딥 러닝 알고리즘을 통한 금융 이상거래 탐지 모델

윤민욱, 설유환, 김진°

중앙대학교 컴퓨터공학부, 시큐레이어 인텔리전스 기술연구소°

A Credit Card Fraud Detection Model based on Deep Learning Algorithms

Yoon Min Wook, Seol Yoo Hwan, Kim Jin°,

Seculayer Co., Ltd.

{ymw032, adaan12}@cau.ac.kr

jinkim@seculayer.co.kr

요 약

본 연구는 높은 오탐률을 보이는 기존의 룰 방식 이상거래 탐지 시스템의 한계를 극복하기 위해 DNN과 GAN을 이용한 딥러닝 기반 이상거래 탐지 시스템을 연구하고, 2013년 9월에 발생한 유럽의 카드 거래 데이터를 이용해 테스트하였다. DNN과 GAN 알고리즘을 이용해 학습 모델을 만들고, 거래 데이터를 이용해 검증하여 정확도와 정탐률, 오탐률을 계산한 결과 기존의 머신 러닝 알고리즘보다 딥러닝 모델이 더 좋은 결과를 보여주는 것을 확인할 수 있었다.

1. 서론

전자금융 거래량의 증가 및 간편 결제 수단의 다양화로 전자금융 이용자 수는 지속해서 증가하고 있다. 최근에는 은행거래의 90%가 비대면을 통해 이루어지고 있으며 이와 함께 보이스피싱이나 악성코드를 이용한 개인 정보 유출 및 자금 이체 사고 또한 증가하고 있어 사회적으로 큰 문제가 되고 있다.

이에 많은 금융회사가 이상 거래 탐지 시스템을 도입 및 운영하고 있지만, 금융 사기 방식은 지속해서 진화하고 끊임없이 이루어지므로 이에 대응하기 위한 더욱 효과적인 방법이 요구된다.

금융권에서 도입한 이상 거래 탐지시스템은 현재까지 사고 거래 분석을 기반으로 한 시나리오 기반 룰 탐지 수준에 머물고 있고, 높은 오탐률로 인한 고객의 불편과 정책 운용의 어려움을 초래하고 있어 한계가 보인다.

룰 탐지의 경우 이상 거래 시나리오를 분석해 룰을 도출하기 때문에 비슷한 시나리오에 대한 정탐률은 높을 수 있으나, 정상거래에 대한 분석이 미흡하게 되어 오탐률 또한 높을 수밖에 없다. 따라서 시나리오 기반 룰 탐지 방식의 한계를 벗어나기 위해서는 새로운 방식의 이상 거래 탐지시스템이 필요하다.

본 연구에서는 최근 부상하고 있는 딥 러닝 알고리즘인 DNN과 GAN을 활용해 거래 로그를 학습하여 정탐률과 오탐률, 정확도 등을 계산하여 비교한다.

2. 관련 연구

딥 러닝을 포함한 머신 러닝 알고리즘을 이용해 이상 거래를 탐지하는 연구는 꾸준히 진행되고 있고, 현 금융업계에서도 조금씩 적용하려 하는 추세이다.

Igor Anohhin의 연구에서는 로지스틱 회귀와 의사결정 트리, 자기 조직화 지도 등의 머신 러닝 알고리즘을 적용하여 이상거래 탐지 모델을 구성하고 실험하였다 [1]. 실제 은행 거래 데이터가 아닌, R을 이용하여 만든 데이터를 이용해 머신 러닝 알고리즘을 적용하여 탐지율 및 오탐률을 비교한 연구도 있다. Lusis Payments에서는 로지스틱 회귀와 랜덤 포레스트 알고리즘을 적용해 이상 거래 탐지율과 오탐률 등을 비교한 연구결과를 제시하였다. 비록 이상 거래 비율이 5%로 구성된 가상 데이터를 사용하여 실세계를 정확하게 반영하지는 못했지만, 비교적 높은 탐지율과 낮은 오탐률을 도출하였다 [2]. 반면 Raghavendra Patidar와 Lokesh Sharma의 연구에서는, 모델은 BPN(Back Propagation Network)를 선택해 학습을 시키고, 알고리즘의 파라미터는 GA(Genetic Algorithm)을 이용하여 선택하는 알고리즘인 GANN(Genetic Algorithm and Neural Network)을 제시했다 [3].

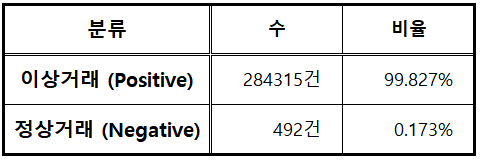
이처럼 이상거래 탐지 분야에 있어서 여러 가지 알고리즘을 적용한 연구가 많이 진행되고 있다. 그러나 아직 모든 거래 로그에 좋은 결과값을 산출하는 알고리즘은 없으므로, 앞으로 더 많은 머신 러닝 알고리즘이 연구될 것으로 생각된다.

3. DNN과 GAN을 이용한 이상거래 탐지 방법

3.1 데이터 셋

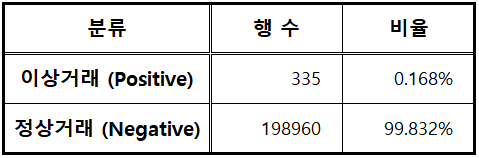
본 연구에 사용된 카드 거래 데이터는 개인 정보의 문제로 사람이 특징들을 식별할 수 없도록 주성분 분석(Principal Component Analysis) 되어 있는 상태이다 [5].

데이터의 구성은 [표 1]과 같다. 실세계에서 이루어지는 거래량 중 이상거래의 비율은 아주 낮다.

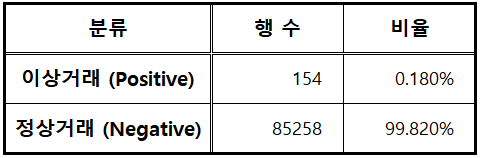


[표 1] Dataset class statistic.

거래 로그 데이터를 훈련 데이터 셋과 테스트 데이터 셋으로 분할하였다. 훈련 데이터 셋의 경우 199,295개의 행을 가지고 있으며, 테스트 데이터 셋의 경우 85,412개의 행을 가지고 있다. 테스트 데이터 셋과 훈련 데이터 셋을 분할함으로써 본 연구의 분류 모델이 훈련 데이터 셋에만 국한된 성능을 보이지 않도록 하였다.



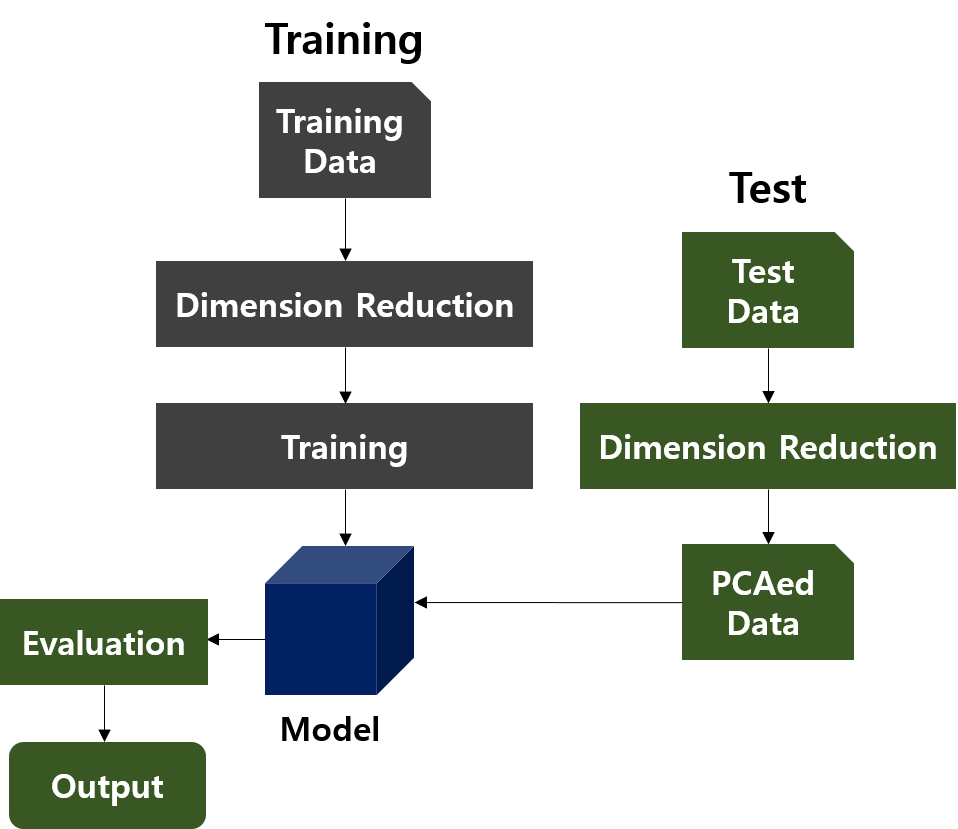
[표 2] Training dataset negative/positive class statistics.



[표 3] Test dataset negative/positive class statistics.

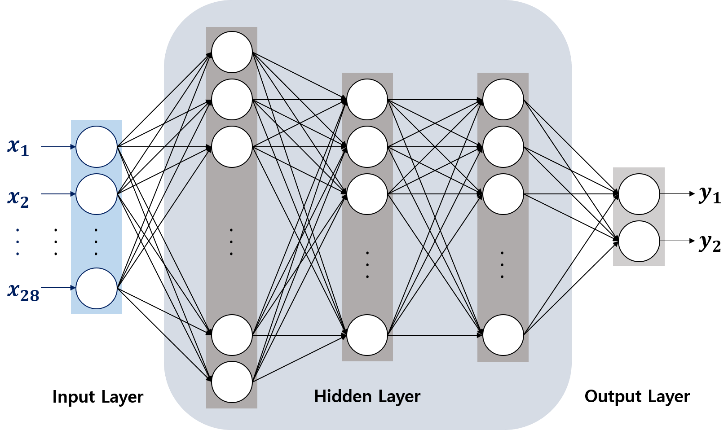
3.2. DNN

본 연구에서 제안하는 DNN을 이용한 이상 거래 탐지 모델의 Flow Chart는 [그림 1]과 같다.



[그림 1] Flowchart of proposed system.

본 연구에 사용된 DNN 모델은 입력층(Input Layer)과 출력층(Output Layer), 그리고 3개의 은닉층(Hidden Layer)으로 구성되어 있으며, 각각의 은닉층마다 112개, 56개, 56개의 Perceptron이 존재한다.



[그림 2] Layout of Deep Neural Network.

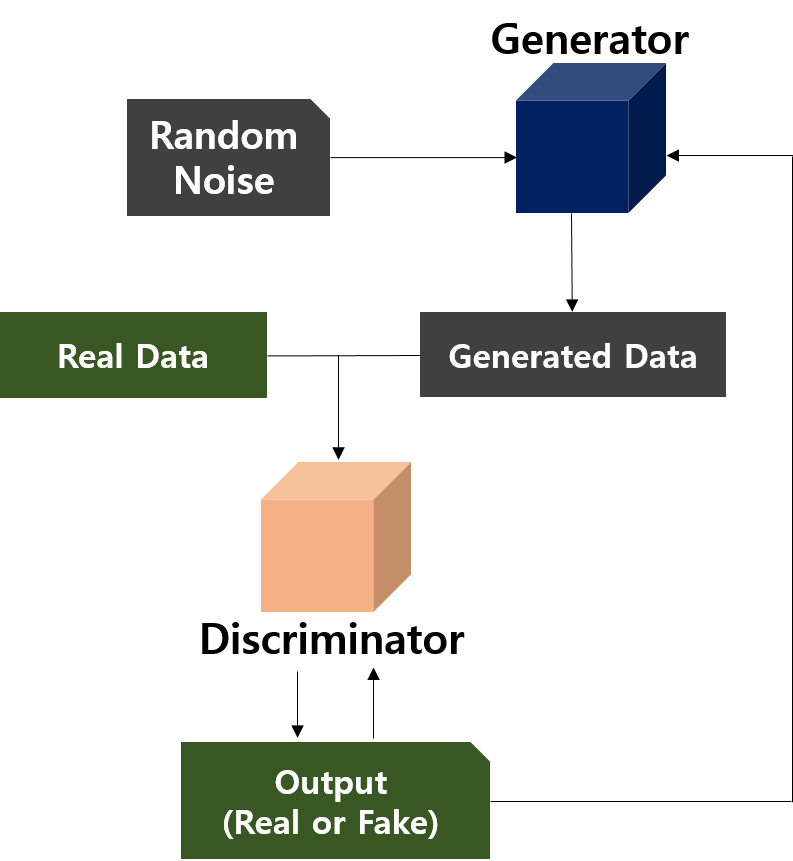
각 층에서 학습된 결과가 비용함수를 통해 오차를 계산하게 되는데, 비용함수로는 Cross Entropy나 Squared Error Criteria를 사용하는 것이 일반적이다. 이때, Cross Entropy를 비용함수로 사용하는 것이 Squared Error Criteria를 이용하는 것보다 Local Optimum을 찾는 데 더 효과적이므로 본 모델의 비용함수는 Cross Entropy를 사용하였다 [4].

한편, 신경망을 구성하는 Perceptron의 출력값은 활성화 함수의 결과값으로 출력이 되는데, 활성화 함수로는 Sigmoid 함수를 주로 사용한다. 그러나 Sigmoid 함수를 사용할 경우, 은닉층이 깊어질수록 오차가 정확히 반영되지 않는 기울기 소실문제가 발생할 수 있다. 따라서 활성화 함수로 Sigmoid함수 대신 ReLu(Rectified Linear Unit) 함수를 사용하여 모델을 구성하였다.

본 연구에 사용한 거래 로그 데이터는 주성분 분석이 이미 되어있는 상태로, [그림 1]의 차원축소 과정이 이미 이루어졌다고 할 수 있다. 따라서 데이터를 바로 학습하여 모델을 구성할 수 있으며, 테스트 또한 데이터의 차원축소 과정 없이 입력할 수 있다.

3.3 GAN

2014년 Ian Goodfellow로부터 제안된 Generative Adversarial Networks는 실제 같은 거짓 데이터를 생성하는 ‘생성기’(Generator)와 실제 데이터와 만들어진 데이터를 더 정밀히 분류하는 ‘분류기’(Discriminator)를 적대적으로 경쟁시켜 상호 발전을 유도하는 방법이다 [6]. 대부분의 기존 기계 학습 방법론들이 분류, 패턴인식 위주의 솔루션만을 제공했지만 GAN은 분류뿐만 아니라 생성의 문제도 해결할 수 있는 가능성을 보이면서 특히 이미지 리 분야에 활발한 연구와 공헌을 보이고 있다.  
 또한, 무의미한 값(Latent vector)으로 시작해 실제 데이터에 가깝게 변형, 생성된 데이터와 실제 데이터를 분류함으로써 분류기를 지속적으로 학습시키고, 분류된 결과는 또다시 생성기를 학습시키는 방식으로 별도의 사용자 지도 없이 다양한 데이터를 유연하게 이용할 수 있는 장점이 있다.



[그림 3] Flowchart of Generative Adversarial Network.

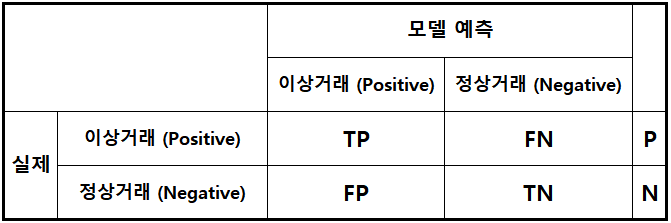
3.2.1. 이상거래탐지에 적용

공격 데이터와 정상 데이터로 구성된 이진클래스의 데이터를 분류하는 것이 이 연구의 최종 목적이므로 GAN의 적용점은 생성보다 분류에 초점이 맞춰져 있다. 준비된 실물 정상 데이터(비공격 데이터)와 생성된 가공 데이터로 분류기를 학습시키는 컨셉으로 분류기는 학습 이후 테스트 단계에 이르러서 입력 받는 데이터가 정상 데이터인지 비정상데이터(공격데이터)인지 분류하게 된다. 따라서 GAN기반 실험에서는 동일한 데이터셋에서 정상거래만을 학습시키고, 테스트 단계는 동일하게 진행했다.

4. 성능 평가

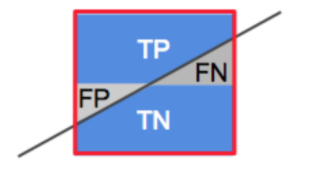
분류 모델에 대한 모델 평가는 주로 혼동 행렬(Confusion Matrix)를 이용한다. 혼동 행렬의 구성은 [표 4]과 같다.

는 실제로 이상거래인 데이터가 이상거래로 분류된 경우, 는 실제로 이상거래인 데이터가 정상거래로 분류된 경우, 은 실제로 정상거래인 데이터가 이상거래로 분류된 경우, 은 실제로 정상거래인 데이터가 정상거래로 분류된 경우를 의미한다.



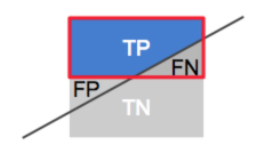
[표 4] Confusion Matrix.

혼동 행렬을 이용해 분류 모델의 성능을 평가할 때, 대표적으로 사용되는 지표는 정확도(Accuracy)로, 모델이 얼마나 정확하게 분류를 하는가를 나타낸다.



[그림 4] Accuracy.

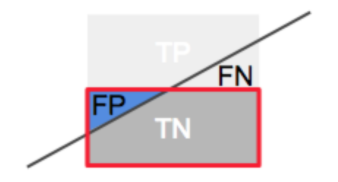
그러나 본 연구에 사용된 데이터의 경우 정상거래가 99%를 넘는 비율로 차지하고 있어 데이터가 매우 불균형(Imbalance) 한 상태이다. 때문에 정확도를 계산하는 것만으로는 분류 모델의 성능을 평가하기에는 다소 부족하다. 특히 제안하는 분류 모델의 궁극적인 목표 중 하나는 이상거래를 최대한 많이 찾아내는 것이기 때문에, 정상거래를 정상거래로 분류하는 것도 중요하지만, 이상거래를 이상거래로 정확하게 판별하는 것이 훨씬 더 중요하다. 이때 사용할 수 있는 지표로는 민감도(TPR or Sensitivity)가 있다.



[그림 5] Sensitivity.

민감도란 분류 모델의 얼마나 정확하게 Positive 값을 찾아내는가를 보여주는 지표로서, 이 지표를 참조하여 본 연구의 분류 모델이 얼마나 정확하게 이상거래를 이상거래로 분류하는가를 확인할 수 있다.

한편, 본 분류 모델의 궁극적인 목표 중 다른 하나는 정상거래를 이상거래로 탐지하지 않는 것이다. 정상거래를 이상거래로 탐지할 경우 고객 충성도에 문제가 생기기 쉬운 금융업에 있어서는 [그림 6]과 같은 지표를 오탐률(FPR)로서 레퍼런스 할 수 있다.

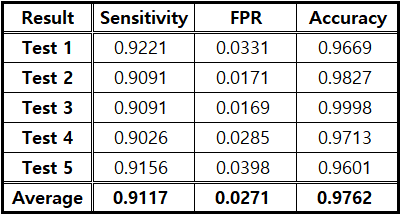


[그림 6] False Positive Rate.

5. 실험결과

5.1 DNN

DNN 모델의 연산은 행렬 곱과 합이 주를 이룬다. 때문에 많은 양의 계산이 필요하므로, GPU를 이용해 모델의 학습시간을 줄이는 것이 필수적이다. 본 연구에서 사용한 훈련 및 테스트 PC는 Intel Xeon CPU E5-2620 v4 @ 2.10GHz, 128GB RAM, 그리고 NVIDIA Tesla P100 16GB를 사용하였다.



[표 5] Test results of the Deep Neural Network.

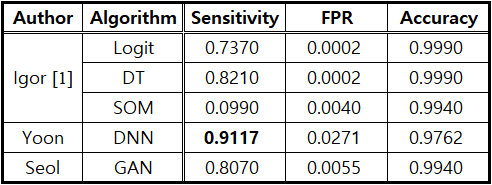
[표 5]는 DNN 모델로 테스트한 결과이다. Learning rate를 0.08, Step을 80으로 주었을 때 가장 좋은 결과값을 도출할 수 있었다.

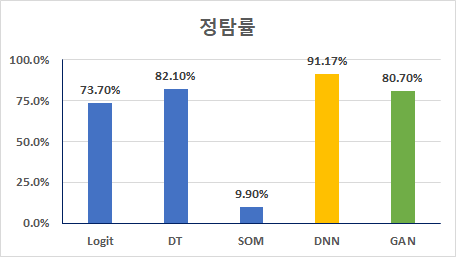
5.2 GAN

DNN을 실험한 동일한 환경에서 진행하였다. 이번 연구에서는 GAN 알고리즘을 이용한 최적의 결과 도출보다 적용가능성에 중점을 두었기 때문에 생성기와 분류기 각각 1개의 은닉 계층을 가지는 기본적인 모델을 구성하였다. 정상 데이터를 1, 공격 데이터를 0이라 하였을 때, 구별기는 실제 데이터와 1값 , 만들어진 데이터와 0값 사이의 loss를 정의하고, 마찬가지로 생성기는 생성된 데이터와 1사이의 loss를 정의했다. GAN 알고리즘의 원문에서는 각 output의 log값을 취한 비용함수를 적용했지만, 본 논문에서 쓰인 데이터셋의 특성상 loss가 무한으로 발산하는 문제가 있어 단순합을 적용하였다

5.3 기존 머신러닝 알고리즘과의 비교

[표 6]은 기존 머신 러닝 알고리즘과 본 연구의 성능을 비교한 표이고, [그림 7]은 [표 6] 중 민감도를 비교해 시각화 한 그래프이다.

[표 6] Mean classification results from each model.



[그림 7] Graph representing sensitivities of each model.

Igor Anohhin의 연구에 의하면 기존의 머신 러닝 알고리즘을 기반으로 한 이상거래 탐지 모델은 불균형 데이터로 인한 높은 정확도 대비 낮은 정탐률을 보여주고 있다.

반면, DNN을 이용한 모델은 평균 91%가 넘는 정탐률을 보여주어 기존의 머신 러닝 알고리즘 기반 모델보다 약 10% 정도 더 높은 탐지율을 도출해 내었다. GAN을 이용한 모델의 경우 정탐률이 평균 80% 내외로 비록 기존의 의사결정 트리보다 조금 낮은 점수를 얻었으나, 단순 적용 확인을 위한 모델이었다는 점에서 후에 여러 최적화를 통한 발전 가능성을 확인할 수 있다.

6. 결론 및 향후 연구과제

금융 이상거래 탐지 시스템이 대부분 정상과 비정상으로 이루어진 이진 클래스 데이터를 분류하는 기능을 수행하기 때문에 일반적인 기계학습 알고리즘과 데이터 마이닝 방법을 통한 이상거래 탐지 모델이 주를 이루었다. 또한, 기존의 인공신경망(Artificial Neural Network) 알고리즘이 가진 과적합(Overfitting)문제와 높은 시간 복잡도(Time Complexity) 등으로 인해 빠른 시간안에 분류를 해야 하는 이상거래 탐지 모델에 있어서 딥 러닝은 적용하기 힘든 알고리즘이었다. 그러나 인공지능 분야의 지속적인 연구와 발전으로 딥 러닝 알고리즘의 대표적인 문제점들이 해결 가능해지면서, 머신 러닝 알고리즘이 주가 되었던 이상거래 탐지 모델에도 딥 러닝 알고리즘이 대체 적용될 수 있게 되었다.

이번 논문을 통해 DNN 기반의 새로운 금융 이상거래 탐지 모델을 수립하여 기존 모델이 가진 불균형한 데이터셋으로 인한 탐지율 한계를 개선하였고, GAN 기반 모델을 수립하여 비지도 학습의 이진 분류문제 적용 가능성을 제시하였다. 신용카드 거래 이력 데이터셋의 두 가지 특징을 고려한 이번 연구는 금융업계에 딥 러닝 기술이 금융 이용자 보호를 위하여 더욱 의미 있게 활용될 수 있다는 점을 보여주었다.

학습 방법에 정형화된 룰이 없기 때문에 GAN기반 모델의 결과 수렴이 불안정적이지만 Latent vector가 더 유의미한 값을 가질 수 있도록 하는 방법인 [7], [8], [9], 그리고 Auto-encoder [11]로 분류기를 구현하여 비용함수에 정•오탐 가중치를 적용한 방법인[10] 등, 두 모델이 경쟁하여 서로를 학습시키는 GAN의 핵심 메커니즘을 유지하며 네트워크를 정교하게 만드는 연구가 활발히 진행되고 있다. 특히, AnoGAN[9]은 일반적인 데이터의 분포에서 벗어나는 값을 찾아내는 방법으로 이번 이상거래 탐지시스템 연구의 향후 연구방향으로 적합하다.

또한, DNN 알고리즘 기반 모델의 정확도(Accuracy)가 기존 머신 러닝 알고리즘 기반 모델의 경우보다 낮다는 점, 즉 정상거래를 이상거래로 분류하는 경우가 조금 더 많은 문제점은 본 연구에서 제안된 모델의 하이퍼 파라미터 변경을 통해 정확도를 개선하거나, 불균형 데이터 셋의 오버 샘플링(Over Sampling) 이나 언더 샘플링(Under Sampling) 이후 테스트를 진행하는 등의 개선 방안이 필요할 것으로 보인다.

4. 참고 문헌

[1] Igor Anohhin, DATA MINING AND MACHINE LEARNING FOR FRAUD DETECTION, TALLINN UNIVERSITY OF TECHNOLOGY, 2017

[2] Lusis, A Comparison of Machine Learning Techniques for Credit Card Fraud Detection, Lusis Payments, April 20, 2017

[3] Raghavendra Patidar, Lokesh Sharma, Credit Card Fraud Detection Using Neural Network, International Journal of Soft Computing and Engineering (IJSCE), June 2011

[4] Pavel Golik, Patrick Doetsch, Hermann Ney, Cross-Entropy vs. Squared Error Training: a Theoretical and Experimental Comparison, August, 2013

[5] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair‡ , Aaron Courville, Yoshua Bengio, Generative Adversarial Network, In Proceedings of NIPS, pages 2672– 2680, 2014

[7] Chen, Xi, et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets." *Advances in Neural Information Processing Systems*. 2016.

[8] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).

[9] Schlegl, Thomas, et al. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery." International Conference on Information Processing in Medical Imaging. Springer, Cham, 2017.

[10] Zhao, Junbo, Michael Mathieu, and Yann LeCun. "Energy-based generative adversarial network." *arXiv preprint arXiv:1609.03126*(2016).

[11] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

4. 기타

이 글은 JCCI 논문 제출 양식에 따라서 작성되었으니 참조하십시오