KIPS Trans. Softw. and Data Eng.
Vol.8, No.7 pp.311~316 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.7.311

A Deep Learning Based Over-Sampling Scheme for Imbalanced Data Classification

Son Min Jae† · Jung Seung Won†† · Hwang Een Jun†††

ABSTRACT

Classification problem is to predict the class to which an input data belongs. One of the most popular methods to do this is training a machine learning algorithm using the given dataset. In this case, the dataset should have a well-balanced class distribution for the best performance. However, when the dataset has an imbalanced class distribution, its classification performance could be very poor. To overcome this problem, we propose an over-sampling scheme that balances the number of data by using Conditional Generative Adversarial Networks (CGAN). CGAN is a generative model developed from Generative Adversarial Networks (GAN), which can learn data characteristics and generate data that is similar to real data. Therefore, CGAN can generate data of a class which has a small number of data so that the problem induced by imbalanced class distribution can be mitigated, and classification performance can be improved. Experiments using actual collected data show that the over-sampling technique using CGAN is effective and that it is superior to existing over-sampling techniques.

Keywords : Imbalanced Data, CGAN, Deep Learning, Over-Sampling

불균형 데이터 분류를 위한 딥러닝 기반 오버샘플링 기법

손 민 재† · 정 승 원†† · 황 인 준†††

요 약

분류 문제는 주어진 입력 데이터에 대해 해당 데이터의 클래스를 예측하는 문제로, 자주 쓰이는 방법 중의 하나는 주어진 데이터셋을 사용하여 기계학습 알고리즘을 학습시키는 것이다. 이런 경우 분류하고자 하는 클래스에 따른 데이터의 분포가 균일한 데이터셋이 이상적이지만, 불균형한 분포를 가지고 경우 제대로 분류하지 못하는 문제가 발생한다. 이러한 문제를 해결하기 위해 본 논문에서는 Conditional Generative Adversarial Networks(CGAN)을 활용하여 데이터 수의 균형을 맞추는 오버샘플링 기법을 제안한다. CGAN은 Generative Adversarial Networks(GAN)에서 파생된 생성 모델로, 데이터의 특징을 학습하여 실제 데이터와 유사한 데이터를 생성할 수 있다. 따라서 CGAN이 데이터 수가 적은 클래스의 데이터를 학습하고 생성함으로써 불균형한 클래스 비율을 맞추어 줄 수 있으며, 그에 따라 분류 성능을 높일 수 있다. 실제 수집된 데이터를 이용한 실험을 통해 CGAN을 활용한 오버샘플링 기법이 효과가 있음을 보이고 기존 오버샘플링 기법들과 비교하여 기존 기법들보다 우수함을 입증하였다.

키워드 : 불균형 데이터, CGAN, 딥러닝, 오버샘플링

1. 서 론

분류 문제는 입력 데이터가 주어졌을 때 해당 데이터의 클래스를 예측하는 문제를 말한다. 일반적으로 이러한 문제를 해결하기 위해 기계학습 알고리즘을 주어진 데이터셋으로 학습시킨다. 이때 이상적인 데이터셋은 분류하고자 하는 클래스의 분포가 균일해야 한다. 그러나 데이터셋 대부분은 클래스마다 데이터 수의 차이가 존재하며 심한 경우에는 클라스 하나에만 데이터가 편중되기도 한다. 기계학습 알고리즘들은 각 클래스의 비율이 비슷한 상황을 가정하기 때문에, 클래스가 불균형한 데이터셋의 경우 전체적인 데이터에 대해 제대로 학습하지 못하고 큰 비중을 차지하는 클래스에 편향되어 학습한다[1]. 그 결과 전체적인 정확도는 높으나 정작 원하는 항목에 대해서는 분류해내지 못하는 클래스 불균형 현상이 발생된다.

예를 들어, 병원에서 암 진단 검사를 받은 환자에 대한 데이터를 수집하는 경우, 검사에서 양성 반응을 보이는 환자의 수보다 음성 반응을 보이는 환자의 수가 훨씬 많으므로 음성 환자에 대한 데이터양이 양성 환자에 대한 데이터양보다 매우 크다. 이 상황에서 양성 환자를 분류해 내기 위해 기계학습 알고리즘을 적용한다면 대부분 음성 환자 중심으로 편향된 학습을 하여 양성 환자에 대한 특징은 학습하지 못하게 된다[2]. 또한, 테스트 시 음성 환자에 대해서는 높은 정확도로 분류해낼 수 있지만, 양성 환자는 대부분 음성 환자로 판단하게 된다. 이 외

※ 본 결과물은 환경부의 재원으로 한국환경산업기술원의 환경정책기반공공기술개발사업의 지원을 받아 연구되었습니다(2017000210001).
† 준 회 원 : 고려대학교 전기전자공학과 석사과정
†† 준 회 원 : 고려대학교 전기전자공학과 석·박사통합과정
††† 종신회원 : 고려대학교 전기전자공학과 교수
Manuscript Received : February 12, 2019
Accepted : May 7, 2019
* Corresponding Author : Hwang Een Jun(ehwang04@korea.ac.kr)

에도 금융 범죄 탐지[3], 고객 수요 탐지[4] 등 다양한 분야에서 이러한 클래스 불균형 현상이 발생한다.

클래스 불균형을 해결하기 위해 주로 사용되는 방법으로 데이터 샘플링(Data Sampling) 기법이 있다. 데이터 샘플링은 불균형한 데이터 집합에서 대부분을 차지하는 클래스인 다수 클래스(Majority Class)와 반대로 적은 부분만 차지하는 소수 클래스(Minority Class)의 샘플 개수를 조정하여 균형 있는 데이터 집합으로 만드는 기법으로, 두 클래스중 어느 클래스의 샘플 개수를 조절하느냐에 따라 언더샘플링(Under-sampling) 기법과 오버샘플링(Over-sampling) 기법으로 분류된다[5].

언더샘플링은 소수 클래스의 샘플 수에 맞도록 다수 클래스의 샘플들을 제거하는 방식이다. 언더샘플링 기법으로는 무작위로 다수 클래스의 샘플을 제거하는 랜덤 언더샘플링(Random Under-sampling), 다수 클래스 샘플에서 독립적으로 뽑은 부분 집합과 소수 클래스 집합으로 분류 모델을 학습시키는 EasyEnsemble[6] 등의 기법이 제안되었다. 그러나 언더샘플링 기법들은 데이터를 제거하기 때문에 정보의 손실을 초래하게 된다는 문제점이 있다.

오버샘플링은 언더샘플링과는 반대로 다수 클래스 샘플 개수에 맞춰 소수 클래스를 위한 샘플을 생성하는 방식으로, 정보 손실을 피할 수 있다. 오버샘플링 방법에는 무작위로 소수 클래스 샘플을 생성하는 랜덤 오버샘플링(Random Over-sampling), k-NN(k-nearest neighbors) 알고리즘을 활용하여 소수 클래스 샘플에서 이웃을 찾고 그 사이에 속하게 될 새로운 샘플을 합성하는 SMOTE(Synthetic Minority Over-sampling Technique)[7] 등이 있다. 그 외에도, SMOTE에 소수 클래스 주변의 다수 클래스 밀도에 따라 가중치를 부여하는 개념을 추가시킨 ADASYN(Adaptive synthetic sampling)[8], 소수 클래스 샘플과 다수 클래스 샘플 간의 경계를 기준으로 SMOTE 방식의 합성을 하는 Borderline-SMOTE[9] 등 다양한 기법이 존재한다.

하지만 오버샘플링 기법들 역시 단점을 가지고 있다. 랜덤 오버샘플링 같은 데이터 복제 방식은 동일한 데이터를 반복 학습하기 때문에 분류 모델이 학습 데이터에 과적합(Overfitting)될 수 있다. k-NN 알고리즘을 활용한 데이터 합성 방식은 소수 클래스 샘플의 밀도가 낮은 데이터에 적용할 경우 샘플을 생성하는 과정에서 다수 클래스 샘플 범위에 속하는 데이터를 생성하게 되고 이러한 데이터가 분류 학습에 노이즈로 작용해 분류 성능을 떨어뜨린다.

본 논문에서는 데이터 클래스 불균형을 해결하기 위한 새로운 방법으로 CGAN(Conditional Generative Adversarial Networks)[10] 기반의 오버샘플링 기법을 제안한다. CGAN은 GAN(Generative Adversarial Networks)[11]에서 발전된 모델로 GAN과의 차이점은 사용자가 원하는 특징을 반영하여 학습시켜 원하는 방향대로 데이터를 생성할 수 있다는 것이다. 이러한 점을 활용하여 CGAN 모델에 소수 클래스 샘플의 특징을 학습시키고 가상의 소수 클래스 샘플을 합성함으로써 다수 클래스와 소수 클래스의 샘플 수 차이를 없앤다.

본 논문의 구성은 다음과 같다. 2장에서는 오버샘플링 기법에 대한 연구를 기술하고, 3장에서는 본 논문에서 제안하는 기법에 대해 설명한다. 4장에서는 제안된 모델을 실제 불균형 데이터를 사용하여 기존의 샘플링 기법과 비교하는 실험을 진행한다. 5장에서는 결론에 대해 요약하고, 향후 연구 방향을 제시한다.

2. 관련 연구

오버샘플링 기법은 데이터를 어떻게 복제할 것인지, 어떻게 생성할 것인지에 관해 주로 연구되어 왔다. Chawla 등[7]은 소수 클래스 샘플을 선택하고 k-NN 알고리즘을 활용하여 이웃을 찾아 그사이에 새로운 데이터를 합성하는 방식인 SMOTE를 제안하였다. Haibo 등[8]은 SMOTE를 발전시켜 소수 클래스 샘플 선택 시 샘플 주변의 다수 클래스 샘플의 밀도에 따라 가중치를 부여하는 개념을 더한 ADASYN을 선보였다. Han 등[9] 역시 SMOTE를 기반으로 연구를 수행하여, 소수 클래스 샘플 집합과 다수 클래스 샘플 집합 간의 경계를 기준으로 SMOTE 방식의 합성을 하는 Borderline-SMOTE을 제안하였다. Nguyen 등[12]은 Support Vector Machine(SVM)[13]을 활용해 소수 클래스 샘플 분포와 다수 클래스 샘플 분포 사이의 경계를 찾고, 이를 중심으로 소수 범주를 샘플링하는 방식을 제안하였다. Japowicz[14]은 클래스별로 학습 데이터를 군집화(Clustering)하고 군집마다 오버샘플링을 수행하는 방식으로, 두 클래스 사이에 발생하는 불균형뿐만 아니라 클래스 내부에 발생하는 불균형도 해결하는 오버샘플링 기법을 제안했다. Macia 등[15]은 최소 신장 트리(Minimum Spanning Tree) 기반으로 두 클래스 사이 경계의 복잡성을 계산하고, 데이터 차원수를 고려하여 데이터를 샘플링하는 기법을 제안하였다. Wang[16]은 B-SVM(Biased Support Vector Machine)[17]의 학습 결과로 나온 소수 클래스 지원 벡터로만 SMOTE를 적용하여 오버샘플링하였다.

언더샘플링과 혼합한 방법도 제안되었다. Batista 등[18]은 Wilson's Edited Nearest Neighbor Rule[19], Tomek links[20] 같은 언더샘플링을 수행 후 SMOTE를 적용하는 샘플링 기법을 제안하였다. Liu 등[21]은 SMOTE로 소수 클래스 샘플을 일정 수 오버샘플링한 후 다수 클래스 샘플을 언더샘플링하여 두 클래스 샘플 수의 균형을 맞추는 방식을 제안하였다. 본 논문에서는 딥러닝 생성 모델 중 하나인 CGAN을 활용하여, 소개한 기존 기법들과 다르게 오버샘플링하였으며 기존 기법들보다 뛰어난 분류 성능을 보였다.

3. 딥러닝 기반의 생성 모델

3.1 CGAN

CGAN은 GAN에서 파생된 모델이다. GAN은 딥러닝 생성 모델 중 하나로, 가상의 데이터를 생성해내는 생성기(Generator)와 생성된 데이터와 실제 데이터를 구분하는 분류기(Discriminator)로 구성되어 있다. 생성기는 분류기가 실제 데이터로 판단할 수 있도록 정교한 데이터를 생성해야 하는 한편, 분류기는 생성기가 만들어낸 데이터와 실제 데이터

를 확실히 판단할 수 있어야 한다. 상반되는 목적으로 인해 이 두 모듈은 서로 경쟁하는 방식으로 학습한다.

이러한 학습을 진행하기 위한 GAN의 목적 함수는 Equation (1)과 같다. p_{data}는 실제 데이터의 분포, x는 실제 데이터 분포인 p_{data}에서 뽑은 표본을 나타낸다. 마찬가지로 p_z는 노이즈의 분포, z는 p_z에서 뽑은 노이즈 표본을 의미한다. G는 생성기를 의미하며, $G(z)$는 생성기가 z를 입력으로 받아 생성한 데이터를 말한다. 한편, D는 분류기 모듈로, 입력 데이터를 실제 데이터라 판단한 경우 1을, 실제 데이터가 아니라고 판단한 경우 0을 출력해야 한다. 예를 들어, $D(x)$는 실제 데이터 표본인 x가 입력되어 1을 출력해야 하는 반면, $D(G(z))$는 생성된 데이터가 입력되었기 때문에 0을 출력해야 한다.

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

분류기 입장에서는 $D(x)$가 1, $D(G(z))$가 0이 되는 것이 이상적이므로 $V(D,G)$를 최대화하려 시도한다. 반대로, 생성기 입장에서는 분류기를 속여 자신이 생성한 데이터를 분류기가 실제 데이터라고 판단하게 만들어야 하므로 $D(G(z))$가 1이 되는 것이 이상적이고 이에 따라 $V(D,G)$를 최소화하려 시도한다.

본 논문에서 사용되는 모델인 CGAN은 이 GAN모델에서 발전된 모델로, 기본적인 학습방식은 GAN과 동일하다. GAN과의 다른 점은 분포와 더불어 특징을 같이 학습시킬 수 있다는 점이다. 특징을 학습시키기 위한 손실 함수는 Equation (2)와 같다. 식에서 볼 수 있듯 y라는 특징을 반영할 수 있도록 조건이 추가되었다. CGAN의 분류기는 GAN의 분류기처럼 입력 데이터가 진짜 데이터인지 만들어진 가짜 데이터인지를 판별하는 역할을 하지만 y를 고려하여 구분하게 된다. 생성기 역시 y의 성질을 가지도록 데이터를 생성하도록 학습한다. 이러한 CGAN의 특징을 활용하여 소수 클래스의 특징을 갖는 샘플을 생성하는 오버샘플링 모델로 활용한다.

$$V(D,G) = E_{x \sim p_z}[\log D(x|y)] + E_{z \sim p_z(z)}[\log(1 - D(G(z|y)))] \quad (2)$$

3.2 CGAN 학습

CGAN은 사용자가 원하는 특징을 반영하여 데이터를 생성할 수 있다. 따라서 본 논문에서는 이를 활용해 소수 클래스 데이터를 생성하도록 하였다. 데이터 생성 과정은 Fig. 1과 같다. 우선 가우시안 분포(Gaussian distribution)에서 임의로 선택한 노이즈와 소수 클래스 또는 다수 클래스를 식별할 수 있는 클래스 정보를 입력 데이터로 넣는다. 생성기는 입력된 데이터를 활용해 각 클래스의 데이터를 생성하도록 학습한다. 이에 반해 분류기는 각 클래스의 실제 데이터와 생성기에서 생성된 데이터를 구별하도록 학습한다. 이런 학습 과정을 반복함으로써 CGAN의 생성기는 소수 클래스의 데이터 특징을 반영하여 데이터를 생성해낼 수 있다.

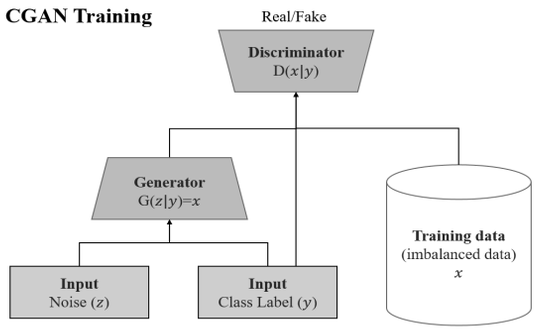

Fig. 1. CGAN Training Process

3.3 CGAN을 활용한 오버샘플링

데이터 특징을 학습한 CGAN을 활용하여 오버샘플링하고 분류기의 학습을 진행하며 그 진행 과정은 Fig. 2와 같다. 학습된 CGAN의 생성기에 노이즈와 소수 클래스라는 클래스 정보를 입력하면, 생성기는 소수 클래스 데이터와 유사한 데이터를 생성한다. 이를 학습 데이터의 다수 클래스 샘플 수와 소수 클래스의 샘플 수 차이만큼 반복하여 실제 소수 클래스 데이터와 생성한 소수 클래스 데이터를 합한 수가 다수 클래스의 수와 같아지도록 만든다. 생성된 데이터와 학습 데이터를 합쳐 데이터 집합을 구성한 후 이 집합을 기계학습 및 딥러닝 알고리즘의 학습 데이터로 사용해 분류를 수행한다.

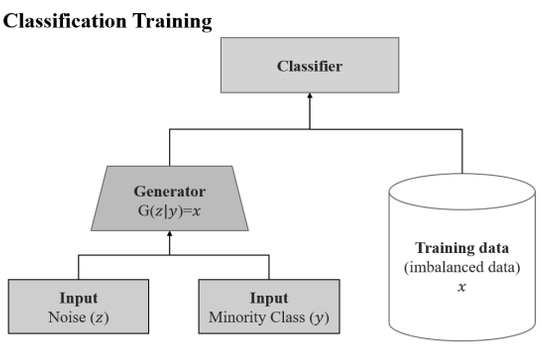

Fig. 2. Classification Training Process

4. 실험 및 결과

4.1 데이터셋

본 논문에서 제안하는 CGAN을 활용한 오버샘플링 기법을 검증하기 위해 불균형한 데이터셋을 수집하고 데이터셋 내 클래스를 분류하는 실험을 진행하였다. 실험에 사용된 데이터 집합은 세 종류로, 첫 번째 데이터는 카드 사기 데이터(Card)[22], 두 번째는 이메일 마케팅 데이터(E-mail)[23], 세 번째는 텔레마케팅 데이터(Tel)[24]이다. 카드 사기 데이터셋은 총 284,807개의 데이터로 구성되어 있고, 보안상 처리된 28개의 변수와 금액과 시간을 포함한 총 30개의 변수를 데이터 특징으로, 사기 여부를 클래스로 가진다. 클래스 분포는 사기 492건, 비사기 284,315건으로 매우 불균형하다. 이메일 마케팅 데이터셋은 쇼핑몰 사이트에서 수집한 64,000개의 데이터로, 사이트에서 지출한 총 금액, 신규 가입 여부, 홍보 이메일 종류 등 9개의 특징

Table 1. Used Datasets

Dataset	Attributes	Instances	Majority Class	Minority Class
Card	30	284807	284315	492
E-mail	9	64000	54606	9394
Tel	20	41188	36548	4640

으로 구성되어있다. 클래스는 이메일을 받은 후의 방문 여부로, 64,000개 중 9,394개만이 방문한 데이터다. 마지막으로 텔레마케팅 데이터셋은 총 41,188개의 데이터로, 나이, 성별, 적금 유무, 연락 일자 등 20개의 특징으로 구성되어 있다. 클래스는 구매 여부로, 4,640개만 구매 데이터이고 나머지는 비구매 데이터다. 이러한 데이터 특성은 Table 1에 요약되어 있다.

4.2 구현

CGAN을 활용한 오버샘플링 기법과의 비교를 위해 본 논문에서 고려한 기법으로는 Random Over-sampling(ROS), SMOTE, Borderline-SMOTE(B-SMOTE), ADASYN, GAN을 활용한 오버샘플링(GAN-OS) 등이 있다. 한편, 분류 모델로는 SVM, Random Forest(RF)[25], 그리고 Multi-Layer Perceptron(MLP)[26]을 사용하였다. 실험은 모두 Python 3.5에서 수행되었고, 라이브러리는 sklearn 0.19.1, imblearn 0.4.3를 사용하였다.

오버샘플링 기법과 분류기에 사용된 하이퍼 파라미터는 다음과 같다. K-NN 기반인 SMOTE와 Borderline-SMOTE, ADASYN 모두 k=5를 사용하였다. GAN과 CGAN의 경우 생성기, 분류기 모두 Hidden Layer 2층으로 구성되었고 Activation Function과 Optimizer는 각각 ReLU[27], Adam[28]을 사용하였다. 분류기로 사용된 SVM은 C=1.0, 커널(Kernel)은 Radial Basis Function으로 설정하였고, RF는 100개 트리, gini index를 불순도 지표로 설정하였다. MLP 모델의 Activation Function과 Optimizer는 CGAN과 동일하나 은닉층은 3층으로 구성하였다. 노드의 수는 데이터 특징의 수에 따라 노드 수를 다르게 설정하였다.

분류기의 정확도를 측정하기 위한 평가 척도는 클래스 불균형 데이터에 가장 많이 이용되는 Area Under the ROC Curve(AUC)를 사용하였다[2]. ROC Curve는 x축을 False positive rate, y축을 True positive rate로 두고, 문턱값 변화에 따라 False positive rate와 True positive rate의 변화를 그린 그래프를 의미하며, ROC Curve의 아래 면적을 AUC라 한다. AUC가 높으면 높을수록 안정적인 모델로 평가된다.

4.3 데이터 생성

Fig. 3은 CGAN이 실제 데이터 분포를 제대로 학습하는지 확인하기 위해 학습 진행에 따라 CGAN에서 생성된 데이터 분포를 시각화한 그림이다. 시각화를 위해 주성분 분석(Principal Component Analysis)[29]으로 주성분을 추출하고, x축에 첫 번째 주성분을, y축에 두 번째 주성분을 표시한 것이다. (a), (b), (c)는 차례로 각각 CGAN 학습 초반, 학습 중반, 학습 후반 데이터 분포를 나타내며, (d)는 원 데이터의 분포이다. 학습 초기에는 원래 데이터 분포와 다른 데이터들이 생성되지만, 학습 중반과 후반에는 기존 분포와 유사하게 생성하고 있는 것을 확인할 수 있다. 데이터셋마다 유사한 분포를 생성할 수 있을 때까지 최대 13,000 Epoch만큼 학습을 진행하였다. 하지만 E-mail 데이터셋의 경우 학습 횟수를 더욱 늘려도 기존 데이터의 분포에서 벗어난 데이터를 생성하였다. 이러한 원인은 데이터 특징 차원이 9로, 다른 데이터셋에 비해 작아서 데이터 분포를 정밀하게 학습하지 못한 것이라 추측된다.

4.4 성능 평가

Fig. 4는 샘플링 적용 없이 분류 학습을 진행했을 때 AUC 결과 대비 CGAN 오버샘플링 후 분류 학습을 진행하여 얻어진 AUC를 퍼센트로 환산해 나타낸 그래프이다. x축은 데이터별 분류기 종류, y축은 AUC 비율을 나타낸다. 분류기별로 모든 데이터가 샘플링 후 성능이 향상됨을 볼 수 있다. Card

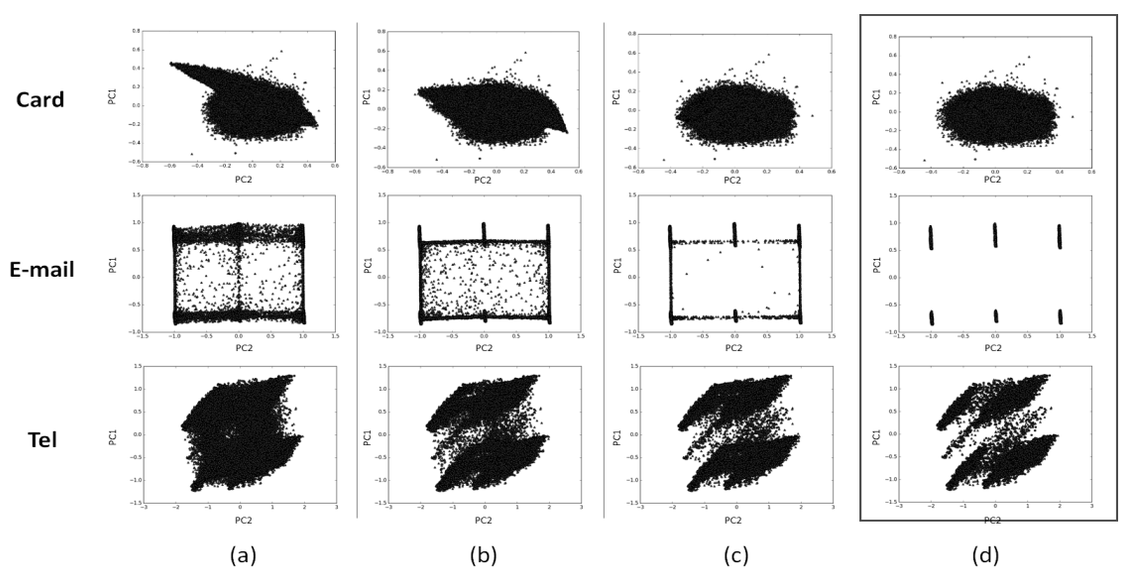

Fig. 3. Data Distribution Generated by Cgan

Table 2. Performance Comparison of Over-sampling Methods and Classification Models

Dataset		Base	ROS	SMOTE	ADASYN	B-SMOTE	GAN	CGAN
Card	SVM	0.910	0.950	**0.954**	0.951	0.952	0.941	0.953
	RF	0.850	0.859	0.867	0.857	0.849	0.857	**0.878**
	MLP	0.930	0.959	0.961	0.964	0.959	0.965	**0.971**
E-mail	SVM	0.495	0.631	0.631	0.636	**0.636**	0.629	0.632
	RF	0.566	0.562	0.559	0.562	0.562	0.563	**0.588**
	MLP	0.564	0.626	0.615	0.630	0.630	0.632	**0.637**
Tel	SVM	0.697	0.784	0.784	0.785	**0.785**	0.631	0.779
	RF	0.770	0.771	0.763	0.766	0.766	0.771	**0.773**
	MLP	0.753	0.779	0.780	0.780	0.780	0.779	**0.789**

데이터의 경우 모든 분류기에서 평균 4%의 AUC 향상을 보였고, E-mail 데이터의 경우 SVM에서 27%, MLP에서 13%로 대폭 향상되었다. Tel 데이터는 SVM에서 12%로 크게 향상했으나, RF는 0.4%로 거의 변화가 없었다.

Table 2는 기존 오버샘플링 기법과의 성능을 비교한 표이다. 표에서 볼 수 있듯이 본 논문에서 제안하는 오버샘플링 기법이 기존 오버샘플링 기법들에 비해 전반적으로 우수한 성능을 보임을 확인할 수 있다. 특히, RF와 MLP에서 CGAN을 활용한 오버샘플링이 효과적이며, SVM에서도 안정된 분류 성능을 보였다. 좀 더 세밀한 비교를 위해 Wilcoxon signed-rank test[30]를 진행하였다. Wilcoxon signed-rank test는 두 분류기 사이에 유의한 차이가 있는지 판단하기 위해 차이가 없다는 것을 귀무가설로 설정한 후 이를 검증하는데 이용되었다. p-value 값이 유의수준 보다 작을 경우에는 귀무가설을 기각하며 두 분류기는 유의한 차이가 있다고 판단하게 된다. 본 연구에서 유의수준을 .05로 설정하고 Wilcoxon signed-rank test를 수행한 결과는 Table 3과 같다.

Table 3. Result of Wilcoxon signed-rank test

Over-sampling Scheme		p-value (< .05)
CGAN	ROS	0.0273
	SMOTE	0.0378
	ADASYN	0.0328
	B-SMOTE	0.0438
	GAN	0.0039

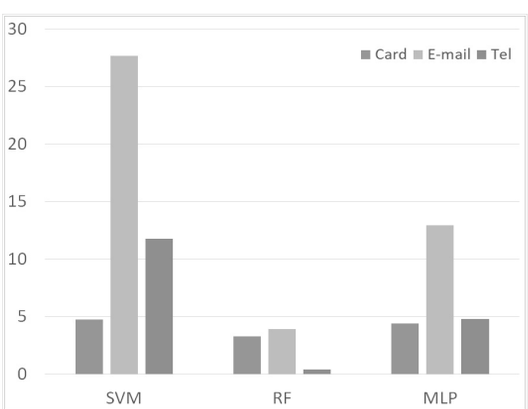

Fig. 4. AUC Comparison of Classification Models

5. 결론 및 향후 계획

본 논문에서는 클래스 분포가 불균형한 데이터의 분류를 위해 클래스별 특징을 반영하여 데이터를 생성할 수 있는 CGAN을 활용한 오버샘플링 기법을 제안하였다. 적절한 CGAN의 학습 횟수를 도출하기 위해 주성분을 뽑아 시각화했으며, 생성된 데이터의 유의미함을 검증하기 위해 SVM, RF, MLP 분류기를 사용해 분류 학습을 진행하였다. 실험 결과 기존의 오버샘플링 기법 대비 우수한 분류 성능을 보였고, Wilcoxon signed-rank test를 통해 성능 향상의 유의미함을 입증하였다.

향후 연구에서는 CGAN의 샘플을 생성하는데 있어, 보다 유의미한 소수 클래스 샘플을 찾아내 그를 대상으로 샘플링하는 연구를 진행할 것이다. 또한 다른 더 많은 불균형 데이터에 실험을 통해 보다 최적화된 모델을 구축할 계획이다.

References

[1] R. O'Brien and H. Ishwaran, "A random forests quantile classifier for class imbalanced data," *Pattern Recognition*, Vol.90, pp.232-249, 2019.

[2] G. Haixiang, et al., "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, Vol.73, pp.220-239, 2017.

[3] S. J. Salvatore, et al., "Cost-based modeling for fraud and intrusion detection: Results from the JAM project," in *Proceedings of the DARPA Information Survivability Conference and Exposition*, Washington, pp.130-144, 2000.

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

[5] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the ACM International Conference on Machine Learning*, New York, pp.935-942, 2007.

[6] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.39, No.2, pp.539-550, 2009.

[7] N. V. Chawla, et al., "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, Vol.16, pp.321-357, 2002.

[8] H. He, et al., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp.1322-1328, 2008.
[9] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Proceedings of the International Conference on Intelligent Computing*, Berlin, pp.878-887, 2005.
[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[11] I. J. Goodfellow, et al., "Generative adversarial nets," in *Proceedings of the Neural Information Processing Systems*, pp.2672-2680, 2014.
[12] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline Over-Sampling for Imbalanced Data Classification," *International Journal of Knowledge Engineering and Soft Data Paradigms*, Vol.3, No.1, pp.4-21, 2011.
[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, Vol.20, No.3, pp.273-297, 1995.
[14] T. Jo and N. Japkowicz, "Class Imbalances versus Small Disjuncts," *ACM Special Interest Group on Knowledge Discovery in Data*, Vol.6, pp.40-49, 2004.
[15] N. Macia, E. Bernado-Mansilla, and A Orriols-Puig, "A.Preliminary Approach in Synthetic Data Sets Generation based on Class Separability Measure," in *Proceedings of the International Conference on Pattern Recognition*, pp.1-4, 2008.
[16] H. Y. Wang, "Combination Approach of SMOTE and Biased-SVM for Imbalanced Datasets," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 228-231, 2008.
[17] C. H. Hoi, et al., "Biased support vector machine for relevance feedback in image retrieval," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2004.
[18] G. E. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explorations Newsletter*, Vol.6, No.1, pp.20-29, 2004.
[19] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Transactions on Systems Man and Communications(SMC)*, Vol.3, pp.408-421, 1972.
[20] I. Tomek, "Two Modifications of CNN," *IEEE Transactions on Systems Man and Communications(SMC)*, Vol.6, pp.769-772, 1976.
[21] Y. Liu, A. An, and X. Huang, "Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles," in *Proceedings of the Tenth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, pp. 107-118, 2006.
[22] A. Dal Pozzolo and G. Bontempi, "Adaptive machine learning for credit card fraud detection." 2015.
[23] MineThatData [Internet], http://www.minethatdata.com.
[24] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, Vol.62, pp.22-31, 2014.
[25] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News,* Vol.2, No.3, pp.18-22, 2002.
[26] S. S. Haykin, "Neural networks and learning machines," Vol. 3, Upper Saddle River:Pearson, 2009.
[27] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines." in *Proceedings of the 27th international conference on machine learning (ICML-10),* Haifa, pp.807-814, 2010.
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." in *Proceedings of 3rd International Conference on Learning Representations,* San diego, 2014.
[29] I. Jolliffe, "Principal component analysis," *International Encyclopedia of Statistical Science*, Springer, Berlin, Heidelberg, pp.1094-1096, 2011.
[30] G. W. Corder and D. I. Foreman, "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach," John Wiley & Sons, 2014.

손 민 재

https://orcid.org/0000-0001-5814-5892
e-mail : smj5668@korea.ac.kr
2017년 동서대학교 컴퓨터공학과(학사)
2017년~현 재 고려대학교 전기전자공학과 석사과정
관심분야 : 불균형 데이터 처리, 수요 예측, 기계 합습, 데이터 마이닝

정 승 원

https://orcid.org/0000-0002-8869-0426
e-mail : jsw161@korea.ac.kr
2016년 고려대학교 전기전자전파공학부 (학사)
2016년~현 재 고려대학교 전기전자공학과 석·박사통합과정
관심분야 : 기계 학습, 데이터 마이닝, 데이터베이스, 추천 시스템

황 인 준

https://orcid.org/0000-0002-0418-4092
e-mail : ehwang04@korea.ac.kr
1988년 서울대학교 컴퓨터공학과(학사)
1990년 서울대학교 컴퓨터공학과(석사)
1998년 Univ. Maryland at College Park 전산학과(박사)
1998년~1999년 Bowie State Univ. 조교수
1999년 Hughes Research Lab. 연구교수
1999년~2004년 아주대학교 정보통신전문대학원 조/부교수
2004년~현 재 고려대학교 전기전자공학과 교수
관심분야 : 데이터베이스, 멀티미디어 시스템, 정보 검색, 빅데이터 처리, 의료 어플리케이션