Deep metric Learning을 사용한 신용 예측 모델

오영민, 이주홍*
인하대학교
oh9008@gmail.com, juhong@inha.ac.kr*

Credit Forecasting Model Using Deep Metric Learning

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

요 약

본 논문은 개인이 신용카드 등으로 대출 행위를 했을 때 정해진 기간 내에 정상적으로 대출을 갚을 수 있을지 혹은 갚지 못하고 연체될 것인지를 예측하는 신용 예측 모형을 개발한다. 신용 데이터는 정상 데이터에 비해서 연체 데이터가 현저히 적은 data imbalance 문제가 있어서 일반 기계학습 모델로는 학습이 잘 되지 않는다. 이 문제를 개선하기 위해서 data imbalance 문제에 강점을 가진 deep metric Learning을 사용하여 신용 예측 모델을 개발한다. deep metric learning 의 여러 기법 중에서 triplet network를 사용하여 개발한 신용 예측 모델이 기존의 신용 예측 모델들보다 향상된 성능을 보였다.

Ⅰ. 서론

신용 예측이란 개인이나 기관이 대출받거나 신용카드를 사용하는 등의 행동 시에 정해진 기간에 갚을 능력이 있는지 과거 연체 내역, 현재 수익, 보유 자산 등의 데이터를 기반으로 예측하는 것으로 금융 기관의 결정을 도와주는 역할을 한다. 이러한 신용 예측은 정상적으로 정해진 기간에 갚을지(정상) 혹은 갚지 못하고 연체될 것인지(연체) 예측하는 binary classification task로써 이를 위해 현재까지 통계적 방법을 사용한 모델과 machine learning 기반의 모델들이 많이 사용되었다[1]. 하지만 신용 예측 데이터는 대체로 정상적으로 정해진 기간 안에 갚은 class에 속하는 데이터의 비율이 갚지 못하고 연체된 class에 속하는 데이터의 비율보다 많은 문제를 가진다. 이를 위해 Random Forest나 xgboost 등의 ensemble 기반의 모델들이 적용되었고 다른 신용 예측 모델과 비교하였을 때 좋은 성능을 나타내었다[2]. 그러나 Random Forest나 xgboost 등의 모델들은 성능 향상을 위해 imbalance 문제를 처리하기 위한 기법이 적용되어야 한다. 따라서 본 논문에서는 기존에 사용된 방법과 다르게 별도의 imbalance 처리가 불필요한 deep metric learning 기반의 모델을 사용하였다. deep metric learning이란 주로 face verification이나 face identification 등의 computer vision 분야에서 사용되는 방법으로 데이터의 label과 같은 주어진 정보와 distance를 사용하는 방법이다. 이러한 방법은 embedding 시킨 데이터 및 데이터의 class 간에 distance를 사용하여 데이터의 similarity를 판단하기 때문에 각 class에 속한 데이터의 비율이 불균형한 imbalance 문제에 영향을 적게 받는 장점이 있다[3]. 또한, deep mertic learning은 label이 같은 데이터는 데이터 사이의 distance를 가깝게 하고 다른 데이터는 distance를 멀리하도록 만들기 위해 데이터를 다른 정해진 공간에 비선형적으로 embedding 시키는 함수를 학습하는 방법을 사용한다. 이때, embedding을 위한 함수를 Convolution Neural Network 등의 deep learning을 통하여 학습시키며 최근 image classification 등의 classification task에서 높은 성능을 보여주었다[4].

Ⅱ. 본론

본 논문에서는 신용 예측 데이터가 보유하고 있는 문제인 imbalance를 개선하기 위해 distance 기반으로 학습하여 imbalance 문제에 강점이 있는 deep metric learning 모델 중 triplet network를 사용하였다. triplet network는 기준이 되는 anchor data(a)를 기반으로 이와 같은 class에 속하는 데이터인 positive data(p)와 다른 class에 속하는 negative data(n)를 한 쌍으로 학습하는 방법을 사용한다. 학습 시에는 3개 data를 embedding 시킨 데이터를 사용하여 anchor와 positive data는 distance를 가깝게 하고 anchor와 negative data는 distance가 멀어지도록 학습한다. 따라서, 3개의 input에 대한 weight를 공유하여 학습하는 구조를 가진다. 또한, neural network를 통해서 embedding(f)을 시킨 데이터 간의 distance(D)를 사용하며 학습을 위한 loss function은 다음과 같다[5].

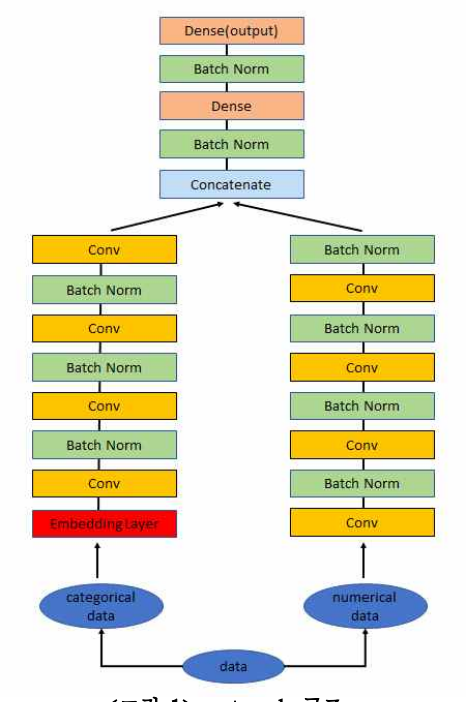
$$D(f(a),f(p)) = \|f(a)-f(p)\|_2^2$$

$$\max\{0, D(f(a),f(p)) - D(f(a),f(n)) + m\}$$

<식 1> triplet loss function

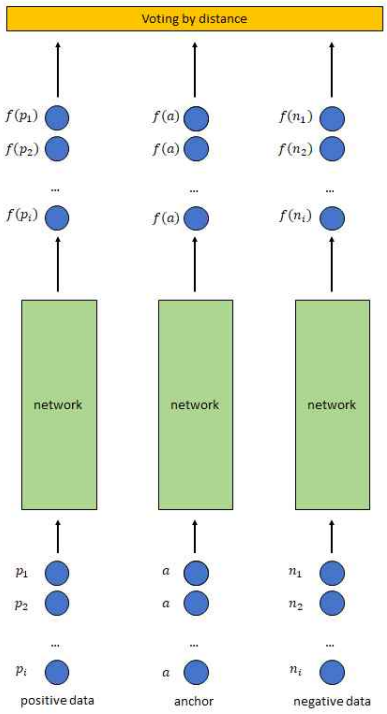
그리고 m은 margin을 의미하는데 이는 anchor와 positive data 사이의 거리와 anchor와 negative data 사이의 거리가 최소한 margin 이상의 차이가 나도록 하여 분류하기 쉽게 만드는 역할을 한다.

하지만, 기존 triplet network는 대부분 이미지 데이터를 사용하는 것에 반해 신용 예측 데이터의 경우 numerical data와 categorical data로 구성된 tabular 형태의 데이터로 구성됨으로써 network 구조의 변경이 필요하다. 따라서, 변경된 network의 구조는 다음과 같다.



<그림 1> network 구조

또한, triplet network는 예측 시에 1개의 pair만을 사용하여 예측하는데 이는 제한된 데이터를 활용하여 positive sample과 negative sample에 따라서 많은 영향을 받아 부정확한 결과를 도출할 수 있다[3]. 따라서 기존 triplet network에 추가로 ensemble 방법인 majority voting을 적용한 방법을 사용하였다. majority voting이란 model을 통하여 다양한 예측이 나왔을 때 이를 다수결의 결정에 따라 최종 결정을 내리는 방법이다. 본 논문에서는 triplet network를 통해 예측 시에 같은 anchor data에 대해서 여러 positive sample과 negative sample로 구성된 각각의 여러 pair를 구성하여 majority voting을 적용해 결과를 도출하였다.



<그림 2> model 구조

실험은 UCI Machine Learning Repository에서 제공하고 있으며 신용카드를 사용한 사람들의 연체 및 연체가 되지 않은 정보를 담고 있는 dataset인 default of credit card clients Data Set을 사용하여 진행하였다. 실험 시에는 dataset에 포함된 정상적으로 갚은 데이터와 연체된 데이터의 비율을 유지하여 각 class의 train set과 test set의 비율을 6:4로 나누어 실험을 진행하였다. 또한, 평가지표는 AUC를 기준으로 평가하였으며 기존에 신용 예측 분야에 사용된 xgboost와 Random Forest 등의 모델과 비교한 결과는 다음과 같다.

AUC = (정상재현율+연체재현율)/2

<식2> AUC 수식

model	AUC
triplet network + majority voting	0.671
xgboost	0.652
Random Forest	0.66

<표 1> 실험 결과 비교

Ⅲ. 결론

해당 논문에서는 대출 등의 행위를 하였을 때 정상적으로 정해진 기간에 갚은 데이터에 비해서 연체된 데이터가 현저히 적은 신용 예측 데이터의 data imbalance 문제를 개선하기 위해 deep metric learning 모델을 제시하였다. deep metric learning은 embedding을 시킨 데이터 간에 distance를 기반으로 학습하고 예측하기 때문에 imbalance 문제에 강점을 보유하고 있다. 또한, 이러한 모델을 이전에 신용 예측 분야에 사용된 xgboost, Random Forest 등의 모델과 비교하였을 때 기존보다 향상된 성능을 보였다.

참 고 문 헌

[1] Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing, 91, 106263.

[2] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.

[3] Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4004-4012).

[4] Lu, J., Hu, J., & Zhou, J. (2017). Deep metric learning for visual understanding: An overview of recent advances. IEEE Signal Processing Magazine, 34(6), 76-84.

[5] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., ... & Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1386-1393).