ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 September: 27(3): 157~173 http://dx.doi.org/10.13088/jiis.2021.27.3.157

Conditional Generative Adversarial Network(CGAN) 기반 협업 필터링 추천 시스템*

강소이
이화여자대학교 일반대학원 빅데이터분석학
(dlthrkd23@ewhain.net)

신경식
이화여자대학교 경영대학
(ksshin@ewha.ac.kr)

소비자의 욕구와 관심에 맞추어 개인화된 제품을 추천하는 추천 시스템은 비즈니스에 필수적인 기술로서의 그 중요성이 증가하고 있다. 추천 시스템의 대표적인 모형 중 협업 필터링은 우수한 성능으로 다양한 분야에서 활용되고 있다. 그러나 협업필터링은 사용자-아이템의 선호도 정보가 충분하지 않을 경우 성능이 저하되는 희소성의 문제가 있다. 또한 실제 평점 데이터의 경우 대부분 높은 점수에 데이터가 편향되어 있어 심한 불균형을 갖는다. 불균형 데이터에 협업 필터링을 적용할 경우 편향된 클래스에 과도하게 학습되어 추천 성능이 저하된다. 이러한 문제를 해결하기 위해 많은 선행연구들이 진행되어 왔지만 추가적인 외부 데이터 또는 기존의 전통적인 오버샘플링 기법에 의존한 추천을 시도하였기에 유용성이 떨어지고 추천 성능 측면에서 한계점이 있었다.

본 연구에서는 CGAN을 기반으로 협업 필터링 구현 시 발생하는 희소성 문제를 해결함과 동시에 실제 데이터에서 발생하는 데이터 불균형을 완화하여 추천의 성능을 높이는 것을 목표로 한다. CGAN을 이용하여 비어있는 사용자-아이템 매트릭스에 실제와 흡사한 가상의 데이터를 생성하여, 희소성을 가지고 있는 기존의 매트릭스로만 학습한 것과 비교했을 때 높은 정확도가 예상된다. 이 과정에서 Condition vector y를 이용하여 소수 클래스에 대한 분포를 파악하고 그 특징을 반영하여 데이터를 생성하였다. 이후 협업 필터링을 적용하고, 하이퍼파라미터 튜닝을 통해 추천 시스템의 성능을 최대화하는데 기여하였다. 비교 대상으로는 전통적인 오버샘플링 기법인 SMOTE, BorderlineSMOTE, SVM-SMOTE, ADASYN와 GAN을 사용하였다. 결과적으로 데이터 희소성을 가지고 있는 기존의 실제 데이터뿐만 아니라 기존 오버샘플링 기법들보다 제안 모형의 추천 성능이 우수함을 확인하였으며, RMSE, MAE 평가 척도에서 가장 높은 예측 정확도를 나타낸다는 사실을 증명하였다.

주제어 : 추천 시스템, 협업 필터링, 오버샘플링, 딥러닝, GAN

논문접수일 : 2021년 6월 1일 논문수정일 : 2021년 7월 5일 게재확정일 : 2021년 7월 14일
원고유형 : 일반논문 교신저자 : 신경식

1. 서론

정보기술의 발달로 정보의 양이 나날이 축적되고 있다. 다양한 아이템에 대한 선택의 폭이 넓어졌으나, 정보 과부하로 인해 사용자가 원하는 정보를 쉽게 찾기 어렵다는 문제점이 있다. 인터넷이 대중화되면서 사용자들은 가용한 여러 정보를 읽어보고 종합적으로 판단하는 대신 정보 검색 및 학습 시간을 절약해줄 수 있는 가시화된 시스템에 의존하고자 한다(Lee et al., 2010). 이에 따라 추천 시스템(recommendation system)은 소비자의 욕구와 관심에 따라 개인화된 제품

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A5A2A03067552)

을 추천하여 비즈니스에 필수적인 기술로서의 그 중요성이 증가하고 있다(Adomavicius et al., 2005).

협업 필터링(CF; Collaborative Filtering)(Goldberg et al., 1992)은 과거 사용자-아이템 관계를 학습하여 사용자 선호도를 예측하며 추천 시스템에서 가장 널리 사용되는 접근 방식 중 하나이다. 협업 필터링은 유사한 사용자간 관심사와 선호도를 바탕으로 추천이 이루어져, 우수한 성능으로 많은 대규모 상업 사이트에서 사용되고 있다. 그러나 실제 데이터에는 협업 필터링의 성능을 저해하는 여러 가지 문제가 있다.

특히 추천의 유용성을 제한하는 주요 문제로는 희소성(sparsity)이 있다. 협업 필터링에서는 사용자 또는 아이템의 유사도를 측정하기 위해 사용자-아이템 매트릭스를 사용하며, 행렬 값은 아이템에 대한 사용자의 평점을 이용한다. 사용자-아이템 매트릭스에서의 평가 값은 상품의 인기도에 따라 데이터가 편중되거나 아직 값을 평가하지 않은 새로운 사용자가 존재할 수 있다(Son et al., 2015). 이와 같이 소비자의 선호를 식별하기에 과거 데이터가 불충분한 문제를 데이터 희소성이라 하며 이러한 문제점을 해결하기 위해 다양한 연구가 진행되었다. 희소성 문제를 해결하기 위한 대부분의 연구는 사용자들의 소셜 네트워크 등 개인정보나 아이템에 대한 특성 등의 데이터를 이용하여 선호도를 예측한다. 그러나 이러한 방법은 추가적인 외부 데이터가 확보되어야만 적용할 수 있어 희소성 문제를 충분히 개선하지 못한다.

또 다른 주요한 문제로는 실제 평점 데이터의 경우 평점이 대부분 높은 점수에 편향되어 있어 심각한 데이터 불균형 문제가 발생한다는 것이다. (Chevalier et al., 2006)에 의하면 Amazon과 Barnes & Noble에 대한 사용자 평점은 평균 4.13-4.45 정도로 압도적으로 긍정적임을 볼 수 있다. 이러한 불균형 분포를 유발하는 한 가지 원인으로는 제품 평가가 높은 사용자만이 제품을 구매하기 때문에 평가가 낮은 사람은 제품을 구매할 가능성이 적고 이에 따라 부정적인 제품 리뷰를 남기지 않게 된다는 purchasing bias를 들 수 있다(Hu et al., 2009). 이러한 특징으로 인해 대부분의 사용자들의 실제 선호와 달리 제품을 구매한 사용자들의 리뷰는 긍정적일 가능성이 더 높다(Schoenmueller et al., 2018). 따라서 실제 평점 데이터는 편향적 특징으로 인해 발생 빈도가 많은 다수 클래스(majority class)에 과도하게 학습되어 시장을 왜곡한다. 이러한 불균형한 평점 매트릭스를 협업 필터링 모델에 적용할 경우 왜곡된 사용자의 선호도를 학습하여 추천의 정확도를 저하시킨다.

한편, 대부분의 기존 데이터 불균형 문제에 대한 전처리 방법은 이진 클래스를 대상으로 설계가 되고 사용된다. 이진 클래스 불균형 기법의 경우 2개 이상의 클래스 간 경계선상의 대상이 발생하거나 여러 클래스가 겹치는 등의 다중 클래스 문제를 모델링할 수 없기 때문에 다중 클래스 문제에 적용이 어렵다(Koziarski et al., 2020). 이를 해결하기 위하여 다중 클래스 문제를 두 개의 클래스 문제로 변환하여 적용하는 연구가 진행되었지만 다중 클래스 문제의 단순화는 선택된 아이템 이상의 관계에 대한 중요한 정보의 손실로 이어지며(Krawczyk et al., 2016), 다른 하위 문제에서 학습한 분류기의 결과를 결합할 경우 잠재적인 분류 오류가 발생할 수 있다(Jin et al., 2007). 따라서 다중 클래스 불균형 문제를 해결하기 위하여 더욱 효과적인 방법 개발이 필요하다.

딥러닝 기법 중 GAN(Generative Adversarial Networks)은 최근 이미지, 음성, 자연어 처리 등 데이터 생성 분야에서 좋은 성능을 나타내고 있다. 본 연구에서는 CGAN(Conditional Generative Adversarial Network)을 사용하여 협업 필터링 구현 시 발생하는 희소성 문제를 해결함과 동시에 실제 데이터에서 발생하는 데이터 불균형을 완화하여 최적의 추천 시스템 모형을 제안하고자 한다. 먼저 CGAN을 이용하여 비어있는 사용자-아이템 매트릭스에 대하여 실제와 흡사한 가상의 데이터를 생성한다. 이는 희소성을 가지고 있는 기존의 매트릭스로만 학습한 것과 비교했을 때 높은 정확도가 예상된다. 이 과정에서 Condition vector y를 이용하여 소수 클래스에 대한 분포를 파악하고 그 특징을 반영한 데이터 생성을 통해 클래스 불균형을 완화하였다. 이후 협업 필터링을 적용하였으며, 하이퍼파라미터 튜닝을 통해 추천 시스템의 성능을 최대화하는데 기여하였다. 나아가 데이터 희소성을 가지고 있는 기존의 실제 데이터뿐만 아니라, 다양한 오버샘플링 기법들과 추천 성과를 비교하여 제안 모형의 유용성을 검증하고자 한다.

2. 관련 연구

2.1. 협업 필터링

본 연구에서는 사용자에게 아이템을 추천할 때 추천의 정확도를 높이기 위해 희소성 문제를 해결하는 방법에 대하여 제안한다. 평점 정보가 부족한 실제 데이터를 기반으로 정확한 요인 행렬을 구할 수 없다는 점은 협업 필터링의 추천 성능을 저하시키는 주요한 원인이다. 희소성 문제를 개선하기 위한 방법에는 크게 두 가지가 있다. 데이터 내의 누락된 정보를 추가적인 선호 정보들을 이용하여 이를 대치하는 방법과 아이템을 유사한 아이템 또는 사용자들을 군집화하여 추천을 진행하는 방법이다. 희소성 문제는 많은 연구자들에 의해 연구되었지만 대부분의 연구는 추가적인 외부 데이터가 확보되어야만 적용할 수 있어 희소성 문제를 충분히 개선하지 못하므로 추가적인 연구가 필요한 실정이다.

(Gong et al., 2009)에서는 BP neural networks 및 아이템 기반 협업 필터링을 기반으로 데이터 희소성 문제를 해결하는 새로운 추천 방법을 제안하였다. 비어있는 매트릭스 정보에 대하여 BP neural networks를 이용하여 빈 등급을 채우고 협업 필터링에 적용하여 추천을 진행하였다. (Billsus et al., 1998)은 특이값 분해(Singular Value Decomposition) 기술이 결합된 학습 알고리즘을 제안하여 대표적이지 않은 정보들을 제거를 통해 희소성 문제를 해결하였다. 영화 데이터를 사용하여 전통적 협업 필터링의 추천 성능을 향상시키는 것을 확인하였으나 이 방법은 특정한 사용자 또는 아이템이 제거됨으로써 유용한 정보를 잃어 추천의 성능이 떨어질 가능성이 있다는 문제점이 존재한다. (Ahn et al., 2009)은 웹에 존재하는 방대한 문화적 메타 데이터를 활용하여 영화 추천 시스템을 제안하였으며, 문화적 메타 데이터의 잠재력을 확인하였다. (Melville et al., 2002)에서는 사용자-아이템 행렬이 희소한 상황에서 내용기반(content-based) 방법을 활용해 빈 항목을 예측하여 채워 넣는 접근법을 제안하였다. 협업 필터링과 내용기반 방법의 구성 요소를 통합한 하이브리드 추천 시스템은 각각의 방법의 고유한 단점을 해결하여 기존의 기법들보다 더 나은 성능을 나타내는 것을 확인하였다. 그러나 이는 아이템

정보에 대한 접근이 필요하지 않은 협업 필터링의 장점을 제공하지 못한다는 문제가 있다. (Lee et al., 2017)에서는 사용자와 아이템의 관계를 정확히 이해하기 힘든 상황에서 무관심한 아이템을 이용하는 방법을 제안하였다. 사용자가 평가하지 않은 상품의 평점을 추론하여 평점 매트릭스를 채우는 접근법을 사용하였다. (Kwon et al., 2013)에서는 사용자의 명시적 선호와 함께 유사도 평가의 대상이 되는 두 객체의 고유한 속성을 특징 벡터로 사용하였다. 영화 아이템 장르 속성을 이용하여 데이터 희소 조건에서 실험을 진행한 결과 전반적인 성능 개선으로 이어짐을 확인하였다. (Zhang et al., 2018)에서는 사용자와 아이템의 특징을 추출하기 위해 태그 및 시간 요소를 기반으로 개선된 하이브리드 협업 필터링을 제안하였으며, (Najafabadi et al., 2017)에서는 클러스터링과 연관 규칙을 사용하여 희소성을 해결하였다. (Lee et al., 2018)에서는 VAE(Variational Autoencoder)과 같은 딥러닝 기법을 사용하여 사용자와 아이템 특징을 추출하였으며, MLP(Multi-layer Perceptron)을 기반으로 이를 학습하여 선호도를 유추하였다.

2.2. GAN 기반 생성 모델

본 연구에서는 CGAN을 통해 실제 데이터의 불균형을 해결하여 협업 필터링의 장점을 극대화하고 추천 시스템의 정확도 향상에 중점을 둔 연구를 진행한다. GAN은 최근 데이터의 분포를 고려하여 유사한 샘플을 생성해 내는데 효과적으로 알려져 있으며 기존 오버샘플링 방법의 일반화 문제를 해결하고 우수한 성능을 보이는 것으로 주목받고 있다(Mirza et al., 2014). 따라서 기존의 이미지, 음성 등 비정형 데이터뿐만 아니라 다양한 분야에서 GAN을 활용하여 불균형 문제를 해결한 연구들이 진행되고 있다.

(Mirza et al., 2014)는 GAN 모델의 입력이 무작위 노이즈이기 때문에 출력 결과를 제어할 수 없다는 문제점을 해결하기 위하여 CGAN을 연구하였다. CGAN을 이용하여 사용자가 임의로 원하는 데이터를 생성할 수 있다. GAN 기반 모델은 학습이 어렵다는 단점이 있지만 한번 학습이 완료되면 뛰어난 성능을 보이기 때문에 많은 관련 연구들이 진행되고 있다. (Douzas et al., 2017)에서는 CGAN을 이용하여 여러 정형 데이터를 오버샘플링 하였다. CGAN을 이용하여 실제 데이터 분포를 학습시키고 다양한 불균형 데이터 세트의 소수 클래스에 대한 데이터를 생성하였으며, 생성된 데이터의 품질이 기존 오버샘플링 기법들에 비해 크게 향상되었음을 확인하였다. (Fiore et al., 2017)에서는 FDM 개발에서 발생하는 불균형 데이터 문제를 완화하기 위하여 GAN을 사용하여 소수 클래스를 오버샘플링 하였다. GAN을 통해 생성된 데이터셋이 기존의 원본 데이터보다 더 높은 분류 예측도를 가진다는 것을 검증하였다. (Son et al., 2019)와 (Vega-Márquez et al., 2019)에서는 CGAN을 활용하여 불균형한 카드 사기 데이터, 이메일 마케팅 데이터, 텔레마케팅 데이터의 소수 클래스에 대한 오버샘플링을 진행하였다. 분류기 별로 모든 데이터가 샘플링 적용 후 성능이 향상됨을 확인하였다. (Kim et al., 2019)에서는 구매 이력이나 평점이 부족한 고객(cold-start users) 문제를 해결하기 위하여 CGAN을 이용하여 cold-start 고객들과 유사한 가상의 고객들을 생성하는 방안을 제안하였다. 생성된 가상의 고객은 UMAP(Uniform Manifold Approximation and Projection)을 이용한 시각화 기법을 통해 적절하게 생성된 것을 확인하였다. (Seo et al., 2017)

은 불균형 빅데이터를 이용한 분류 모델 학습에서 분류 성능이 저하되는 문제를 해결하기 위하여 GAN을 사용하였다. GAN 기반 오버샘플링 기법을 적용한 방법론이 기존의 오버샘플링 기법을 적용한 방법론에 비해 소수 클래스에 대한 분류를 보다 효율적으로 수행할 수 있음을 보였다.

불균형 데이터에 대한 우수한 성능으로 다양한 분야에서 CGAN을 활용한 연구들이 진행되고 있다. 그러나 CGAN을 통해 희소성과 데이터 편향성을 해결한 추천 시스템의 성과를 향상에 대한 연구는 미미하다. 따라서 본 연구에서는 CGAN을 활용하여 협업 필터링의 문제점을 개선시키고 추천 시스템의 정확도를 높이는 것에 중점을 두고자 한다.

3. 딥러닝 기반의 생성모델

3.1. GAN

GAN은 생성기(G, Generator)모델과 판별기(D, Discriminator)모델이 서로 적대적으로 경쟁하며 학습하여 실제에 가까운 샘플을 생성하는 비지도 학습 기반의 학습 모델이다(Goodfellow et al., 2014). 생성기는 원본 데이터와 유사한 가상의 데이터를 생성하는 모델이며 판별기는 생성된 데이터와 원본 데이터를 판별하는 모델이다. 생성기의 목적은 판별기를 혼동시키는 것으로 충분한 학습 후에는 판별기가 구별하기 어려울 정도로 원본과 유사한 샘플을 생성하게 된다. 임의의 노이즈 z를 생성기에 입력값으로 넣으면 가상의 데이터가 생성되고, 이를 판별기에 넣으면 해당 데이터가 실제 데이터일 확률이 출력된다. 출력된 확률값을 통해 생성기와 판별기가 서로 대립적으로 경쟁하며 학습을 진행한다. GAN의 목적 함수는 다음 식과 같다.

$$\min_G \max_D V(D,G) = E_D + E_G$$

Where :

$$E_D = \mathrm{E}_{x\sim Pdata}(x)[\log D(x)]$$

$$E_G = E_{z\sim Pz}(z)[\log(1 - D(G(z)))]$$

P_{data}는 원본 데이터의 분포이며 x는 원본 데이터 P_{data}에서 추출한 표본을 나타낸다. P_z는 노이즈의 분포를 나타내며, z는 P_z에서 추출한 노이즈 표본을 나타낸다. G는 생성기를 나타내며 $G(z)$는 생성기가 임의의 노이즈 z를 입력으로 받아 생성한 가상의 데이터를 의미한다. D는 판별기 모델로 $D(x)$는 판별기가 생성된 샘플을 실제 데이터와 비교하여 정확한 구분을 할 확률이다. 입력 데이터가 실제 데이터라고 판단할 경우 1을 출력하고, 생성된 데이터라고 판단할 경우 0을 출력한다. 판별기의 입장에서 $D(x)$와 $G(z)$를 잘 구별해내는 것이 목적이므로 $D(x)$는 1, $D(G(z))$는 0이 되어야하며 $\log(1 - D(G(z)))$의 값은 0이 되도록 학습을 진행하는 것이 이상적이다. 반면 생성기의 입장에서는 판별기가 $G(z)$를 실제 데이터라고 판단하도록 만드는 것이 목적이므로 $D(G(z))$가 1로, $\log(1 - D(G(z)))$은 무한대가 되도록 최대화해야 한다. 결과적으로 D는 $V(D, G)$를 최대화하려고 하는 반면 G는 $V(D, G)$를 최소화하기 위해 상반되는 목적을 가지고 서로 경쟁하며 실제 데이터와 샘플을 구별할 수 없을 정도($D(x) = 0.5$)에 도달하게 된다.

3.2. CGAN

본 논문에서 사용되는 CGAN 모델은 GAN 모델에서 파생된 모델로 학습 과정은 GAN과 동일하다. GAN과의 차이점은 원본 데이터의 분포와 동시에 사용자가 원하는 특징을 반영한 학습이 가능하다는 것이다. 따라서 추가 정보를 통해 데이터 생성 과정을 제어하여 효과적으로 특정 조건의 데이터를 생성할 수 있다(Mirza et al., 2014).

먼저 가우시안 분포(Gaussian distribution)로 생성된 무작위 소음 z와 소수 클래스의 특징 y가 생성기에 입력되어 소수 클래스의 샘플 데이터를 생성한다. 해당 샘플은 판별기에 입력되며, 판별기는 원본 데이터 P_{data}에서 추출한 표본 x와 사용자가 지정한 Condition vector y를 반영하여 입력된 데이터가 실제 데이터인지 가짜 데이터인지 구별한다. 이를 목적함수로 나타내면 다음 식과 같다.

$$\min_G \max_D V(D, G) = E_D + E_G$$

Where :

$$E_D = E_{x \sim Pdata}(x)[\log D(x|y)]$$

$$E_G = E_{z \sim Pz}(z)\,[\log(1 - D(G(z|y)))]$$

$D(x|y)$을 보면 판별기에 원본 데이터 P_{data}에서 추출한 표본 x 뿐만 아니라 Condition vector y의 조건이 추가되었다. CGAN의 생성기는 실제 데이터로부터 샘플을 생성할 때, y의 성질을 가지도록 생성한다. 판별기 또한 입력된 데이터가 실제 데이터인지 가상의 데이터인지 구별하는 과정에서 y를 고려하여 구분하게 된다. 이러한 특징으로 본 논문에서는 소수 클래스의 특징을 추가하여 데이터를 생성하고자 한다.

4. Sampling method

4.1. Synthetic Minority Oversampling TEchnique:SMOTE

SMOTE는 소수 클래스의 데이터를 무작위로 선택해 반복적으로 추가하여 데이터 수를 증가시키는 ROS의 과적합 문제를 극복하면서 소수 클래스 샘플의 데이터를 확장하기 위해 제안된 방법이다. SMOTE는 데이터를 그대로 복제하는 것이 아니라 기존 데이터를 적절히 조합하여 새로운 샘플을 만드는 방법을 사용한다. 소수 범주의 데이터를 대상으로 k-Nearest Neighbor 알고리즘에 기반하여 k개를 합성해 새로운 데이터를 생성한다(Chawla et al., 2002). SMOTE는 소수 범주의 데이터를 복제하여 사용하는 것이 아닌 새로운 데이터를 생성하기 때문에 과적합의 문제가 감소한다. 그러나 합성된 데이터가 오히려 노이즈로 작용하거나 소수 클래스의 경계가 다수 클래스 영역에 가깝게 확산되어, 새로운 다수 클래스에 대한 예측에 있어서 성능 저하가 나타날 수 있다.

4.2. Borderline SMOTE

Borderline SMOTE는 앞서 설명한 SMOTE 알고리즘을 변형한 방법이다. Borderline SMOTE는 두 집단을 구분하는 경계선상의 샘플들에 대해서만 정확히 분류하면 이외의 자료 또한 분류가 잘 될 것이라는 가정을 기반으로 한다(Han et al., 2005). 기존의 SMOTE는 소수 클래스의 데이터 전체를 대상으로 샘플링을 진행하는 반면, Borderline SMOTE는 소수 클래스 중에서도 특히 다수 클래스와의 경계에 위치한 샘플들에 대해서만 새로운 샘플을 생성한다.

〈Table 1〉 MovieLens 100K data

userID	movieID	Ratings
1	1339	3.5
287	25014	5.0
213	59369	4.0
547	8528	3.5
644	785	3.0
…	…	…

4.3. SVM-SMOTE

SVM-SMOTE는 Borderline SMOTE와 같이 클래스 간의 경계선 근처에서 새로운 소수 클래스 샘플을 생성하는데 중점을 둔다. 그러나 SVM-SMOTE의 경계선 영역은 Support Vector Machine(SVM) 분류기를 사용하여 얻은 support vector에 의해 계산된다는 차이점이 있다(Nguyen et al., 2009). (Ghorbani et al., 2020)에 의하면 불균형 데이터 문제에 있어서 SVM-SMOTE는 다른 오버샘플링 기법들보다 효율적이며 가장 좋은 성능을 나타낸다.

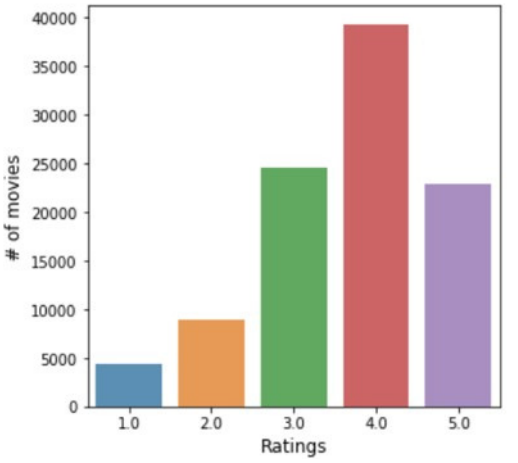
4.4. ADAptive SYNthetic Sampling: ADASYN

ADASYN은 밀도 분포를 고려하여 특정 샘플에 대해 생성할 합성 샘플의 수를 결정함으로써 SMOTE 기법에 비해 보다 체계적으로 데이터를 생성시키는 방법이다(He et al., 2008). SMOTE에 비해 상대적으로 균형 잡힌 데이터가 생성되는 것으로 보이며, 학습하기 어려운 샘플에 집중하여 학습을 진행할 수 있기 때문에 성능을 향상시킨다. 그러나 소수 클래스의 다양성을 향상시키는 관점에서는 큰 성능을 보이지 못한다.

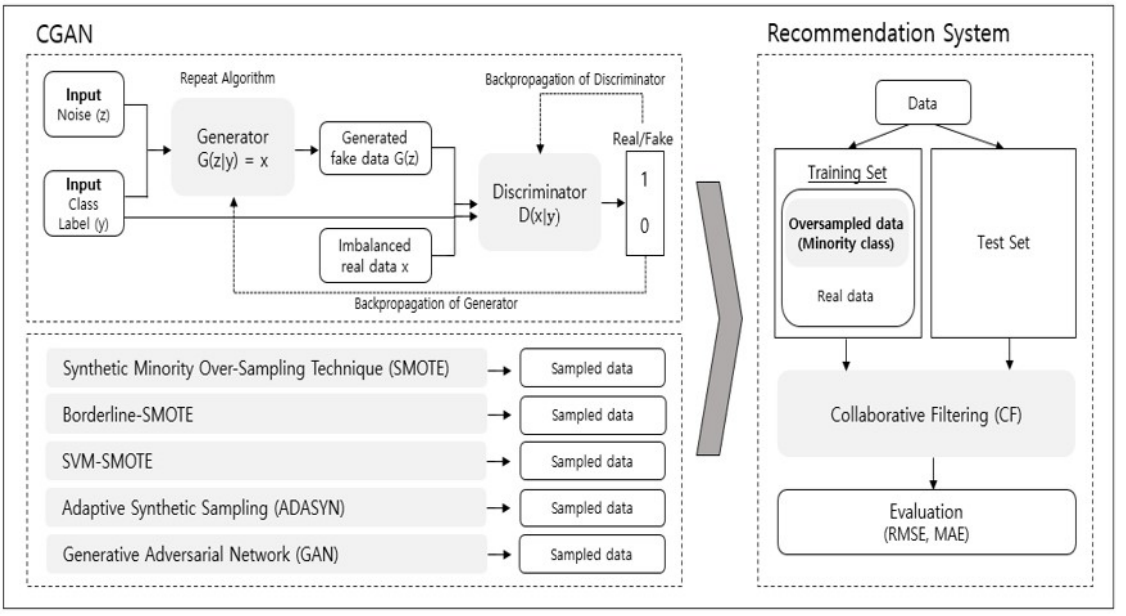
5. 연구 설계

5.1. 데이터셋

본 연구에 활용된 데이터는 Minnesota 대학의 GroupLens 프로젝트에 의해 개발된 Movielens 데이터이다. 해당 데이터는 943명의 userID, 1682편의 movieId, 100,000건의 ratings, timestamp으로 구성되어 있으며 timestamp를 제외한 나머지를 사용한다. 선호도 등급은 0부터 5까지의 척도로 이루어져 있으며, 0.5씩 증가한다. <Table 1>는 연구에서 활용된 데이터의 일부를 예시로 나타낸 것이다.



〈Figure 1〉 MovieLens distribution of rating



〈Figure 2〉 Basic structure

데이터의 평점별 분포를 1을 간격으로 변환하여 막대그래프로 나타내면 <Figure 1>와 같다. 평점의 분포를 보면, 4점이 34.2%로 가장 많으며 가장 적은 1점의 분포인 6.1%와 5배 이상 차이가 나타난다는 것을 알 수 있다. 또한 3-5점의 긍정적인 평점이 전체의 80% 이상으로 데이터 자체에 심한 불균형을 가지고 있다는 특징이 있다.

5.2. 분석 절차

제안하는 모형의 성능을 테스트하기 위해서 선행연구의 검토를 통해 비교 기법을 선택하였다. 전체적인 분석 과정과 비교 모형은 <Figure 2>와 같다.

본 연구에서는 CGAN을 기반으로 협업필터링 구현 시 발생하는 희소성 문제를 해결함과 동시에 실제 데이터에서 발생하는 데이터 불균형을 완화하여 모형의 정확도 향상에 얼마나 효과적인가를 확인하고자 한다. 다양한 기법들과의 비교를 위해 희소성이 존재하는 기존 사용자-아이템 데이터에 협업 필터링을 적용한 모델과 오버샘플링 기법인 SMOTE, Borderline SMOTE, SVM-SMOTE, ADASYN 및 condition vector을 추가하지 않은 GAN을 사용하여 데이터셋을 설정하고 협업 필터링을 적용한다. 최종적으로 각 방법론 별로 협업 필터링 추천 성능을 비교하였으며, 협업 필터링 모델로는 KNN, SVD를 사용하였다. 모든 실험에 대한 성능 평가는 오버샘플링 자체에 대한 분류 정확도가 아닌 협업 필터링 모델의 성과로 측정하여 추천 시스템에 대한 성능 향상을 목적으로 하였다. 본 연구는 다음 절차에 따라 진행된다.

우선, 효과적인 CGAN 학습을 위해 userId와 movieId에 대하여 각각 사용자의 특성과 아이템

의 특성으로 이루어진 잠재요인 벡터를 도출한다. 도출된 user-vector와 item-vector을 CGAN 알고리즘에 적용하여 학습시킨다. 데이터의 특징이 학습된 CGAN 알고리즘에 random noise vector z와 소수 클래스라는 클래스 정보를 conditional vector y에 입력하여, 원본 데이터와 유사한 가상의 데이터를 생성한다. 이 과정을 반복하여 소수 클래스의 데이터 수가 다수 클래스의 데이터 수와 같아지도록 새로운 데이터 집합을 만든다.

또한, 최적의 추천 적중률을 찾기 위하여 CGAN 모델의 하이퍼 파라미터 튜닝을 진행하고자 하며, 구체적인 하이퍼 파라미터 array list는 다음 <Table 2>과 같이 정의하였다.

〈Table 2〉 Hyperparameter array list

Parameter	Values
Epochs	120
Batchs	256
Activation functions	G : LeakyRELU, Tanh D : LeakyRELU, Sigmoid
Optimization algorithm	Adam
Number of layers	G : 4 D : 4
Loss functions	Binary cross entropy
Learning rate	0.0002

이후 생성된 가상의 user-vector와 item-vector에 대하여 기존 user-vector와 item-vector와 각각의 코사인 유사도를 산출한 뒤 가장 유사한 사용자와 아이템으로 대체하여 희소한 사용자-아이템 매트릭스를 보완한다. 또한 이 과정에서 소수 범주의 샘플수가 다수 범주의 샘플 수와 동일하게 되도록 새로운 데이터를 기존의 데이터셋에 추가한다. 이렇게 만들어진 새로운 데이터셋에 협업 필터링 기법을 적용하여 사용자에게 아이템을 추천한다. 협업 필터링에서 사용자 간 유사도 계산 시, 피어슨 상관계수 (pearson correlation coefficient)가 코사인 유사도(cosine similarity)보다 좋은 성과를 나타내는 것으로 알려져 있다 (Herlocker et al., 1999). 따라서 본 실험에서 협업 필터링 적용 시 사용자 간 유사도를 측정에는 피어슨 상관계수를 이용한다.

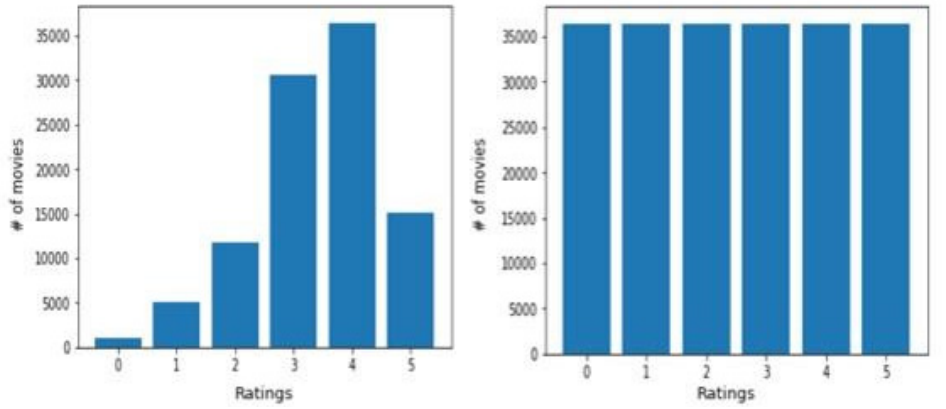
마지막으로 제안 모형의 성능을 보다 정확하게 검증하기 위하여 기존 데이터셋을 기반으로 한 전통적인 협업 필터링 모델뿐만 아니라, 보편적으로 많이 사용되고 있는 오버샘플링 기법인 SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN과 GAN 알고리즘을 기반으로 데이터를 생성한 뒤 협업 필터링에 적용한다.

본 연구에서는 추천 시스템에 대한 예측력 향상을 확인하기 위하여 오버샘플링 자체에 대한 분류 정확도가 아닌 협업 필터링 모델의 성과를 기반으로 성능을 평가한다. 또한 5-fold cross validation에 의하여 협업 필터링의 예측을 실시하고, RMSE, MAE 평가 방법을 사용하여 평가를 진행한다.

5.3. 평가 척도

추천 시스템의 주요 목적은 사용자의 미래 관심사과 선호도를 예측하는 것이다. 이에 따라 본 연구에서는 RMSE(Root Mean Squared Error)와 MAE(Mean Absolute Error)를 비교하여 추천 성능을 평가했다.

RMSE와 MAE의 경우 수치와된 평점에 대한 평가를 하기에 용이하며, 결과값의 크기가 작을수록 높은 예측 정확도를 나타낸다. 예측 결과가 맞고 틀리고의 정확도가 중요한 일반적인 비감



〈Figure 3〉 Example of datasets before and after oversampling

〈Table 3〉 Comparison of RMSE, MAE

Evaluation	RMSE		MAE	
CF	KNN	SVD	KNN	SVD
Base	0.9670	0.8969	0.7434	0.6904
SMOTE	1.4544	1.3290	1.0999	1.0640
B-SMOTE	1.2202	1.1540	0.8693	0.8825
SVM-SMOTE	1.2145	1.1460	0.8987	0.8743
ADASYN	1.4539	1.3361	1.1002	1.0700
GAN	0.8601	0.8336	0.6125	0.6387
CGAN	0.7777	0.7486	0.5513	0.5592

독, 강화 학습과 달리, 평점의 예측은 근접도가 중요하기 때문에 RMSE와 MAE를 사용하여 실제 등급에 대한 예측 등급의 근접도를 측정하고자 한다. Ep가 검증 데이터, 사용자 i의 아이템 a에 대한 실제 평점은 r_{ia}, 예측 평점은 $\tilde{r}_{ia}$라고 하면, 수식은 다음과 같다(Lü et al., 2012).

$$\text{MAE} = \frac{1}{|\text{Ep}|}\sum_{(i,a)\epsilon\text{Ep}}|r_{ia} - \tilde{r}_{ia}|,$$

$$\text{RMSE} = \left(\frac{1}{|\text{Ep}|}\sum_{(i,a)\epsilon\text{Ep}}(r_{ia} - \tilde{r}_{ia})^2\right)^{1/2}$$

5.4. 분석 결과

<Figure 3>은 훈련용 데이터에 각 오버샘플링 기법들을 적용한 후 최종적으로 구성된 데이터를 예시적으로 나타내고 있다. 그림을 살펴 보면 생성된 데이터가 추가된 새로운 데이터셋에서는 각 클래스가 균형을 이룬 것을 확인할 수 있다.

실험 데이터셋에 대하여 KNN, SVD 알고리즘 기반 협업 필터링을 적용한 후 성능을 평가한 결과는 <Table 3>과 같다. CGAN 알고리즘을 기반으로 생성된 가상의 데이터에 협업 필터링 기법

을 적용하여 추천 성능을 실험한 결과 CGAN + KNN CF의 경우 RMSE는 0.7777, MAE는 0.5513으로, CGAN + SVD CF의 경우 RMSE는 0.7486, MAE는 0.5592로 가장 낮게 계산되었다. 제안모형의 RMSE는 기존 협업 필터링 모델보다 0.1893, GAN기반 협업 필터링 모델보다 0.0824, SVM-SMOTE 기반 협업 필터링 모델보다 0.4368 낮은 값을 보였다.

결과적으로 KNN기반 협업 필터링의 RMSE를 기준으로 보았을 때, 본 연구에서 제안한 모형은 가장 정확도가 높았던 SVM-SMOTE 오버샘플링 기법보다 RMSE가 약 35%, GAN 모델에 비하여 약 10%, 원본 데이터에 비하여 약 20% 성능이 개선됨을 확인하였다.

기존의 전통적 오버샘플링 기법들로 생성한 데이터의 경우 오히려 원본 데이터보다 추천의 정확도가 낮아지는 모습을 보인다. 기존 오버샘플링 기법은 고차원 데이터에 효율적이지 않으며, 클래스가 겹치거나 노이즈를 생성하여 데이터의 편향을 줄임과 동시에 추천의 정확도를 높이는 것은 어렵다. 이진 클래스는 클래스들 사이에 잘 정의된 관계가 존재한다. 반면 다중 클래스의 경우 클래스 간의 관계가 간단하지 않기 때문에 다중 클래스에 대한 기존의 오버샘플링은 덜 효과적이거나 심지어 부정적인 영향을 미치는 경우가 있다(Zhou et al., 2006). 실험을 통해, 본 연구에서 사용된 다중 클래스 불균형 데이터의 경우 이진 분류에 사용되는 기법을 직접 적용할 수 없으므로 새로운 샘플링 기법이 필요하다는 것을 확인하였다.

GAN을 기반으로 생성된 모델의 경우에는 기존에 비하여 추천의 성능은 다소 개선되었지만 원본 데이터의 편향성을 해결하지 못하고 다수의 편향된 샘플에 과도하게 학습된다. 결과적으로 대부분 다수의 범주에 해당하는 샘플이 생성된 것을 확인할 수 있었다. 이러한 결과는 실제 사용자들의 선호와는 거리가 있어 추천의 유용성을 저해한다.

이에 반해 본 연구에서 제안한 CGAN + CF 기법은 모든 기법들 보다 향상된 성능을 보임을 확인하였다. CGAN이 그럴듯한 사용자와 아이템의 평점을 생성하여 데이터 희소성과 편향성을 완화하기 때문이다. 따라서 본 연구의 실험 결과, CGAN의 활용으로 협업 필터링 구현 시 발생하는 희소성 문제와 실제 데이터에서 발생하는 데이터 불균형 문제를 해결하고 추천의 정확도를 높일 수 있음을 입증하였다.

6. 결론

본 연구에서는 실제 데이터에서 협업 필터링의 추천 성능을 저해하는 문제를 해결하기 위해 CGAN과 협업 필터링을 결합하여 기존 문제점을 개선하고 정확도를 높이는 모형을 제시하고자 하였다. 그 결과, CGAN 기반 협업 필터링이 기존의 다양한 기법들에 비해 향상된 성능을 보임을 증명하였다. 딥러닝을 기반으로 실제 데이터와 유사한 가상의 데이터를 생성하고 이러한 특징을 추천 시스템에서의 다중 클래스 불균형 문제 해결을 위한 전처리 과정에 활용할 경우 다음과 같은 장점이 있다.

첫째, 소수 클래스와 다수 클래스의 비율을 맞추어 각 클래스의 균형을 이룰 수 있다. Condition vector y를 통해 소수 범주에 대한 데이터만을 생성하여, 부족한 소수 범주 데이터만큼의 샘플 추출을 통해 불균형을 해소하는 효과를 얻을 수 있다.

둘째, 다중 클래스 불균형 문제에 대한 기존 오버샘플링 기법들의 한계를 해소할 수 있다. 기존의 랜덤 오버샘플링 기법은 과적합으로 인해 추천 성능이 크게 저하될 수 있으며, 데이터 겹침이나 추가적인 노이즈가 발생할 수 있고 고차원 데이터에서는 효과적이지 않을 수 있다. 또한 데이터에 대한 전반적인 분포를 고려하기 보다 임의의 점과 그 주변의 점들만을 고려하여 비현실적인 데이터를 생성할 가능성이 있다. 반면 CGAN은 어떠한 분포의 데이터도 모방하는 학습이 가능하기 때문에 고차원 데이터에서도 우수한 성능을 보인다.

마지막으로 데이터가 희소한 상황에서 빈 매트릭스를 채워넣어 협업 필터링의 희소성 문제를 해결할 수 있다. CGAN으로 생성된 가상의 사용자-아이템-평점 데이터를 실제 데이터와 가장 유사한 항목으로 각각 대체하여 희소한 매트릭스를 보완할 수 있다.

본 연구를 통해 도출할 수 있는 시사점은 다음과 같다. 첫째, 딥러닝에 기반한 오버샘플링이 실제 데이터를 사용한 추천 시스템의 성능을 더욱 정교하게 하여 비즈니스 추천 시스템 구축 등에 활용될 수 있을 것이다. 또한, 기존 이미지, 음성, 자연어 처리 등의 분야에서만 활발히 연구되던 CGAN의 연구 범위를 확장하여 평점 데이터에 적용하였다는 점에서 의의를 가지며 해당 모형을 다양한 도메인에 활용할 경우 효과적인 추천 시스템을 구축하는데 기여할 수 있을 것으로 기대한다.

그러나 본 연구는 다음과 같은 한계점을 갖는다. 첫째, 본 연구에서 제안한 방법과의 비교 대상으로 Base, SMOTE, B-SMOTE, ADASYN, GAN만을 사용했다는 점이다. One-versus-one(OVO) 또는 One-versus-all(OVA) 기반 SMOTE, AdaBoost 등 다양한 기법들을 추가적으로 활용한다면 다중 클래스 오버샘들링에 대한 더욱 효과적인 성능 비교가 가능할 것이다. 둘째, 협업 필터링의 유사도 계산 과정에서 임계치와 KNN 기반 최적 이웃의 수(k)등 feature의 수를 다양하게 설정하여 적용해보지 못했다는 점에서 한계가 있다. 마지막으로 제안 모형을 다양한 종류의 데이터로 실험을 진행하지 못하였다는 점도 한계점으로 볼 수 있다. 향후 연구에서는 다양한 분야에서 발생되는 실제 데이터와 비교 모델을 활용하여 그에 따른 성능 확인을 진행해야 할 것으로 판단된다. 또한, 추천 시스템에 대한 딥러닝 기반 오버샘플링 전처리의 유용성을 확인하였으므로, 다양한 딥러닝 알고리즘을 활용한 오버샘플링을 통해 보다 높은 추천 모델을 구현할 수 있을 것으로 기대된다.

참고문헌(References)

B. Vega-Márquez, C. Rubio-Escudero, J. C. Riquelme, and I. Nepomuceno-Chamorro, “Creation of synthetic data with conditional generative adversarial networks,” International Workshop on Soft Computing Models in Industrial and Environmental Applications, (2019), 231~240.

Chevalier, J. and Mayzlin, D, “The effect of word of mouth on sales: Online book reviews,” J. of Marketing Research, 43(2006), 3.

D. Billsus and M. . Pazzani, “Learning collaborative information filters,” *Proceedings of the Fifteenth International Conference on Machine Learning*, Vol. 54(1998), 48.

Douzas, G., & Bacao, F, “Effective data generation for imbalanced learning using conditional

generative adversarial networks," Expert Systems with Applications, 91(2017), 464~471.

Fiore, U., A. D. Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," Information Sciences, 479(2017), 448~455.

G. Adomavicius, A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, 17(2005), 734~749.

Goldberg, D., Nichols, D., Oki, B. M., and Terry, D, "Using collaborative filterin g to weave an information tapestry," Communications of the ACM, 35(12) (1992), 61-70.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D.,Ozair, S., Courville, A., & Bengio., Y, "Generative Adversarial Nets," Advances in neural information processing systems, (2014), 2672~2680.

H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," Proceedings of the International Conference on Intelligent Computing, Berlin(2005), 878~887.

H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," Proc. IEEE Int. Joint Conf. Neural Netw. IEEE World Congr. Comput. Intell.(2008), 1322~1328.

H. M. Nguyen, E. W. Cooper, K. Kamei, "Borderline over-sampling for imbalanced data classification," International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1)(2009), 4~21.

Kim, Chae, Kim, "Generating and Visualizing Neighbors of Cold-start Users in Recommender System," Journal of the Korea Institute of Information and Communication Engineering, (2019), 134~135.

Krawczyk B, "Learning from imbalanced data: open challenges and future directions," Progr. AI, 5 (4) (2016), 221~232.

Kwon, Hong, "Method to Improve Data Sparsity Problem of Collaborative Filtering Using Latent Attribute Preference," *Journal of Internet Computing and Services*, Vol.14, No.5(2013), 59~67.

Lee Y, Kim SW, Park S, Xie X, "How to impute missing ratings? Claims, solution, and its application to collaborative filtering," In: Proceedings of the 2018 World Wide Web conference on World Wide Web(2018), 783~792.

Lee, Kim, "Analysis of Data Imputation in Recommender Systems," *Journal of KIISE*, Vol. 44, No. 12(2017), 1333~1337.

Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K., & Zhou, T, "Recommender systems," Physics Reports, 519(1) (2012), 1~49.

M. K. Najafabadi, M. N. Mahrin, S. Chuprat, and H. M. Sarkan, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," *Comput. Hum. Behav*, Vol. 67(2017), 113~128.

Michał Koziarski, Michał Woźniak, Bartosz Krawczyk, "Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise," *Knowl.-Based Syst*, Vol 204(2020), 106223.

Mirza, M. and Osindero, S, "Conditional Generative Adversarial Nets," arXiv preprint arXiv:1411.1784, (2014).

N. Hu, J. Zhang, and P. A. Pavlou. “Overcoming the J-shaped Distribution of Product Reviews,” *Communications of the ACM*, Vol. 52, No. 10(2009), 144~147.

N. V. Chawla, et al, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, Vol.16(2002), 321~357.

P. Melville, R. J. Mooney, and R. Nagarajan, “Content-Boosted Collaborative Filtering for Improved Recommendations,” American Association for Artificial Intelligence (www.aaai.org) (2002), 187-192.

R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting Students’ performance using machine learning techniques," *IEEE Access*, Vol. 8(2020), 67899~67911.

R. Jin and J. Zhang, “Multi-class learning by smoothed boosting", *Mach. Learn.*, Vol. 67, No. 3(2007), 207~227.

Schoenmueller, V., Netzer, O., & Stahl, F, “The extreme distribution of online reviews: Prevalence, drivers and implications,” Columbia Business School Research Paper(2018), No. 18~10.

Seo, Jeon, Lee, Jung, Kim, “An Over-sampling Method based on Generative Adversarial Networks for Effective Classification of Imbalanced Big Data,” Journal of the Korea Institute of Information and Communication Engineering(2017), 1030~1032.

Shinhyun Ahn and Chung-Kon Shi, “Exploring Movie Recommendation System Using Cultural Metadata,” *Transactions on Edutainment II, Lecture Notes in Computer Science*, Vol. 5660(2009), 119~134.

Son, Kim, Kim, Cho, “Review and Analysis of Recommender Systems,” Journal of the Korean Institute of Industrial Engineers, 41(2) (2015), 185~208.

SongJie Gong, HongWu Ye, “An item based collaborative filtering using BP Neural Networks prediction,” International Conference on Industrial and Information Systems(2009), 146~148.

Z.-H. Zhou and X.-Y. Liu, “Training cost-sensitive neural networks with methods addressing the class imbalance problem,” *IEEE Trans. Knowl. Data Eng.*, Vol. 18, No. 1(2006), 63~77.

Zhang, C. X., Yang, M., Lv, J., & Wang, W. Q, “An improved hybrid collaborative filtering algorithm based on tags and time factor,” Big Data Mining and Analytics, 1(2018), 48~56.

Abstract

Conditional Generative Adversarial Network based Collaborative Filtering Recommendation System

Soyi Kang* · Kyung-shik Shin**

With the development of information technology, the amount of available information increases daily. However, having access to so much information makes it difficult for users to easily find the information they seek. Users want a visualized system that reduces information retrieval and learning time, saving them from personally reading and judging all available information. As a result, recommendation systems are an increasingly important technologies that are essential to the business.

Collaborative filtering is used in various fields with excellent performance because recommendations are made based on similar user interests and preferences. However, limitations do exist. Sparsity occurs when user-item preference information is insufficient, and is the main limitation of collaborative filtering. The evaluation value of the user item matrix may be distorted by the data depending on the popularity of the product, or there may be new users who have not yet evaluated the value. The lack of historical data to identify consumer preferences is referred to as data sparsity, and various methods have been studied to address these problems. However, most attempts to solve the sparsity problem are not optimal because they can only be applied when additional data such as users' personal information, social networks, or characteristics of items are included.

Another problem is that real-world score data are mostly biased to high scores, resulting in severe imbalances. One cause of this imbalance distribution is the purchasing bias, in which only users with high product ratings purchase products, so those with low ratings are less likely to purchase products and thus do not leave negative product reviews. Due to these characteristics, unlike most users' actual preferences, reviews by users who purchase products are more likely to be positive. Therefore, the actual rating data is over-learned in many classes with high incidence due to its biased characteristics, distorting the market. Applying collaborative filtering to these imbalanced data leads to poor recommendation performance due

* Department of Big Data Analytics, Ewha Womans University
** Corresponding author: Kyung-shik Shin
School of Business, Ewha Womans University
52 Ewhayeodae-gil, Seodaemun-gu, Seoul, 120-750, Korea
Tel: +82-02-3277-2799, Fax: +82-2-3277-2766, E-mail: ksshin@ewha.ac.kr

to excessive learning of biased classes.

Traditional oversampling techniques to address this problem are likely to cause overfitting because they repeat the same data, which acts as noise in learning, reducing recommendation performance. In addition, pre-processing methods for most existing data imbalance problems are designed and used for binary classes. Binary class imbalance techniques are difficult to apply to multi-class problems because they cannot model multi-class problems, such as objects at cross-class boundaries or objects overlapping multiple classes.

To solve this problem, research has been conducted to convert and apply multi-class problems to binary class problems. However, simplification of multi-class problems can cause potential classification errors when combined with the results of classifiers learned from other sub-problems, resulting in loss of important information about relationships beyond the selected items. Therefore, it is necessary to develop more effective methods to address multi-class imbalance problems.

We propose a collaborative filtering model using CGAN to generate realistic virtual data to populate the empty user-item matrix. Conditional vector y identify distributions for minority classes and generate data reflecting their characteristics. Collaborative filtering then maximizes the performance of the recommendation system via hyperparameter tuning. This process should improve the accuracy of the model by addressing the sparsity problem of collaborative filtering implementations while mitigating data imbalances arising from real data. Our model has superior recommendation performance over existing oversampling techniques and existing real-world data with data sparsity. SMOTE, Borderline SMOTE, SVM-SMOTE, ADASYN, and GAN were used as comparative models and we demonstrate the highest prediction accuracy on the RMSE and MAE evaluation scales. Through this study, oversampling based on deep learning will be able to further refine the performance of recommendation systems using actual data and be used to build business recommendation systems.

Key Words : Recommendation System, Collaborative filtering, Oversampling, Deep learning, GAN

Received : June 1, 2021 Revised : July 5, 2021 Accepted : July 14, 2021
Corresponding Author : Kyung-shik Shin

저 자 소 개

강 소 이
이화여자대학교에서 빅데이터분석학 석사학위를 취득하였다. 주요 관심분야는 추천 기술, 딥러닝, 빅데이터 분석, 비즈니스 애널리틱스(Business Analytics) 등이다.

신 경 식
현재 이화여자대학교 경영대학 경영학부 교수로 재직 중이다. 연세대학교 경영학과를 졸업하고, 미국 George Washington University에서 MBA, KAIST에서 경영공학 Ph.D.를 취득하였다. 주요 연구분야는 데이터 마이닝과 비즈니스 인텔리전스, 빅데이터 분석, 비즈니스 애널리틱스(Business Analytics), 인공지능 응용과 지식공학 등이다.