JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 20, No. 7, pp. 1409-1416, Jul. 2019

딥러닝 기법을 이용한 P2P 소셜 대출 채무자 부도 예측모델에 관한 연구

이 현 진
숭실사이버대학교 ICT공학과

A Study on Prediction Model of Peer-to-Peer (P2P) Social Lending Debtor Default using Deep Learning Technique

Hyun-Jin Lee
Department of ICT Engineering, Korea Soongsil Cyber University, Seoul 03132, Korea

[요 약]

국내외적으로 다양한 P2P 소셜 대출 서비스가 등장하면서 플랫폼 서비스 업자나 투자자 입장에서는 연체나 부도가 발생하지 않을 대출 요청을 찾아서 투자하는 것이 중요하다. 하지만, 대출 요청자의 특성 상 은행에서 대출받는 사람들보다는 높은 연체율을 보이는 것이 사실이다. 따라서, 대출 요청 내용을 분석하여 연체나 부도가 발생하지 않을 대출을 선별하는 것이 중요하다. 본 논문에서는 렌딩 클럽 데이터를 이용하여 딥러닝 기법을 이용한 P2P 소셜 대출 채무자 부도 예측 모델을 개발하였다. 정확도를 높이기 위하여 전체 변수를 사용하고, 학습 속도를 높이기 위하여 계층적 오토 인코더로 특징을 추출하였다. 인공지능 기반 균등 부분 표본 추출로 데이터 클래스를 균일하게 하고, 다층 퍼셉트론을 이용하여 부도 예측을 수행하였다. 렌딩 클럽 데이터에 적용하여 정확도와 정밀도 모두 기존 방법보다 높은 성능을 보이는 것을 확인 할 수 있었다.

[Abstract]

It is important for platform service providers and investors to find and invest in loan requests that do not cause delinquency or default. However, due to the characteristics of the loan requestor, it is true that the default rate is higher than those who are borrowed from the bank. Therefore, it is important to analyze the contents of the loan request to select the loans for which no delinquency or default will occur. In this paper, we developed a prediction model of P2P social loan debtor 's default using deep learning method for lending club database. We used all parameters to increase the accuracy and extracted features using stacked auto encoder to increase learning speed. Using AI based balanced sub sampling the data class is uniformized and the default prediction is performed using the multi-layer perceptron. As a result of applying it to the lending club database, it was confirmed that both the accuracy and the precision outperforms other classifiers.

색인어 : 딥러닝, P2P 소셜대출, 대출 부도, 예측 모델, 이진 분류

Key word : Deep Learning, P2P Social Lending, Lending Default, Prediction Model, Binary Classification

http://dx.doi.org/10.9728/dcs.2019.20.7.1409

Received 13 June 2019; **Revised** 10 July 2019
Accepted 25 July 2019

***Corresponding Author; Hyun-Jin Lee**

Tel: +82-2-708-7863
E-mail: hjlee@mail.kcu.ac

Ⅰ. 서 론

인터넷 등 통신 기술의 발달과 아이디어를 빠르게 실현할 수 있는 정보 통신기술의 발달되어 벤처 기업은 전 세계적으로 증가하고 있다. 벤처 기업이 성공적으로 성장하기 위해서 필요한 것은 초기 자본을 확보이다. 하지만 실적이 없는 벤처 기업이 전통적인 금융 시장에서 자본을 확보하는 것은 쉽지 않다. 벤처 캐피탈이나 벤처 투자 업체를 통해 자금을 확보할 수 있지만, 각 업체에서 기준에 따라 벤처 기업을 평가하기 때문에 적시에 자금을 확보하는 것은 쉽지 않다. 최근 몇 년 동안 벤처 기업들에게는 P2P 소셜 대출 이나 크라우드 펀딩 (Crowdfunding)이 금융 업체를 거치지 않고 빠르게 자금을 확보할 수 있는 새로운 방법으로 부상했다[1].

전통적으로 대출에 있어서 연체율에 대한 관리는 중요한 문제중 하나였다. 대출에 대한 상환이 모두 이루어지는 것을 희망하지만, 다양한 원인에 의해 채무불이행은 발생할 수밖에 없고, 채무 불이행율을 최소화하기 위해 금융 기관들은 심사 단계에서부터 상환 능력을 다양한 관점에서 살펴본다. 그렇기 때문에 은행 등의 제1금융권에서 대출을 못 받는 사람들은 저축은행, 캐피탈 등 제 2금융권에서 대출을 신청하고, 이 곳에서도 대출을 못 받는 사람들은 대부업체나 사채를 사용하게 된다. 일반적으로 제1금융권의 대출 이자보다 대부업체의 대출 이자가 높기 때문에 저 신용자들은 대출의 악순환에 빠지기 쉽다.

P2P 대출은 이러한 저 신용자들이 대출을 받을 때 저렴한 이자를 책정하여 대출자의 부담은 줄이고, 투자자들은 은행 이자보다 높은 이자를 제공받음으로써 높은 투자 수익을 실현할 수 있도록 하는 서비스이다. 하지만, 대출자가 저 신용자라는 특성상 연체나 부도의 가능성이 높고, 서비스의 신뢰도를 높이기 위해서는 연체나 부도의 가능성이 적은 대출자들을 선택하여 대출을 진행하여야 한다.

본 논문에서는 P2P 소셜 대출에서 대출자의 부도 여부를 딥러닝에 기반하여 예측하는 방법을 제안한다. 먼저 부도가 발생한 대출이 전체 대출 대비 적은 비중을 차지하기 때문에 분류 성능이 떨어지는 요인이 된다. 따라서 인공지능 기반 균등 부분 표본 추출 (AI based balanced sub sampling)을 적용하여 데이터의 균형도를 높였다. 전통적인 방법에서는 속도 및 이해도 측면에서 데이터 속성을 선별하여 10개 이내만 사용하였지만, 전체 속성을 활용하여 사람이 인식할 수 없는 숨겨진 속성을 활용하여 성능을 향상시키고자 한다. 그러기 위해서 딥러닝 기반의 계층적 오토 인코더 (Stacked Auto Encoder)를 이용하여 특징 벡터를 추출하고, 이를 입력 데이터로 사용하여 신경회로망 (Neural Networks)을 학습시켜서 P2P 소셜 대출에서의 부도를 예측한다.

본 논문의 구성을 다음과 같다. 제 2장에서는 P2P 소셜 대출과 관련된 선행 연구를 정리한다. 제 3장에서는 제안하는 딥러닝 기반 P2P 소셜 대출의 부도 예측 모형을 설명한다. 제 4장에서는 제안하는 부도 예측 모형을 실 데이터에 적용하여 결과를 비교 분석하고, 제 5장에서는 결론 및 향후 연구방향에 대하여 살펴본다.

Ⅱ. 관련연구

2-1 표본 추출 방법 (Sampling Methods)

불균형한 데이터로 학습을 수행하면 학습 결과의 정확도가 떨어진다. 불균형한 데이터에서의 문제는 일반적인 학습 알고리즘은 자주 발생하는 클래스만을 정답으로 인식하고, 희박한 클래스는 무시하는 것이다.

데이터 관점에서 불균형한 클래스를 해소하는 방법은 클래스의 분포를 변형시켜 균형 잡힌 표본을 생성하는 것이다. 데이터를 표본 추출하는 다양한 방법이 있다. 가장 일반적인 방법은 희박한 클래스를 재 선택하여 중복 표본 추출하는 추가 표본 추출 (over sampling) 방법과 자주 발생하는 클래스의 데이터에서 표본 추출하는 축소 표본 추출 (under sampling) 방법이 있다. 추가 표본 추출 방법은 가장 단순한 형태로 소수의 클래스를 복제함으로써 중복된 값이 과다 생성되어 과적합이 발생할 수 있다. 축소 표본 추출은 빈번한 클래스에서 데이터를 제거하는 것으로 중요한 데이터를 잃어버릴 수 있다. 하지만, 이해하기 쉽고 구현하기 편리하기 때문에 다양한 분야에서 연구되고 있다[2][3].

단순한 추가 표본 추출과 축소 표본 추출의 단점을 보완하기 위하여 다양한 연구가 진행되고 있다. Lee는 희박한 클래스의 데이터에 표준 잡음(noise)를 추가하여 고정된 수의 반복이 생성되게 하였다. 다수의 반복을 통해 클래스의 발생 확률을 평준화 시킬 수 있었다[4]. Chawla 등은 합성 소수 추가 표준 추출 기법 (Synthetic Minority Oversampling Technique : SMOTE)을 제안했다[5]. 희박한 학습 데이터에 대해 특징 공간에서 데이터에 가까이 있는 이웃들을 연결하는 선상에 존재하는 데이터를 무작위로 선택하여 새로운 데이터를 생성하였다. Menardi 등은 무작위 추가 표본 추출 (Random Over Sampling Examples : ROSE)을 제안했다[6]. 이 방법은 커널 함수로 특징 공간을 변경하고, 희박한 학습 데이터에 대한 이웃들을 선택하여 인공적인 데이터를 생성하였다.

이와 같이 관찰되지 않은 새로운 인공적인 데이터를 생성하여 과적합의 위험을 줄이고 추가 표본 추출 방법에 의해 손상된 일반화 능력을 향상시키는 방법이 연구되고 있다.

2-2 딥러닝(Deep Learning)

오토 인코더는 딥러닝 (Deep Learning) 알고리즘의 하나로 비지도 학습 (Unsupervised Learning) 방법이다[7]. 오토 인코더는 입력 데이터를의 차원을 축소하는 인코더 (Encoder)와 압축 코드를 원본 데이터와 가깝게 재구성하는 디코더 (Decoder)로 구성된다. 오토 인코더는 신경망의 목표 값이 학습하는 비지도

학습 방법이다. 하나의 은닉층 (Hidden Layer)이 있는 오토 인코더의 구성은 (그림 1)과 같다.

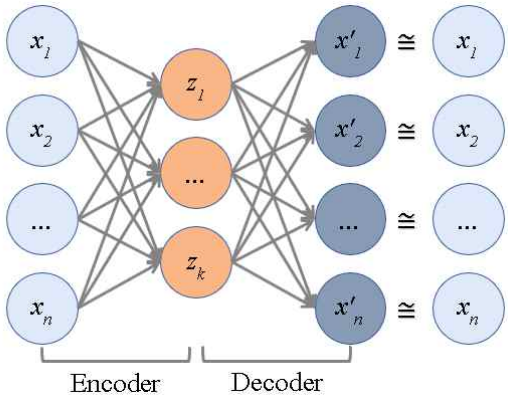

그림 1. 1층의 은닉층을 가지는 오토인코더
Fig. 1. Autoencoder with One Hidden Layer

오토 인코더는 수식 (1)과 같이 입력 데이터 X를 그대로 모사하는 함수 F를 의미한다.

$$F(X) \cong X \tag{1}$$

함수 F는 수식 (2)와 같이 인코더에 해당하는 함수 E와 디코더에 해당하는 함수 D로 구성된다.

$$F = E \bullet D \tag{2}$$

수식 (3)에 있는 인코더의 출력 Z는 X의 특징 벡터가 되고, 이것을 통해 입력 데이터 X의 차원을 축소할 수 있고, Z를 디코더에 대입하여 X의 원형을 모사할 수 있게 된다.

$$E(X) = Z,\ D(Z) = X',\ X' \cong X \tag{3}$$

오토 인코더의 학습 방법은 오류 역전파로 (Error Backpropagation) 은닉층이 하나일 때의 성능은 좋지만, 은닉층이 두 개 이상일 때는 은닉층이 하나일 때와 비교하여 학습하는 시간이 오래 걸리면서, 성능 차이가 크지 않다. 또한, 층을 따라서 계속 작은 가중치의 오차 값들이 수정되는 과정에서 처음 인식됐던 가중치의 오차의 기울기가 마지막 층에서는 사라지는 문제(Vanishing Gradient)가 발생할 수 있는데, 계층적 오토 인코더 (Stacked Auto Encoder)로 문제를 해결할 수 있다[8].

2-3 채무자 부도 예측

P2P 소셜 대출과 관련된 연구는 P2P 소셜 대출과 기존 은행 금융거래의 비교연구, 차입자의 신용정보와 대출정보를 이용하여 대출 상환의 성공 요소에 대한 연구, 차입자와 투자자의 사회적 활동 변수를 도출한 연구등이 수행되었다. 전통적으로 금융기관의 대출상환 성공요인에 대한 연구는 인구통계학적 변수와 재무적 변수에 집중하여 연구가 수행되었다. 인구통계학적 변수는 연령, 교육, 결혼유무, 성별, 직업 등이며, 재무적 변수는 소득, 주거형태, 신용등급, 사금융 사용유무, 신용카드 연체율 등이 있다.

P2P 소셜 대출 플랫폼을 통한 대출의 부도 여부를 예측하는 연구도 다양하게 이루어져 왔다. Malekipirbazari 등은 랜덤 포레스트 (Random Forests) 알고리즘을 사용하여 채무자의 부도 여부를 판단하였다[9]. Byanjankar 등은 유럽 P2P 대출에 대해 인공 신경망을 적용하여 사용자에게 신용점수를 제공하는 방법을 제안하였다[10]. 채무자의 기본 정보와 총부채 상환 비율과 수입, 지출 비율 등의 특징을 추가하여 학습하였고, 기존 신용 평가 방법보다 좋은 결과를 보였다. Yanghong등은 사례(instance)기반의 로지스틱 방법을 적용하여 유사한 대출의 상환과 부도를 측정하였다[11]. Kumar 등은 세개의 앙상블 분류기를 학습하여 채무이행 여부를 판단하는 방법을 제안하였다[12]. Lin 등은 중국 P2P 대출 플랫폼 데이터에 로지스틱 회귀모형을 적용하여 채무 불이행에 영향을 미치는 요소를 파악 하였다[13]. 채무자의 인구통계학적 특성과 재정적인 특성이 채무 불이행에 영향을 주는 요소라고 하였다. Serrano-Cinca 등은 P2P 플랫폼인 렌딩 클럽(Lending Club)에서 제공하는 데이터에 의사결정 나무를 적용하여 채무이행 여부를 예측하였다[14].

Ⅲ. 제안하는 시스템

3-1 제안하는 시스템의 구성

소셜 대출의 채무자 부도 예측을 위해 제안하는 방법은 데이터 전처리 (Preprocessing)와 인공지능 기반한 특징추출, 균등 부분 표본 추출 (AI based balanced sub sampling), 다층퍼셉트론을 이용한 예측으로 으로 구성되며 그 구성은 (그림 2)와 같다. 먼저, 데이터를 전처리한 후 특징의 개수를 줄이기 위한 차원 축소 알고리즘을 적용한다. 데이터의 특성상 부도의 비율이 낮기 때문에 불균형 한 데이터 분포를 클래스 별로 균등하게 분포하도록 인공지능 기반 균등 표본 추출을 수행한다. 마지막으로 다층 퍼셉트론을 이용하여 학습을 한다.

3-2 데이터 전처리

데이터 전처리는 데이터 값의 범위를 조정하는 것이다. 채무 데이터의 속성 값은 성별은 남성, 여성의 범주형이고, 연봉은 [0, 20,000]의 범위를 가지는 연속형 수이고, 나이는 [0, 120]을 가지는 연속형 수로 각 속성들의 값의 범위는 일정하지 않다. 이러한 경우 머신 러닝의 분류 결과는 값이 큰 속성에 영향을 크게 받기 때문에 결과가 좋을 수 없다. 따라서, 정규화 (Normalization)을 통해 속성의 값의 범위에 의한 영향을 최소

화한다. 먼저, 범주형 변수는 이산형 변수로 변환하였다. 이산형 변수와 연속형 변수는 최소-최대 정규화 (Min-Max Normalization)를 하여 데이터를 [0, 1]의 연속형 변수로 생성하였다.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (4)$$

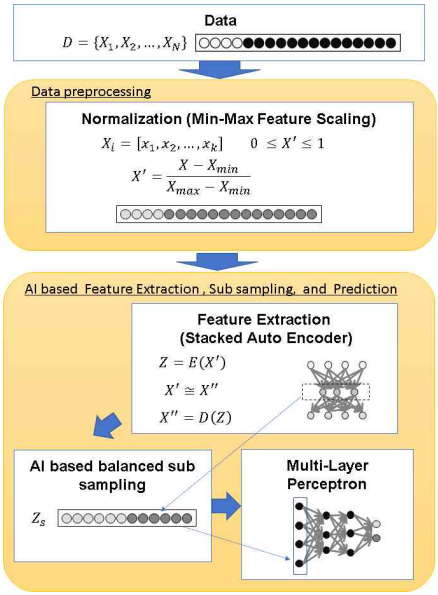

그림 2. 제안하는 분석 시스템의 구성
Fig. 2. Suggested Analysis System

3-3 딥러닝 기반 차원 축소 학습 방법

P2P 소셜 대출에서 데이터의 속성은 모두 145개이다. 기존 분석에서 사용한 기법들은 알고리즘 특성상 이 속성들을 모두 사용하지 못하고, 성별, 나이, 지역 등 인구통계학 정보와 연봉, 주거형태 신용도 등 재정적 정보의 10여개 내외의 속성만 사용하였다. 하지만, 데이터에 숨겨진 정보를 발견하기 위해서는 활용할 수 있는 모든 속성을 사용하는 것이 좋다. 그러나, 모든 속성을 다 사용하면 알고리즘의 학습 시간이 오래 소요되고, 학습 데이터에 과적합이 되어 성능이 떨어질 수 있다. 따라서, 딥러닝에서는 데이터의 속성 중 중요한 요소를 선택하기 위한 특징 추출 (Feature Extraction) 기법을 사용한다. 특징의 개수를 줄이는 방법 중 특징 선택 (Feature Selection)은 원본 데이터에서 불필요한 요소들을 제거하여 간결한 특징 벡터를 만드는 것으로 여러 특징 중 중요한 특징이 무엇인지 알 수 있다. 하지만, 사용자의 경험적 지식이 필요하거나, 특징의 수가 많을 때 특징 선택의 결과가 좋지 않은 경우가 많다. 특징 추출은 원본 특징들을 조합하여 새로운 특징을 생성하는 것으로, 겉으로 드러난 형태는 일부 특징을 선택한 것처럼 보이지만, 전체 특징에 대한 비선형 변환이 이루어 진 것으로 모든 특징이 사용되는 방법이다.

데이터가 이미지일 때 많이 사용하는 컨볼루션 신경회로망 (Convolutional Neural Networks: CNN)의 경우에는 컨볼루션 층 (Convolution Layer)을 조합하여 특징 추출이 이루어지지만 연체 데이터와 같은 비즈니스 데이터의 경우에는 컨볼루션 신경회로망을 적용하기 어렵기 때문에 계층적 오토 인코더 (Stacked Auto Encoder)를 적용하여 특징 벡터를 추출하였다. 계층적 오토 인코더는 3층으로 오토 인코더를 3번 적층하여 145개의 속성에서 20개의 특징을 추출하였다.

3-4 인공지능 기반 균등 부분 표본 추출 (AI based balanced sub sampling)

분류 문제에서 분류 성능에 영향을 끼치는 요소중의 하나는 분류할 클래스 (Class)의 균형이다. 일반적으로 대출의 연체율은 2%에서 5% 정도로 정상 대출이 연체 대출보다 더 많은 불균형한 데이터이다. 이러한 데이터의 경우 데이터의 균형 정도를 높이면, 분류 성능을 더 높일 수 있기 때문에 인공지능 기반 균등 부분 표본 추출을 제안하였다.

표본 추출로 많이 사용하는 무작위 표본 추출 (random sampling)은 중요한 정보를 잃어버릴 가능성이 크고, ROSE 알고리즘은 기존에 없던 새로운 데이터를 생성하는 것이기 때문에 불필요하거나 실세계에서는 실현 불가능한 데이터가 생성될 수 있다. 따라서, 서포트 벡터 머신(Support Vector Machine: SVM)의 경계에 있는 서포트 벡터(Support Vector)를 선택해서 클래스를 구분하는데 영향을 주는 중요 데이터를 많이 선택하고, 임의의 데이터가 생기는 것을 최소화한다. 상세한 알고리즘은 <표 1>과 같다.

표 1. 인공지능기반 균형 부분 샘플링 알고리즘
Table 1. AI based Balanced Sub Sampling Algorithm

X : dataset

Run SVM (Support Vector Machine)
X_s:get all dataset which is smaller class
X_l:get Support Vectors which is larger class

If (number of X_s> number of X_l)
    get random sampling from larger class
else
    get a ROSE balanced sample from X_s

데이터가 적은 클래스의 데이터는 모두 사용하고, 데이터가 많은 클래스의 데이터는 서포트 벡터를 사용한다. 두 클래스의 데이터 개수가 여전히 불균형하다면, 데이터가 적은 클래스에는 ROSE 알고리즘을 적용하고, 데이터가 많은 클래스에서는 임의 추출을 적용하여 두 클래스의 균형도를 맞춘다.

3-5 모델 학습

인공지능 기반 균등 부분 표본 추출을 적용하여 균형 잡힌 데이터 집합이 이루어지면, 다층퍼셉트론(Multi Layer Perceptron:MLP)을 적용한다. 다층퍼셉트론의 구조는 입력층은 20 개의 노드이고, 은닉층은 40 X 10개의 노드의 2 층(Layer)으로 구성된다. 출력층은 2개의 노드이다. 일반화 성능을 향상시키기 위하여 드롭아웃(dropout)을 모든 층에 10%씩 적용하였다.

Ⅳ. 실험 환경 및 결과

4.1 실험 데이터

실험에 사용한 데이터는 P2P 소셜 대출 업체인 렌딩 클럽(Lending Club)의 공개용 데이터베이스를 활용하였다[15]. 렌딩 클럽은 2007년부터 모든 대출 거래에 대한 데이터를 151개의 변수와 함께 공개하고 있다. 그중 2008년 4월부터 2018년 9월까지 약 10년간의 대출 거래 데이터 2,260,668 건을 사용하였다.

<표 2>는 본 연구에서 목표 변수로 사용하는 채무 상태(Loan Status) 별 데이터 수 이다. 렌딩 클럽은 채무 상태를 연체 정도와 상환 여부에 따라 진행중인 대출 (Current), 연체일이 15일 이하인 대출 (In Grace Period), 연체일이 16~30일 사이인 대출 (Late (16-30 days)), 연체일이 31일 ~ 120일 사이인 대출 (Late (31-120 days)), 상환 완료 대출 (Fully Paid), 채무 불이행 (Default)과 사용할 수 없는 데이터 (Does not meet the credit policy)의 일곱 가지로 구분하고 있다.

본 연구의 목적은 채무자 부도 예측 모형 개발이므로 현재 진행중인 데이터 (Current)와 채무 불이행이 발생 할지 아직 알 수 없는 데이터 (In Grace Period, Late (16-30 days), Late (31-120 days)), 사용할 수 없는 데이터를 제외하고, 상환 완료 (Fully Paid)와 채무 불이행 (Default)을 목표 값으로 사용하였다. 상환 완료된 데이터는 1,041,952 건이고, 채무 불이행 데이터는 261,686 건이다.

표 2. 재무상태 (2008.4 ~ 2018. 9)
Table 2. Loan Status (from Apr. 2008 to Sep. 2018)

Status	Count
Current	919,695
In Grace Period	8,952
Late (16-30 days)	3,737
Late (31-120 days)	21,897
Fully Paid	1,041,952
Default	261,686
Does not meet the credit policy	2,749

렌딩 클럽 데이터의 151 변수 중 아이디(Id), 고객 이름 등 개별 데이터를 구분하기 위한 변수와 채무 불이행 여부를 확인하는데 큰 영향을 끼치는 연체 일수, 연체 금액 등 예측에 사용할 수 없는 변수를 제외한 73개 변수의 데이터에 전처리를 수행하였다. 또한, 전통적인 분류 알고리즘에서 성능의 문제로 소수의 변수를 사용하기 때문에 비교를 위해 핵심 변수로 대출 총금액, 이율, 대출 기간 등의 대출 정보 변수와 신용등급, 재직기간, 주택 소유 여부, 연봉, 대출 기간 등 대출자의 인구통계학정 정보 변수를 포함한 13개의 독립 변수를 선정하였다. 전체 대출 건수 1,303,638 건 중 70% (912,546 건)을 학습용 데이터 집합 (train set)으로 사용하고, 30% (391,092 건)을 검증용 데이터 집합 (test set)으로 사용하였다. 또한, 연구결과의 일반성능을 비교하기 위하여 10회의 상호 검증 방법 (10-fold cross-validation method)을 사용하였다.

4-2 실험결과

본 논문에서 제시하는 알고리즘의 성능을 비교하기 위하여 다양한 벤치마크 모델을 비교하였다. 분류 알고리즘은 로지스틱 모델 (LR, Logistic Regression), 의사 결정 나무 (DT, Decision Tree), 다층 퍼셉트론 (MLP, Multi-Layer Perceptron)의 3가지를 이용하였다. 데이터 집합은 13개의 독립 변수를 선별하여 사용한 경우 (IV, Independent Variables), 73개의 전체 변수를 사용한 경우 (AV, All Variables)와 ABS 알고리즘에 기반하여 20개의 변수를 추출하여 (Feature Selection) 데이터의 균형을 맞춘 제안하는 방법이다. 실험 결과는 다음 <표 3>과 같다.

독립 변수에 로지스틱을 적용했을 때 (LR - IV) 정확도는 62.58 %이고, 전체 변수에 로지스틱을 적용했을 때 (LR - AV) 정확도는 81.81 %로 전체 변수를 사용하는 것이 정확도가 높았다. 마찬가지로 독립 변수에 의사 결정 나무를 적용했을 때 (DT – IV) 정확도는 79.00 %이고, 전체 변수에 의사 결정 나무를 적용했을 때 (DT – AV) 정확도는 83.62 %로 전체 변수를 사용하는 것이 정확도가 높았다. 다층 퍼셉트론의 경우에도 독립 변수에 적용했을 때 (MLP – IV) 81.19 %, 전체 변수에 적용했을 때 (MLP – AV) 86.53 %로 전체 변수를 사용하는 것이 높은 정확도를 보였다. 이처럼 일부의 독립 변수를 사용하는 것보다 전체 변수를 사용하는 것이 학습 시간은 오래 걸리지만, 모든 알고리즘에 대해서 최소 5 % 이상 높은 성능을 보였다. 알고리즘 별로 살펴보았을 때, 독립 변수를 사용한 경우와 전체 변수를 사용한 경우 모두 다층 퍼셉트론의 성능이 우수하였다. 제안하는 알고리즘의 성능을 보면, 89.76 %로 전체 변수에 다층 퍼셉트론을 적용한 것보다 우수한 성능을 보였다. 따라서, 클래스의 데이터의 분포를 균등하게 만드는 것이 분류 알고리즘의 성능을 향상시키는데 중요한 영향을 준 것을 확인할 수 있었다.

표 3. 테스트 데이터에 대한 기본예측 모델의 정확도
Table 3. Accuracy of Default Prediction Models for test set

	LR - IV	DT - IV	MLP - IV	LR - AV	DT - AV	MLP - AV	Suggested
1 Fold	0.6134	0.7962	0.8286	0.8250	0.8108	0.8630	0.8878
2 Fold	0.6316	0.7704	0.8202	0.8397	0.8501	0.8424	0.9066
3 Fold	0.6457	0.7889	0.8105	0.8357	0.8316	0.8657	0.9154
4 Fold	0.6266	0.8075	0.8236	0.8046	0.8478	0.8866	0.8992
5 Fold	0.6498	0.7707	0.8314	0.8318	0.8386	0.8368	0.8843
6 Fold	0.6022	0.7994	0.8101	0.7962	0.8210	0.8722	0.9166
7 Fold	0.6288	0.7697	0.7923	0.8388	0.8270	0.8678	0.8812
8 Fold	0.6155	0.7994	0.7806	0.8156	0.8599	0.8955	0.8767
9 Fold	0.6302	0.8124	0.8207	0.8196	0.8475	0.8490	0.8984
10 Fold	0.6142	0.7854	0.8010	0.7740	0.8277	0.8740	0.9098
Average	0.6258	0.7900	0.8119	0.8181	0.8362	0.8653	0.8976

본 연구의 목적은 채무자의 부도 여부를 예측하는 것이기 때문에 정확도 (Accuracy) 뿐만 아니라 정밀도(Precision)도 높은 것이 좋다. 정확도는 식(5)와 같이 계산되며, 전체데이터 중 예측한 클래스가 실제 클래스와 일치하는 비율이다.

$$Accuracy = \frac{True\ Positive + True Negative}{True\ Positivie + False Negative + False\ Positive + True Negative} \quad (5)$$

정밀도는 식(6)과 같이 계산되며 한 클래스의 데이터 중 해당 클래스를 예측한 클래스가 실제 해당 클래스와 일차하는 비율이다.

$$\text{Precision} = \frac{True\ Positive}{True\ Positivie + False\ Positive} \quad (6)$$

정밀도는 연체로 예측된 데이터 중 실제 연체 데이터의 비율로 <표 4>의 오차 행렬 (Confusion Matrix)을 통해 계산한다.

표 4. 오차행렬
Table 4. Confusion Matrix

		True Condition	
		True Condition Positive	True Condition Negative
Predicted Condition	Predicted Condition Positive	True Positive	False Positive
	Predicted Condition Negative	False Negative	True Negative

전체 데이터에 대한 로지스틱 모델, 의사 결정 나무, 다층 퍼셉트론과 제안하는 알고리즘에 의한 오차 행렬은 <표 5>와 같다. 전체 데이터에 대한 로지스틱 모델의 정밀도는 83.35 %, 의사 결정 나무의 정밀도는 86.65 %, 다층 퍼셉트론의 정밀도는 87.12 %, 제안하는 알고리즘의 정밀도는 89.23 %로 제안하는 알고리즘이 정확도와 정밀도 측면에서 우수하였다.

표5. 테스트 데이터에 대한 기본예측 모델의 정확도
Table 5. Confusion Matrix of Default Prediction Models for test set

		Actual	
		Paid	Default
Predicted	Paid	254,342	12,958
	Default	58,196	65,596

(a) LR, Logistic Regression

		Actual	
		Paid	Default
Predicted	Paid	259,062	10,586
	Default	53,476	67,968

(b) DT, Decision Tree

		Actual	
		Paid	Default
Predicted	Paid	269,977	10,116
	Default	42,561	68,438

(c) MLP, Multi-Layer Perceptron

		Actual	
		Paid	Default
Predicted	Paid	280,939	8,463
	Default	31,599	70,091

(d) Suggested Method

Ⅴ. 결 론

본 논문에서는 대출자의 연체 가능성을 예측하기 위하여 딥러닝 기법을 이용한 채무자 부도 예측 모델을 제안하였다. 먼저

대출 신청할 때 사용하는 정보가 많을수록 예측 성능이 올라가기 때문에 전체 변수를 이용한다. 하지만, 변수가 많을수록 학습 시간이 오래 걸리고, 특정 변수에 영향을 받을 수 있기 때문에 딥러닝 기법을 이용한 특징 추출 기법을 적용하여 변수의 개수를 축소하였다. 또한, 데이터의 특성 상 정상 고객과 연체 고객의 비율은 차이를 보이고, 이 차이가 클수록 성능은 비율이 높은 고객만을 잘 분류할 것이기 때문에 인공지능 기반 균등 부분 표본 추출 (AI based balanced sub sampling)을 적용하여 정상 고객과 연체 고객의 비율을 동일하게 만들었다. 제안하는 방법을 렌딩 클럽 (Lending Club)의 실 데이터에 적용한 결과 정확도와 정밀도가 향상된 것을 확인할 수 있었다.

본 연구에서는 대출자의 인구통계학적 정보와 재무적 정보와 같은 객관적인 신용정보를 기반으로 채무자 부도 예측 모델을 개발하였다. 하지만, 인간은 같은 상황에서도 개인의 성향에 따라 행동의 우선 순위가 달라질 수 있다. 예를 들어, 동일한 재산을 가지고, 동일한 직장에 동일한 연봉을 받고 있던 사람이 해당 직장에서 퇴직을 하게 되었을 때 개인적인 성향에 따라 어떤 사람은 퇴직금을 활용해서 기존 대출을 갚고 새로운 직장을 찾기도 하고, 어떤 사람은 기존 대출에 더하여 추가 대출을 받고 해외로 나가는 경우도 있다. 이처럼 개인의 주관에 따라 다른 경향을 보이기 때문에 개인의 주관적 성향을 파악할 수 있는 변수를 고려한 예측 모형 개발이 필요하다.

참고 문헌

[1] Y. Zhao, and P. Harris, W. Lam, “Corwdfunding industry – History, development, policies, and potential issues,” *Journal of Public Affairs*, Vol. 19, Issue 1, 14 March 2019.

[2] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, “Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance”, *Neural Networks*, Vol. 21, No. 2, pp. 427-436, 2008.

[3] J. Burez, and D. Van den Poel, “Handling class imbalance in customer churn prediction”, *Expert Systems with Applications*, Vol. 36, No. 3, pp. 4626-4636, 2009.

[4] S. S, Lee, “Noisy replication in skewed binary classification, “ *Computational Statistics and Data Analysis*, Vol. 34, No. 2, pp. 165-191, 2000.

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

[6] G. Menardi, and N. Torelli, “Training and assessing classification rules with imbalanced data,” Data Mining and Knowledge Discovery, Vol. 28, No. 92 – 122, 2014.

[7] H. J. Lee, “Study for Prediction System of Learning Achievements of Cyber University Students using Deep Learning based on Autoencoder,” *Journal of Digital Contents Society*, Vol. 19, No. 6, pp. 1115 – 1121, 2018.

[8] P. Domingos, "*The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*," Basic Books, 2015.

[9] M. Malekipirbazari, and V. Aksakalli, “Risk assessment in social lending via random forests,” *Expert Systems with Applications,* Vol. 42, No. 10, pp. 4621-4631, 2015.

[10] A.Byanjankar, M. Heikkilä, and J. Mezei, “Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach,” *in Proceeding of the* 2015 *IEEE Symposium Series on Computational Intelligence*, pp. 719-725, 2015.

[11] Y. Guo, W. Zhou, C. Luoa, C. Liu, and H. Xiong, “Instance-based credit risk assessment for investment decisions in P2P lending,” *European Journal of Operational Research*, Vol. 249, No. 2, pp. 417-426, 2016.

[12] L. Vinod Kumar, S. Natarajan, S. Keerthana, .K. M. Chinmayi, and N. Lakshmi, “Credit Risk Analysis in Peer-to-Peer Lending System,” *in Proceeding of the2016 IEEE International Conference on Knowledge Engineering and Applications*, pp. 193-196, 2016.

[13] X. Lin, X. Li, and Z. Zheng, “Evaluating borrower’s default risk in peer-to-peer lending: evidence from a lending platform in China,” *Applied Economics*, Vol. 49, No. 35, pp. 3538-3545, 2017.

[14] C. Serrano-Cinca and B. Gutiérrez-Nieto, “The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending,” *Decision Support Systems*, Vol. 89, pp. 113-122, 2016.

[15] Lendingclub. LendingClub Statistics [Internet]. Available: https://www.lendingclub.com/info/download-data.action

이현진(Hyun-Jin Lee)

1996년: 순천향대학교 전산학과
공학사

1998년: 연세대학교 대학원
컴퓨터과학과 공학석사

2002년: 연세대학교 대학원
컴퓨터과학과 공학박사

2003년~현재: 숭실사이버대학교 ICT공학부 부교수

※관심분야 : 기계학습, 빅데이터, 온라인교육 등