 야기하는 단점이 있는 반면 GAN은 생성기와 판별기의 불균등한 성능 및 불안정한 수렴 등의 단점이 있다. 이러한 문제를 해결하기 위해 본 연구에서는 SMOTE와 SDOCGAN을 혼합함으로써 두 방법의 장점은 유지하면서 단점은 극복하고자 하였다. 특히 일반적으로 딥러닝과 혼합하여 사용하는 경우가 드물었던 SMOTE를 딥러닝 기법과 혼합하였을 뿐만 아니라 소수 클래스 데이터에는 잘 적용하지 않았던 CycleGAN을 변형하여 소수 클래스 데이터에 적용하였다(Mullick *et al.*, 2019). 노이즈 값을 입력하는 기존 연구와 달리 SMOTE를 사용하여 생성된 데이터를 변형된 CycleGAN의 입력 값으로 활용함으로써 보다 우수한 성능을 확인할 수 있었다. 특히 차원이 커질수록 성능이 저하되는 전통적인 오버샘플링 기법과 달리 본 연구에서 제안하는 기법은 차원 증가에 따른 성능의 하락폭이 상대적으로 매우 적어 안정적인 성능을 나타냈다(<표 2> 참조).

넷째, 다양한 기존 오버샘플링 기법과 혼합하여 사용할 수 있는 확장성 높은 기법을 제시하였다. 본 연구에서는 SMOTE 기법뿐만 아니라 ADASYN, Polynom_fit_SMOTE 등의 다른 기법들과 SDOCGAN의 혼합을 통한 성능 개선의 가능성도 함께 제시하였다(<표 5> 참조). 불균형 데이터의 중요성이 증대함에 따라 이를 해결하기 위한 다양한 기법들이 진행되고 있다(Sharma *et al.*, 2022). SDOCGAN의 경우 오버샘플링을 통해 생성된 데이터를 입력 값으로 받아 CycleGAN을 통해 다시 오버샘플링을 진행하는 방식이기 때문에 다른 기법과의 혼합이 용이하다. 특히 실무에서는 이미 구축된 데이터 분석 파이프라인으로 인해 성능이 우수한 오버샘플링 기법이 새롭게 제안된다 하더라도 이를 즉시 도입하여 사용하기가 쉽지 않다. 그러나 본 연구에서 제안하는 하이브리드 오버샘플링의 경우 기존 분석 파이프라인을 변경하지 않고 SDOCGAN을 추가하는 방식이기 때문에 실무에 쉽게 적용할 수 있을 것이다. 본 연구는 새롭게 개발되는 다양한 기법을 SDOCGAN과 혼합하는 방법을 제시함으로써 오버샘플링 성능 개선을 위해 쉽게 활용

가능한 대안을 제시하였다는 점에서 의의가 있다.

5.2 한계점 및 향후 연구 방향

본 연구는 다음과 같은 한계점이 있으며 향후 연구를 통해 이를 해결해야 할 것이다. 첫째, 본 연구에서는 오버샘플링 이후 성능 비교를 위한 분류기를 활용함에 있어 가장 대중적으로 사용되며 우수한 성능을 보이는 것으로 알려진 랜덤 포레스트만을 활용하였다(Fernández-Delgado *et al.*, 2014). 그러나 랜덤 포레스트 이외에도 다양한 분류기가 존재하고 있으며 각각의 장단점이 상이한 만큼 다양한 분류기를 활용하여 성능을 비교함으로써 본 연구에서 제안한 기법의 강건성(robustness)을 검증할 필요가 있다. 둘째, 실제 비즈니스에서 활용된 데이터를 분석하지 못했다. 프라이버시와 기업 정책과 같은 다양한 원인으로 인해 실제 기업 현장에서 발생한 데이터를 활용하지 못하고 인공적으로 생성한 데이터와 캐글에서 활용된 데이터를 기반으로 본 연구에서 제안한 기법의 성능을 비교하였다. 이로 인해 본 연구 결과를 실제 기업 현장에 적용하는 데 주의가 필요하다. 향후 연구에서는 기업으로부터 관련 데이터를 직접 수집함으로써 실제 상황을 반영한 정밀한 분석을 수행할 필요가 있다. 셋째, 딥러닝을 이용한 오버샘플링 기법과의 정량적 비교를 하지 못했다. 실험 환경의 구성 문제로 인해 본 연구에서는 직접적인 비교 대신 AUC 값의 증가량을 통해 딥러닝 기반 오버샘플링 기법들과의 성능을 간접적으로 비교하였다(<표 6> 참조). 따라서 향후 연구에서는 동일한 실험 환경하에서 딥러닝을 활용한 오버샘플링 기법과 본 연구에서 제안한 기법을 직접적으로 비교할 필요가 있다. 마지막으로, 다양한 유형의 정형 데이터를 고려하지 못했다. 본 연구에서는 정형 데이터 가운데 신용카드 사기 거래 데이터를 기반으로 오버샘플링의 성능을 비교하였다. 그러나 동일한 정형 데이터라 할지라도 데이터의 특징 (예: 연속형 또는 이산형)에 따라 오버샘플링 기법의 성능이 달라질 수 있기 때문에 이를 정확하게 고려할 필요가 있다. 따라서 향후 연구에서는 보다 다양한 특징을 갖는 정형 데이터를 분석함으로써 연구의 일반화 가능성을 향상시킬 필요가 있다.

참 고 문 헌

[1] 김예원, 유예림, 최홍용, “생성적 적대 신경망과 딥러닝을 활용한 이상거래 탐지 시스템 모형”, *Information Systems Review*, 제22권, 제1호, 2020, pp. 59-72.

[2] 최형욱, 이승현, 김형훈, 서용철, “CycleGAN을 활용한 항공영상 학습 데이터 셋 보완 기법에 관한 연구”, *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, 제38권, 제6호, 2020, pp. 499-509.

[3] Chen, H., J. Chen, and J. Ding, “Data evaluation and enhancement for quality improvement of machine learning”, *IEEE Transactions on Reliability*, Vol.70, No.2, 2021, pp. 831-847.

[4] Zhou, F., S. Yang, H. Fujita, D. Chen, C. Wene, “Deep learning fault diagnosis method based on global optimization GAN for unbalanced data”, *Knowledge-Based Systems*, Vol.187, 2020, 104837.

[5] Arjovsky, M., S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks”, *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 214-223.

[6] Aydilek, I. B. and A. Arslan, “A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm”, *Information Science*, Vol.233, 2013, pp. 25-35.

[7] Ba, H., “Improving detection of credit card fraudulent transactions using generative adversarial networks”, arXiv, 2019, Available at https://doi.org/10.48550/arXiv.1907.03355.

[8] Bai, S., J. Z. Kolter, and V. Koltun, “An empirical

evaluation of generic convolutional and recurrent networks for sequence modeling", arXiv, 2018, Available at https://doi.org/10.48550/arXiv.1803.01271.

[9] Bosu, M. F. and S. G. MacDonell, "A taxonomy of data quality challenges in empirical software engineering", *Proceedings of the 22nd Australian Software Engineering Conference*, 2013, pp. 97-106.

[10] Cao, Q. and S. Wang, "Applying over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning", *Proceedings of the 2011 International Conference on Information Management, Innovation Management and Industrial Engineering*, 2011, pp. 543-548.

[11] Chandola, V., A. Banerjee, and V. Kumar, "Anomaly detection: A survey", *ACM Computing Surveys*, Vol.41, No.3, 2009, pp. 1-58.

[12] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, Vol.16, 2002, pp. 321-357.

[13] Chawla, N. V., N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets", *ACM SIGKDD Explorations Newsletter*, Vol.6, No.1, 2004, pp. 1-6.

[14] Deepa, T., and M. Punithavalli, "An E-SMOTE technique for feature selection in high-dimensional imbalanced dataset", *Proceedings of the 3rd International Conference on Electronics Computer Technology*, Vol.2, 2011, pp. 322-324.

[15] Dlamini, G., and M. Fahim, "DGM: A data generative model to improve minority class presence in anomaly detection domain", *Neural Computing and Applications*, Vol.33, No.20, 2021, pp. 13635-13646.

[16] Douzas, G. and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks", *Expert Systems with Applications*, Vol.91, 2018, pp. 464-471.

[17] Engelmann, J., and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning", *Expert Systems with Applications*, Vol.174, 2021, 114582.

[18] Fangyu, W., Z. Jianhui, B. Youjun, and C. Bo, "Research on imbalanced data set preprocessing based on deep learning", *Proceedings of the 2021 Asia-Pacific Conference on Communications Technology and Computer Science*, 2021, pp. 75-79.

[19] Fernández, A., S., del Río, N. V. Chawla, F. Herrera1, "An insight into imbalanced big data classification: Outcomes and challenges", *Complex & Intelligence Systems*, Vol.3, 2017, pp. 105-120.

[20] Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?", *Journal of Machine Learning Research*, Vol.15, No.1, 2014, pp. 3133-3181.

[21] Fiore, U., A. De Santis, F. Perla, P. Zanetti, F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection", *Information Science*, Vol.479, 2019, pp. 448-455.

[22] Gangwar, A. K., and V. Ravi, "WiP: Generative adversarial network for oversampling data in credit card fraud detection", *Proceedings of the 15th International Conference on Information Systems Security*, 2019, pp. 123-134.

[23] Gazzah, S., and N. E. B. Amara, "New oversampling approaches based on polynomial fitting for imbalanced data sets", *Proceedings of the Eighth IAPR International Workshop on Document Analysis Systems*, 2008, pp. 677-684.

[24] Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courvile, and Y. Bengio, "Generative adversarial nets",

Advances in Neural Information Processing Systems, Vol.27, 2014, pp. 2672-2680.

[25] Gui, J., Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications", *IEEE Transactions on Knowledge and Data Engineering*, in press, 2021.

[26] Han, H., W. Y. Wang, B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning", *Lecture Notes in Computer Science*, Vol.3644, No.5, 2005, pp. 878-887.

[27] He, H., and E.A. Garcia, "Learning from imbalanced data", *IEEE Transactions on Knowledge and Data Engineering*, Vol.21, No.9, 2009, pp. 1263-1284.

[28] He, H., Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks*, 2008, pp. 1322-1328.

[29] IBM, "Inforgraphic-Extracting business value form the 4Vs of big data", 2020, Available at https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data.

[30] Islam, A., S. B. Belhaouari, A. U. Rehman, and H. Bensmail, "KNNOR: An oversampling technique for imbalanced datasets", *Applied Soft Computing*, Vol.115, 2022, 108288.

[31] Johnson, J. M., and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance", *Journal of Big Data*, Vol.6, 2019, p. 27.

[32] Kate, P., V. Ravi, and A. Gangwar, "FinGAN: Generative adversarial network for analytical customer relationship management in banking and insurance", arXiv, 2022, Available at https://doi.org/10.48550/arXiv.2201.11486.

[33] Khoshgoftaar, T. M., A. Fazelpour, D. J. Dittman, and A. Napolitano, "Ensemble vs. data sampling: Which option is best suited to improve classification performance of imbalanced bioinformatics data?", *Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence*, 2015, pp. 705-712.

[34] Kingma, D. P., and P. Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions", *Proceedings of the Advances in Neural Information Processing Systems 31*, 2018, Available at https://doi.org/10.48550/arXiv.1807.03039.

[35] Kovács, G., "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets", *Applied Soft Computing*, Vol.83, 2019, 105662.

[36] Krawczyk, B. "Learning from imbalanced data: Open challenges and future directions", *Progress in Artificial Intelligence*, Vol.5, No.4, 2016, pp. 221-232.

[37] Krizhevsky, A., I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", *Communications of the ACM*, Vol.60, No.6, 2017, pp. 84-90.

[38] Leevy, J. L., T. M. Khoshgoftaar, R. A., Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data", *Journal of Big Data*, Vol.5, 2018, 42.

[39] Ling, C. X. and C. Li, "Data mining for direct marketing: Problems and solutions", *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998, pp. 73-79.

[40] Liu, Y., H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach", *Expert Systems with Applications*, Vol.36, 2009, pp. 690-701.

[41] Mirza, M., and S. Osindero, "Conditional generative adversarial nets", arXiv, 2014, Available at https://doi.org/10.48550/arXiv.1411.1784.

[42] Mohammed, R., J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results", *Proceedings of the 11th International Conference on Information and Communication Systems*, 2020, pp. 243-248.

[43] Mullick, S. S., S. Datta, and S. Das, "Generative adversarial minority oversampling", *Proceedings of IEEE/CVF International Conference on Computing Vision*, 2019, pp. 1695-1704.

[44] Nazari, E., P. Branco, "On oversampling via generative adversarial networks under different data difficulty factors", *Proceedings of Machine Learning Research*, Vol.154, 2021, pp. 76-89.

[45] Pathak, D., P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting", *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536-2544.

[46] Quintana, M., and C. Miller, "Towards class-balancing human comfort datasets with GANs", *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2019, pp. 391-392.

[47] Radford, A., L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks", *Proceedings of the International Conference on Learning Representations*, 2016, Available at https://doi.org/10.48550/arXiv.1511.06434.

[48] Refinitive, "Smarter humans. Smarter machines", 2019, Available at https://www.refinitiv.com/content/dam/marketing/en_us/documents/gated/reports/refinitiv-ai-ml-survey-report.pdf#form?utm_source=Press_release&utm_medium=web&utm campaign=107263 AISurveyReport&utm_term=&utm_content=Reglp&elqCampaignId=6848.

[49] Sáez, J. A., J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering", *Information Sciences*, Vol.291, 2015, pp. 184-203.

[50] Sambasivan, N., S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. ""Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI", *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1-15.

[51] Saxena, D., and J. Cao, "Generative adversarial networks (GANs) challenges, solutions, and future directions", *ACM Computing Surveys*, Vol.54, No.3, 2022, pp.1-42.

[52] Sharma, A., P. K. Singh and R. Chandra, "SMOTified-GAN for class imbalanced pattern classification problems", *IEEE Access*, Vol.10, 2022, pp. 30655-30665.

[53] Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play", *Science*, Vol.362, No.6419, 2018, pp. 1140-1144.

[54] Soltanzadeh, P., and M. Hashemzadeh, "RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem", *Information Science*, Vol.542, 2020, pp. 92-111.

[55] Statista, "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025", 2021, Available at https://www.statista.com/statistics/871513/worldwide-data-created/.

[56] Tek, F. B., A. G. Dempster, and I. Kale, "Parasite detection and identification for automated thin

blood film malaria diagnosis", *Computer Vision and Image Understanding*, Vol.114, 2010, pp. 21-32.

[57] Thejas G. S., Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, and S. Chennupati, "An extension of synthetic minority oversampling technique based on Kalman filter for imbalanced datasets", *Machine Learning with Applications*, Vol.8, 2022, 100267.

[58] Tomek, I., "Two modifications of CNN", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.6, No.11, 1976, pp. 769-772.

[59] Wang, J., and L. Yao, "Unrolled GAN-based oversampling of credit card dataset for fraud detection", *Proceedings of the 2022 IEEE International Conference on Artificial Intelligence and Computer Applications*, 2022, pp. 858-861.

[60] Wang, Z., Q. She, and T. E. Ward, "Generative adversarial networks: A survey and taxonomy", *ACM Computing Surveys*, Vol.54, No.2, 2022, pp.1-38.

[61] Wilson, D. L., "Asymptotic properties of nearest neighbor rules using edited data", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.2, No.3, 1972, pp. 408-421.

[62] Wise, J., "How much data is created every day in 2022?", 2022, Available at https://earthweb.com/how-much-data-is-created-every-day/#Key_Data_Creation_Statistics_2022.

[63] Xu, L., *Synthesizing Tabular Data using Conditional GAN* (Master's thesis), Massachusetts Institute of Technology, 2020.

[64] Yang, Y., K. Zheng, B. Wu, Y. Yang, X. Wang, "Network intrusion detection based on supervised adversarial variational auto-encoder with regularization", *IEEE Access*, Vol.8, 2020, pp. 42169-42184.

[65] Yap, B. W., K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets", *Proceedings of the First International Conference on Advanced Data and Information Engineering*, 2013, pp. 13-22.

[66] Zhou, B., C. Yang, H. Guo, and J. Hu, "A quasi-linear SVM combined with assembled SMOTE for imbalanced data classification", *Proceedings of the 2013 International Joint Conference on Neural Networks*, 2013, pp. 1-7.

[67] Zhu, B., X. Pin, S. van den Broucke, and J. Xiao, "A GAN-based hybrid sampling method for imbalanced customer classification", *Information Science*, Vol.609, 2022, pp. 1397-1411.

[68] Zhu, J. Y., T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks", *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223-2232.

[69] Zhu, J.-Y., P. Krahenbuhl, E. Shechtman, and A. A. Efros. "Generative visual manipulation on the natural image manifold", *Proceedings of European Conference on Computer Vision*, 2016, pp. 597-613.

Information Systems Review
Volume 24 Number 4
November 2022

A Hybrid Oversampling Technique for Imbalanced Structured Data based on SMOTE and Adapted CycleGAN

Jung-Dam Noh* · Byounggu Choi**

Abstract

As generative adversarial network (GAN) based oversampling techniques have achieved impressive results in class imbalance of unstructured dataset such as image, many studies have begun to apply it to solving the problem of imbalance in structured dataset. However, these studies have failed to reflect the characteristics of structured data due to changing the data structure into an unstructured data format. In order to overcome the limitation, this study adapted CycleGAN to reflect the characteristics of structured data, and proposed hybridization of synthetic minority oversampling technique (SMOTE) and the adapted CycleGAN. In particular, this study tried to overcome the limitations of existing studies by using a one-dimensional convolutional neural network unlike previous studies that used two-dimensional convolutional neural network. Oversampling based on the method proposed have been experimented using various datasets and compared the performance of the method with existing oversampling methods such as SMOTE and adaptive synthetic sampling (ADASYN). The results indicated the proposed hybrid oversampling method showed superior performance compared to the existing methods when data have more dimensions or higher degree of imbalance. This study implied that the classification performance of oversampling structured data can be improved using the proposed hybrid oversampling method that considers the characteristic of structured data.

Keywords: Imbalanced Data, Oversampling, Cycle Generative Adversarial Network, SMOTE, Structured Data

* Junior, Afreeca TV
** Corresponding Author, Professor, Kookmin University, College of Business Administration

◐ 저 자 소 개 ◑

노 정 담 (wjdeka93@kookmin.ac.kr)

현재 Afreeca TV VOD 데이터 팀에 재직 중이다. 국민대학교 데이터사이언스 석사를 취득하였다. 주요 연구분야는 데이터사이언스, 딥러닝, 등이며 경영정보학회 학술대회에서 발표를 하였다.

최 병 구 (h2choi@kookmin.ac.kr)

현재 국민대학교 경영대학 경영학부 교수로 재직 중이다. KAIST경영공학 석사 및 박사학위를 취득하였다. 국민대학교에 부임하기 전에는 University of Sydney, School of Information Technologies에서 조교수로 재직하였다. 주요 연구분야는 지식경영, 소셜미디어 어낼리틱스, 데이터사이언스 등이며 지금까지 이와 관련하여 Journal of Association for the Information Systems, Journal of MIS, IEEE Transactions on Engineering Management, I&M, APJIS, Information Systems Review, 지식경영연구 등을 포함한 다수의 국내 외 학술지에 논문을 게재하였다.

논문접수일 : 2022년 10월 19일 **게재확정일 : 2022년 11월 17일**
1차 수정일 : 2022년 11월 10일