신용 예측 데이터의 불균형 문제 해결을 위한 가상 데이터 생성 기법

오상민, 이주홍*
인하대학교

5hsangmin@gmail.com, juhong@inha.ac.kr*

Virtual Data Generation Techniques for Imbalance problem of credit prediction data

Sang Min Oh, Ju Hong Lee*
Inha Univ.*

요 약

본 논문은 신용 예측 데이터의 불균형 문제를 GAN 을 사용한 Over sampling 을 적용하여 해결하고자 한다. 신용 예측 데이터는 정상과 연체가 매우 불균형하게 이루어져 있어, 이것을 해결하는 기법이 필수적이다. 기존 논문에서는 불균형 문제를 kNN 기반의 Oversampling 을 사용하여 해결하였으나 신용 예측 데이터에는 적합하지 않다. 신용 예측 데이터는 정상, 연체 데이터가 매우 혼재 되어 있고, 많은 범주형 자료들을 가지고 있기 때문이다. 본 논문에서는 범주형 자료의 Probability Space 로의 전이를 통해 이 문제를 해결하고, 나아가 GAN 으로 데이터를 생성하여 신용 예측을 진행하였다. 진행 결과 기존의 kNN 기반의 Oversampling 방법과 비교하여 더 좋은 성능을 보였다.

Ⅰ. 서 론

신용 예측이란 개인이나 기관이 대출을 받을 때 정해진 기간 안에 갚을 능력이 있는지를 과거의 신용거래 경험이나 현재의 신용거래 상태를 기반으로 예측하는 것으로 대출신청 시 금융 기관의 결정을 보조해주는 역할을 하고 있다.

신용 예측은 전통적으로 대출자들의 통계를 기반으로 하여, 신용과 관련된 모든 사항을 항목별로 점수화해 이 점수에 따라 대출 가능여부와 대출금액을 산정한다. 위와 같은 점수는 사람의 주관적인 판단에 의해 결정되어 일정기간동안 그 결과가 바뀌지 않는다는 단점이 있다. 특히 과거의 거래 경험이 적거나 없는 씬파일러들의 경우 해당 항목의 점수가 낮아 대출을 해줄 수 없는 문제를 가지고 있다.

전통적인 방법의 이러한 문제점을 최근에는 인공지능을 활용하여 해결하려는 연구가 많이 진행되고 있다. 신용 예측에서 이런 인공지능 기법들을 적용하는데 가장 큰 문제는 데이터 불균형을 해결하는 것이다. 데이터가 불균형 할 때 신용 예측 모델은 다수 범주에 해당하는 정상으로 분류하는 경향으로 학습한다. 이러한 문제를 기존 논문에서는 kNN 기반의 oversampling 인 SMOTE, ADASYN 등을 사용하여 해결한다. 하지만 신용 예측 데이터는 결정경계면에서 다수의 데이터가 혼재되어 있어, kNN 기반의 oversampling 방식을 적용 시 이 경계면이 더욱 불분명해진다. 특히 범주형 자료의 경우 순서가 없고, sparse 하기 때문에 데이터 사이의 거리 측정이 어려워 실제 데이터 표현의 어려움을 가지고 있다.

본 논문에서는 이러한 문제를 해결하기 위해 Probabilistic distance 를 적용한 GAN 을 통한 oversampling 방법을 제안한다.

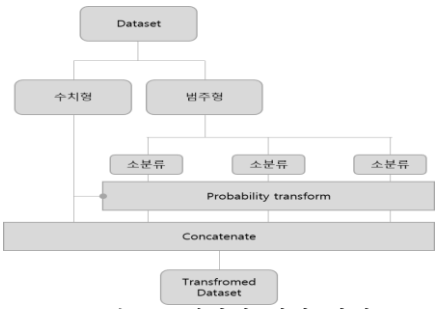
Ⅱ. 본론

범주형 자료들은 대부분 순서가 없고 Sparse 하기 때문에 데이터 생성 연구에서 많은 어려움을 겪고 있다. 그래서 이 범주형 자료들을 연속적인 수로 변환하기 위해 Probability space 를 적용하였다. 범주형 자료들은 이 Probability space 전이를 통해 순서를 가지게 되고, 연속적인 수로 변환된다. 같은 특성을 갖는 데이터로 소분류하는 것은 데이터를 더욱 다양하게 한다. 데이터는 다음과 같은 식을 통해 확률 공간으로 전이 되며, 기존의 다른 데이터를 통해 거리를 학습하게 된다.

$$Dis(x, \pi_k) = \sum [p(X = \pi_k | x_k) - p(X = x_k | \pi_k)^2$$

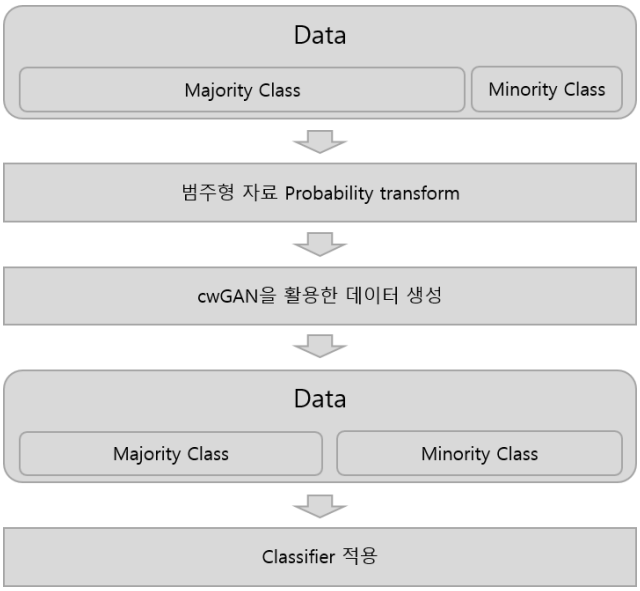
<식 1> Probability distance

π 는 정상 / 연체 결과를 의미하며, x 는 범주형 자료의 값을 뜻하고, X 는 한 사람 전체의 데이터를 의미한다. 확률 거리는 범주형 자료가 각각의 값이 정상 / 연체에 얼마나 영향을 주는 가를 확률로써 표현되며, 범주형 자료 사이의 거리가 커지도록 학습한다.



<그림 1> 데이터 변환 과정

범주형 자료의 확률공간 전이가 끝나면 GAN 을 통하여 Oversampling 을 진행한다. 본 논문에서는 minority class 를 선택적으로 생성하기 위해 conditional GAN 과 학습의 안정성을 위해 wGAN 을 사용하여 데이터를 생성하였다.



<그림 2> 신용 예측 전체 구조도

실험 실험은 A 저축은행 2017 년 3 월부터 2018 년 8 월까지의 데이터를 사용하였다. 데이터는 총 2,569 명의 대출자 정보이며, 거주지, 직업, 접수경로, 나이 등 범주형 자료 30 가지, 대출 금액, 금리, 신용카드 사용 내역등 연속형 자료 150 가지로 이루어진다. 실험 시에는 dataset 에 포함된 정상 데이터와 연체 데이터의 비율을 유지하였으며, test set 은 최종 3 개월간의 데이터를 사용하여 실험을 진행하였다. 평가지표로는 Positive recall, Negative recall, AUC 를 기준으로 평가 하였으며, 기존에 신용예측 분야에 사용된 Sampling 방법인 ADASYN, SMOTE, ENN, ROS, RUS 를 각 분류기 별로 비교 하였다.

Sample 수	정상	연체	Feature수	
			범주형	연속형
2,569	2,418 (94%)	151 (5.8%)	30	150

<표 1> A 저축은행 대출 데이터 특성

실험결과

4 개의 Classifier 를 통해 4 개의 샘플링 방식 총 16 가지를 비교 하였다. 실험 데이터는 총 2,569 개로 데이터가 많이 부족한 환경에서 진행되었으며, 그 결과는 아래 표와 같다. GAN 을 사용한 oversampling 방식이 가장 높은 AUC 를 보이고 있고, 정상 재현율의 경우 다른 sampling 에 비해 조금 낮지만 신용 예측의 특성상 대부분의 손해는 연체를 예측하지 못해서 나오기 때문에 가장 좋은 결과로 보인다.

분류기	평가지표	GAN	ADASYN	SMOTE	ENN	RUS
RF	AUC	**57.24**	**52.78**	**56.6**	**51.35**	**54.46**
	Negative Recall	**39.9**	26.25	38.06	37.27	49.15
	Positive Recall	**74.58**	79.3	75.13	65.43	59.78
XGB	AUC	**58.31**	53.88	53.75	48.38	48.38
	Negative Recall	**55.01**	22.23	21.67	16.52	16.52
	Positive Recall	**61.6**	85.52	85.84	80.21	80.21
DNN	AUC	**55.48**	52.53	55.27	50.87	54.93
	Negative Recall	**60.95**	68.39	60.54	15.15	66.53
	Positive Recall	**50**	36.67	50	86.6	43.33
Average	AUC	**57.01**	53.06	55.21	50.20	52.59
	Negative Recall	**51.95**	38.96	40.09	22.98	44.07
	Positive Recall	**62.06**	67.16	70.32	77.41	61.11

<표 2> 실험 결과

Ⅲ. 결론

본 논문에서는 정상 데이터에 비해 연체된 데이터가 현저히 적은 신용 예측 데이터의 데이터 불균형을 개선하기 위해 GAN 을 통해 데이터 Oversampling 모델을 제시하였다. Euclidean 거리로는 해결할 수 없는 범주형 데이터 생성을 Probability space 로의 전이를 통해 해결하였고, 그 이후 GAN 을 통해 데이터를 생성하였다. 기존의 sampling 방법 3 개와 4 개 Classifier 를 3 가지 평가지표를 사용하여 비교하였고 그 결과 GAN 을 사용한 oversampling 이 다른 샘플링 방식에 비해 유리하다는 것을 보였다.

참 고 문 헌

[1] Ian J. Goodfellow." Generative Adversarial Nets", 2014

[2] Martin Arjovsky, "Wasserstein Gan," 2017,

[3] Lei Xu. "Modeling Tabular Data using Conditional GAN, 2019

[4] Lifei Chen. "Soft subspace clustering of categorical data with probabilistic distance" 2015

[5] Xolani Dastile. "Statistical and machine learning models in credit scoring: A systematic literature survey" 2020