J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

Tabular Data 학습을 위한 강화형 생성자 GAN Mode

Reinforced Generator GAN Model for Tabular Data Learning

성 찬 식[1] 임 준 식[2*]
Chan-sik Sung Joon-sik Lim

요 약

Tabular Data는 수치형과 범주형 데이터의 혼합 데이터로, 이러한 Tabular Data를 이용한 학습을 수행함에 있어, 주로 머신러닝 모델이 생성형 모델보다 그 동안 적합하다고 평가되어 왔다. 이러한 평가는 생성형 모델이 Tabular Data의 특성인 수치형의 다봉분포와 범주형의 빈도 불균형 때문에 과도하게 매개변수가 많아지거나 학습의 방향을 찾지 못하는 문제가 있었기 때문이다. 그러나 데이터가 점차 빅데이터화 되고 실시간으로 이루어 지면서 기존의 머신러닝 모델들은 그 적용에 한계를 보여 왔다. 본 논문에서는 Tabular Data에 생성형 모델을 적용하기 위한 방법론으로, 켤레사전분포를 이용한 군집화 샘플링과 가워계수와 상호 정보량으로 손실함수를 개선한 생성자 강화형 적대적 신경망인 **RGGAN(Reinforced Generator GAN)**을 제안한다. 본 논문이 제안한 RGGAN으로 학습한 판별자들로 이상탐지기를 구성하여, IEEE-CIS Fraud Detection Dataset에서의 사기거래를 탐지하여 AUC를 측정해본 결과, 기존 생성형 모델들 보다 1~7%의 성능 개선 효과를 보임으써, 제안된 모델이 Tabular Data 학습에 유효하고 또한 사기거래 탐지에 효과적인 모델임을 증명하였다.

☞ 주제어 : 강화형 생성자, 가워 계수, 상호 정보, 클러스터링 샘플링, 표 형식의 데이터 학습

ABSTRACT

Tabular Data is a mixture of numerical and categorical data, and machine learning models have been evaluated to be more suitable than generative models in performing learning using such tabular data. This evaluation is because the generative model had a problem of excessively increasing parameters or not finding the direction of learning due to the numerical multimodal distribution and categorical frequency imbalance, which are characteristics of Tabular Data. However, as data gradually becomes big data and becomes real-time, existing machine learning models have shown limitations in their application. In this paper, as a methodology for applying generative models to tabular data, we propose RGGAN (Reinforced Generator GAN), a reinforced generator adversarial neural network that Clustering sampling that leverages conjugate prior distributions and the loss function improved with Gower coefficients and mutual information. As a result of measuring the AUC by detecting fraudulent transactions in the IEEE-CIS Fraud Detection Dataset by constructing an anomaly detector with the discriminators learned from the RGGAN proposed in this paper, it showed a performance improvement effect of 1-7% over the existing generative models, proving that the proposed model is effective for learning tabular data and also effective in detecting fraudulent transactions.

☞ keyword : Reinforced Generator, Gower coefficients, Mutual information, Clustering sampling, Tabular Data Learning

1. 서 론

이미지, 텍스트, 오디오 등의 비정형 데이터 처리에 있어 생성형 모델들은 그동안 중요한 성과들을 보여왔다. 그러나 Tabular Data 처리에 서만큼은 그동안 낮은 성능을 보여왔다 [1]. Tabular Data는 추천 시스템, 온라인 광고, 포트폴리오 최적화와 같은 많은 실제 응용 분야에서 쓰이는 가장 일반적인 데이터 유형이라고 할 수 있다 [2]. Tabular Data는 수치형 데이터와 범주형 데이터의 혼합 데이터 형태를 가지며, 샘플을 나타내는 행과 이종 특성을 나타내는 열로 구성된 구조를 가진다 [1].

생성형 모델들이 이러한 Tabular Data에서 그동안 낮은 성능을 보여온 이유는 Tabular Data가 가지는 수치형 데이터의 다봉분포와 범주형 데이터의 빈도 불균형한 데이터 특성 때문이었다 [3]. 이러한 이유로 머신러닝 모델들이 Tabular Data 처리에 있어 지배적인 모델[4]로 자리 잡아왔다. 그러나 데이터가 점차 빅데이터화 되고 실시간으로 발생 되면서, 피쳐 중요도의 잦은 변경과 계산 복잡성의 증대로 인하여, 기존 머신러닝 모델들은 그 적용에 한계

1 Department of IT Convergence, Gachon University, Seongnam-Si, 13120, Korea.
2 Department of Computer Engineering, Gachon University, Seongnam-Si, 13120, Korea.
* Corresponding author (jslim@gachon.ac.kr)
[Received 3 September 2024, Reviewed 25 September 2024, Accepted 29 September 2024]

를 보여왔다. 머신러닝 모델이 일정 정도의 성능을 보임에도 불구하고 Tabular Data에 생성형 모델을 적용해야 하는 이유는 다음과 같다.

- 대규모 데이터셋의 처리가 가능하다.
- 피쳐 엔지니어링의 필요성을 경감시킨다.
- 실시간 거래에 적합한 지속적 학습이 가능하다.
- 다양한 도메인 데이터 적용이 가능하다.
- 복잡하고 다양한 패턴 학습에 효율적이다.

이러한 다섯 가지 이유로 Tabular Data에 생성형 모델을 적용하고자 하는 여러 연구와 제안들이 활발하게 이루어져 왔다 [4].

TGAN [5], CTGAN [6], TabNet [1], TabTrans-former [2] 등의 Tabular Data 처리를 위한 생성형 모델들은 성능 개선이라는 공통의 성과가 있었지만 이에 대한 반대급부로 모델 복잡도와 계산 비용 증가라는 공통의 문제 또한 안고 있었다. 모델 복잡도 증가는 필연적으로 과적합을 발생시키고 계산 비용 증가는 훈련 시간을 증대시켜, 실시간 대용량으로 발생하는 애플리케이션에서의 실제적 적용을 어렵게 하는 근본적 문제를 발생시켰다.

본 연구에서는 생성형 모델 적용 시 Tabular Data의 특성으로 발생하는 과도한 매개변수 증가와 올바른 학습의 방향을 찾지 못하는 전통적인 문제 외에도, 다양한 생성형 모델에서 고르게 나타났던 모델 복잡성과 계산 비용 증가 문제를 해결하기 위한 개선 방안을 다음과 같이 제안한다.

- Tabular Data가 가지는 수치형 데이터의 다봉분포와 범주형 데이터의 빈도 불균형한 다항분포를 고르게 학습시키는 샘플러를 가진 모델
- Tabular Data의 패턴 학습을 위해, 별도의 특성 추출 또는 특성 계산 알고리즘 없이 손실함수만으로 패턴 학습이 가능한 모델

본 논문은 이러한 목표에 부합하는 모델로 RGGAN (Reinforced Generator GAN)을 제안한다. 본 논문이 제안한 모델로 학습한 판별자들로, 이상 탐지기를 구성하여, IEEE-CIS Fraud Detection Dataset에서의 사기 거래를 탐지하여 AUC를 측정해본 결과, 기존 생성형 모델보다 1~7%의 성능 개선 효과를 보임으로써, RGGAN 모델이 제안한 학습 방법론이 유효하며 또한 신용카드 사기 거래 탐지에도 효과적인 모델임을 증명하였다.

2. 관련 연구

2.1 Tabular Data의 데이터 전처리

데이터 전처리는 학습 모델의 성능을 향상시키는 중요한 과정으로, 일반적으로 수치형 데이터는 스케일링을 통해 표준화와 정규화를 거치고, 범주형 데이터는 레이블 인코딩을 통해 수치화하거나 원-핫 인코딩으로 이산화 전처리를 수행한다 [7].

Tabular Data의 경우, 수치형과 범주형 데이터를 각각 분리하여 전처리를 수행하는데 [8], TGAN [5]은 최소-최대 스케일링을 적용하여 수치형 데이터를 [0,1] 범위로 정규화하였고 범주형 데이터는 원-핫 인코딩을 사용하여, 모델 학습을 용이하게 하였다. 반면, CTGAN [6]은 데이터의 분포를 더 잘 포착하기 위해서, 가우스 혼합모델을 사용하여, 수치형 데이터를 정규화하였고 범주형 데이터의 경우는, 각 범주의 확률을 추정한 원-핫 인코딩된 값으로, 조건부 벡터를 만들어, 조건부 생성 학습이 가능하게 하였다.

이러한 범주형 데이터의 원-핫 인코딩은 범주가 상호 배타적이라는 가정 아래서 이루어지는데, 이러한 가정은 편향된 확률 추정으로 이어질 수 있다. 특히 여러 범주가 동시에, 관계하는 Tabular Data의 경우 편향된 확률 추정의 문제는 필연적으로 과적합 문제를 발생시킬 수 있다.

2.2 Tabular Data의 패턴 학습

Tabular Data에서, 범주형 데이터는 빈도 불균형한 다항분포의 형태를 가진다. 이러한 다항분포는 여러 범주형 변수가 있을 때, 각 범주의 여러 클래스 중 하나 클래스가 선택될 확률로 나타내어진다 [9].

CGAN [10]은 이러한 다항분포를 학습하기 위해서 조건부 벡터를 활용한다. 조건부 벡터는, 범주형 데이터의 원-핫 인코딩된 값으로, 생성자는 조건부 벡터를 생성하고, 판별자는 조건 정보를 받아, 생성된 데이터가 조건을 만족하는지를 판별한다. CGAN은 이러한 조건부 생성을 통해, 생성자가 각 범주의 여러 클래스가 선택될 확률을 학습하게 하여, 이를 기반으로 데이터를 생성하게 된다.

CTGAN은 기존 CGAN을, 범주를 매개로, 수치형 데이터의 복잡한 분포를 처리할 수 있도록, 개선한 버전이라고 할 수 있다. 조건부 벡터는, 범주형 데이터의 이산화된 값과 수치형 데이터의 다봉분포 중 그 범주에 해당하는 정규분포와 그 분포의 구간으로 구성된다. 조건부 벡터를

입력받는 생성자는 해당 조건을 만족하는 데이터를 생성하고, 판별자는 생성된 데이터가 주어진 조건을 만족하는지를 판별하여, 범주와 해당 범주와 관계있는 수치형 데이터의 여러 분포를 학습하게 된다.

이러한 조건부 벡터는 학습 데이터의 피쳐 간의 종속성을 캡처하기 위해 사용된다. 그러나 Tabular Data와 같이 고차원 데이터이면서 비선형 관계를 가진 데이터를 처리하는 경우, 피쳐 간 종속성을 제대로 캡처하지 못할 가능성이 크다. 조건부 벡터가 피쳐 간 종속성을 제대로 캡처하지 못할 경우, 데이터의 모든 분포를 완전하게 포착할 수 없게 되므로 필연적으로 과적합 문제가 발생하게 된다.

CGAN이 생성자의 조건부 벡터를 이용하는 방식이라면 SGAN [11]는 판별자가 생성자가 생성한 데이터가 특정 범주의 어느 클래스에 해당할지를, 클래스에 대한 확률 벡터로 출력하는 방식을 사용한다. 각 확률은 해당 데이터가 그 클래스에 속할 가능성을 나타내며, SGAN은 판별자가 각 범주의 여러 클래스의 레이블을 예측하게 함으로써, 판별자가 다항분포를 학습하게 한다. SGAN은 제한된 특정 범주의 클래스만을 생성하도록 특화된 모델로, Tabular Data와 같이 피쳐 간의 복잡한 상관관계와 종속성이 포함되는 데이터의 경우, SGAN은 이러한 피쳐 간 관계를 정확하게 포착하지 못할 가능성이 크다. 전체 데이터의 분포를 포괄하지 못하는 이러한 문제는 필연적으로 과적합 문제로 이어지게 된다.

TabTransformer는 Tabular Data를 수치형과 범주형 변수로 분리하여, 수치형 변수는 정규화를 거쳐 최종 레이어로 바로 투입하고 범주형 변수는 컬럼임베딩 과정을 거쳐 Transformer 레이어를 통과한 후에 최종 레이어로 투입되게 한다. 컬럼임베딩은 전체 데이터 세트 내에서 다양한 피쳐 간 관계를 맥락화하여 범주형 데이터를 인코딩하는 알고리즘으로, 컬럼임베딩으로 생성된 벡터는 Transformer의 어텐션 매커니즘 레이어들을 통과하여 최종 레이어로 투입되기 전 정규화된 수치형 변수와 합쳐지게 된다. TabTransformer는 정규화된 수치형 변수와 맥락화된 범주형 변수를 최종 연결하여, 데이터 세트 내의 다양한 피쳐 간 패턴을 학습함으로써, 피쳐 간 상호 관계를 맥락 있게 분석하여, 더 정확하고, 해석 가능한 예측을 할 수 있게 되었다.

TabNet는 Tabular Data에 강점을 지닌 트리 기반 모델의 변수 선택 특징을 딥러닝 구조에 반영한 모델로써, 순차적 어텐션 메커니즘을 사용한다. 각 결정 단계마다 어텐션 스코어로, 어떤 피쳐를 사용할지를 선택하게 하여, Tabular Data의 학습을 보다 효율적으로 수행하고 이를 통하여 각 단계를 설명할 수 있게 한다. 또한 마스킹 처리된 인코딩 데이터를 원본대로 복원하는 자기 지도 학습을 병행하여 예측 성능을 크게 향상시켰다.

TabTransformer과 TabNet 모델은 계산 비용이 많이 드는 Transformer 아키텍처의 어텐션 메커니즘를 기반으로 구축되었다. 이러한 높은 계산 비용은 매우 큰 데이터 또는 계산 리소스가 제한된 인프라 환경일 경우 실용성이 현저히 떨어질 수 있다.

2.3 Tabular Data의 특성 추출 및 계산

Tabular Data는 수치형의 다봉분포와 범주형의 빈도 불균형한 다항분포라는, 서로 상이하고 이질적인 분포의 결합으로, 이러한 피쳐 간의 복잡한 상호 작용이 데이터의 특성으로 포함되어 있다. 이러한 상호 작용을 식별하고, 측정하여, 모델에 반영시키기 위해서 생성형 모델들은 조건부 벡터를 사용하거나 어텐션 메커니즘을 사용하여 상호 작용을 캡쳐 하고자 하였다. 조건부 벡터는 피쳐 간 관계를 제대로 파악하지 못하는 경우 과적합 문제를 발생시켰고 어텐션 메커니즘는 높은 계산 비용으로 실제적 적용의 어려움을 발생시켰다.

Tabular Data의 특성 추출 및 계산을 위한 올바른 목적함수의 채택은 모델의 성능과 직결되는 중요한 문제라 할 수 있다 [12]. 가워계수와 상호정보량은 Tabular Data의 피쳐 간 관계를 포착하고 측정하는 대표적인 지표들로, 조건부 벡터나 어텐션 메커니즘의 훌륭한 대안이 될 수 있다.

가워계수 [13]는 서로 다른 유형의 변수를 포함한 데이터 세트 간의 유사성을 측정하는 방법으로, 주로 수치형과 범주형 데이터가 혼합된 경우에 사용되었다. 가워계수는 각 피쳐 별로 부분 유사도를 계산하고 이를 평균 내는 방식으로 전체 유사도를 구하는데, 수치형 변수는 최댓값과 최솟값으로 정규화한 후 거리를 구하고, 범주형 변수는 일치 여부에 따라 유사도를 결정하였다. 가워계수는 주로 수치형과 범주형이 혼합된 데이터 유형의 군집화 문제에서 불일치를 측정하는 목적함수 [14]로 그동안 주로 사용되어왔다. 이 계수를 통하여 클러스터 간의 유사성을 평가하여, 군집화 성능을 크게 향상시켜, 그 성능을 입증하였다.

상호정보량(Mutual Information) [15]은 두 확률변수 간의 상호 의존성을 측정하는 방법으로, 한 변수의 값을 알았을 때 다른 변수에 대한 불확실성이 얼마나 줄어드는지를 나타낸다. 그동안 상호정보량은 기계 학습에서 중요한

특징을 선택하거나, 텍스트 마이닝에서 변수 간의 상호 의존성을 측정하거나, 이미지 처리에서는 유사성 또는 변환 관계를 평가하는 데 사용되어왔다. InfoGAN [16]는 이러한 상호정보량을 특정 특징을 선택하는 목적함수로 사용한 모델로, 생성된 데이터와 잠재 코드 간의 상호정보량을 최대화하는 것을 핵심 아이디어로 하여, 변분 하한(variational lower bound)을 사용하여, 상호정보량을 근사하고, 변분 하한을 최대화함으로써, 생성된 데이터가 잠재 코드에 대한 유용한 정보를 포함하도록 만들어, 이를 통해 InfoGAN이 직접 특정 특징을 제어할 수 있게 함으로써, 해석 가능한 데이터를 생성할 수 있게 되었다.

3. 방법론

3.1 RGGAN의 구성

본 논문에서는 RGGAN 모델의 구성 방안을 Figure 1과 같이, 다음과 같이 제안한다.

- 데이터를 군집의 그룹별로 균등하게 추출하여 학습의 방향을 제시하는 샘플러 **(3.1-A)**
- 가워계수를 손실함수로 사용하여, 조건부 생성으로 수치형과 범주형 가짜 데이터를 생성하는 생성자 **(3.1-B)**
- 상호정보량을 손실함수로 사용하여, 샘플링 데이터가 속한 군집의 다봉분포와 다항분포를 각각 학습하는 수치형과 범주형 판별자로 이루어진 이중 판별망 **(3.1-C)**

학습을 마친 두 개의 판별자로 앙상블 된 이상 탐지기 모델 레이어는 다음과 같이 구성한다.

- 생성자는 LeakyReLU, BatchNormalization, 활성화 함수 tanh를 사용하여 모델 레이어를 구성한다.
- 범주형 판별자는 LeakyReLU, BatchNormalizat-ion, Dropout 및 활성화 함수로 한 개의 Sigmode와 다수의 Softmax를 사용하여 모델 레이어를 구성한다.
- 수치형 판별자는 LeakyReLU, BatchNormalization, Dropout 및 활성화 함수로 Sigmode와 Softmax를 사용하여 모델 레이어를 구성한다.

3.2 데이터 전처리와 샘플러 구현

데이터 전처리 전에 매개 변수의 증가와 과적합을 방지하기 위해서 단변량 테스트를 통한 피쳐 자동 선택을, Figure 2와 같이, 다음과 같이 수행한다.

- 수치형 데이터의 각 피쳐에 ANOVA F-value, 상호정보량, SelectKBest, SelectPercentile의 네 가지 단변량 테스트를 수행하여 순위별 점수를 부여한다. **(3.2-A)**
- 범주형 데이터의 각 피쳐에 ANOVA F-value, 상호정보량, 카이제곱, SelectKBest, SelectPercentile으로 다섯 가지 단변량 테스트를 수행하여 순위별 점수를 부여한다. **(3.2-B)**
- 최종적으로 종합 점수를 채점하여, 상위 20개의 피쳐가 자동으로 선택되도록 한다. **(3.2-C)**

군집 별 균등 샘플러를 구현하기 위하여, Figure 2와 같이, 다음과 같은 데이터 전처리 과정을 수행하여, 샘플러를 구현한다.

- 신용카드 거래 데이터를 데이터 타입에 따라 수치형과 범주형 데이터로 각각 분리한다. **(3.2-D)**
- 수치형 데이터는 결측치를 0으로 처리하고 피쳐 스케일링 정규화를 수행한다. **(3.2-E)**
- 범주형 데이터는 레이블 인코딩을 통해서 수치화한다. **(3.2-F)**
- 정규화된 수치형 데이터에 대해 디리클레 프로세스 가우시안 혼합모델 [17]로 군집화를 수행하여 군집 그룹 정보를 수치형 데이터에 추가한다. **(3.2-G)**
- 수치화된 범주형 데이터에 대하여 디리클레 프로세스 혼합모델 [18]로 군집화를 수행하여 군집 그룹 정보를 범주형 데이터에 추가한다. **(3.2-H)**
- 샘플러는 학습 모델의 배치 사이즈를 군집 그룹 수로 나누어, 나누어진 수만큼 해당 군집 그룹별로 데이터를 무작위로 추출한다. **(3.2-I)**
- 매학습마다 샘플러는 군집 그룹별로 균등한 샘플링을 수행하여, 수치형의 다봉분포와 범주형의 빈도 불균형한 다항분포를, 고르게 학습하게 하여 학습의 방향을 잃지 않게 한다. **(3.2-J)**

3.3 강화형 생성자와 가워계수

생성자가 Tabular Data의 피쳐 간 관계를 학습하여 Tabular Data를 생성하게 되는 과정은 다음과 같다.

- 생성자는 수치형 랜덤 변수와 범주형 랜덤 변수를 각각 입력받는다. **(3.3-A)**
- 입력받은 랜덤 변수로, 생성자는 Tabular Data를 생성한다. **(3.3-B)**
- 생성자는 샘플러로부터 Tabular Data를 목표 데이터로

입력받는다. **(3.3-C)**

- 다중 피드백 레이어는 샘플링된 Tabular Data와 생성자가 생성한 수치형과 Tabular Data의 피쳐 간 유사도를 가워계수로 각각 측정한다. **(3.3-D)**

$$Dij = 1 - \frac{\sum_{k=1}^{p} w_k d_{ij}^k}{\sum_{k=1}^{p} w_k} \quad (1)$$

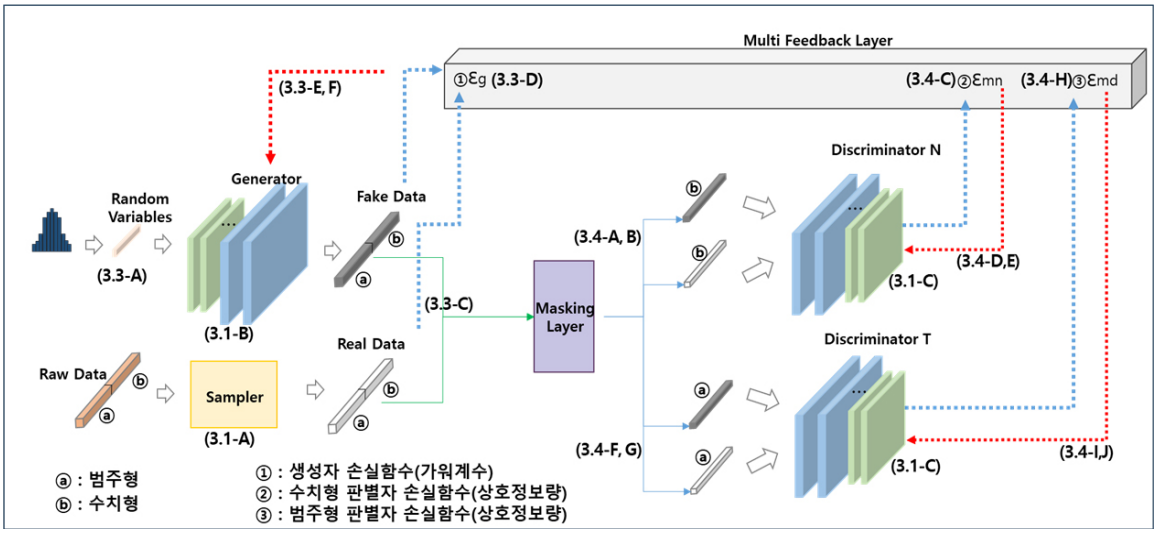
수식 (1):

d_{ij}^k는 데이터 포인트 i과 j 간의 Gower 유사성

w_k는 i과 j간의 k번째 특성에 대한 유사성 점수

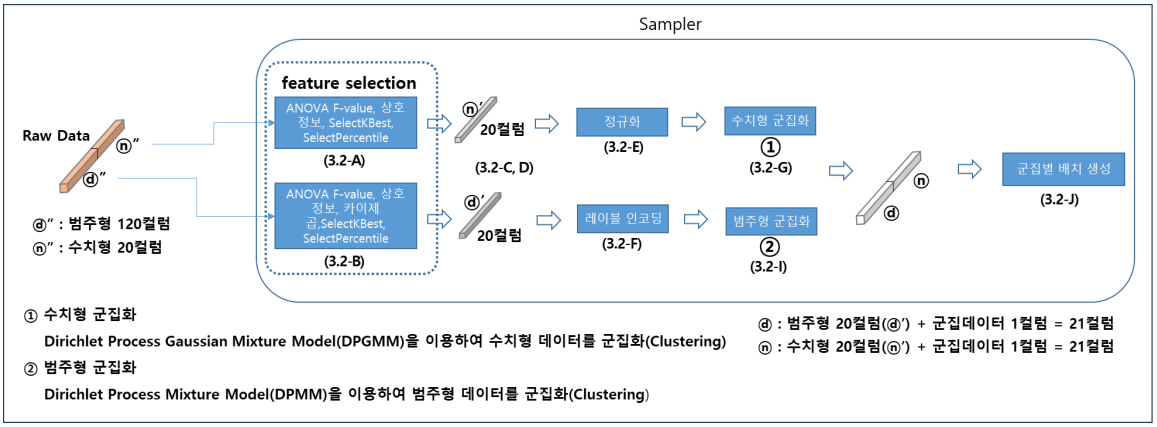
p는 모든 픽쳐 유사성의 총합

- 다중 피드백 레이어는 두 데이터의 가워계수를 비교하여, 두 데이터의 가워계수 차이가 최소가 되는 방향으로 생성자의 학습을 진행한다. **(3.3-E)**
- 생성자는 최종적으로 Tabular Data의 유사도에 기반한, 피쳐 간 관계를 학습하게 되어, 그 관계에 맞는 Tabular Data를 생성하게 된다. **(3.3-F)**



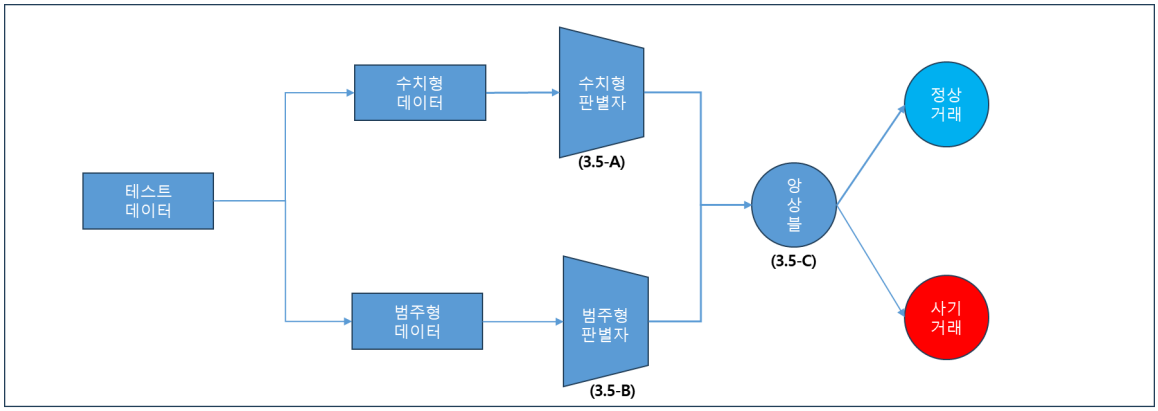
(그림 1) RGGAN의 구성

(Figure 1) Configuring RGGAN



(그림 2) 데이터 전처리 및 샘플러 구현

(Figure 2) Data preprocessing and sampler implementation



(그림 3) 이상탐지기의 구성
(Figure 3) Configuration of the anomaly detector

3.4 이중 판별망과 상호정보량

수치형 데이터의 다봉분포와 범주형 데이터의 빈도 불균형한 다항분포를 판별자의 데이터 매니폴드에서 효과적으로 학습할 수 있도록, 두 개의 판별자로 분리하여 이중 판별망을 구성한다. 이중 판별망은 다음과 같은 과정을 통하여 Tabular Data의 다봉분포와 다항분포를 학습하여 사기 거래 데이터를 판별한다.

- 수치형 데이터는 시간 축을 가지는 다변량 정규분포로써, 다봉분포의 형태를 가지는 정규분포의 결합 분포로 정의한다.
- 범주형 데이터는 카테고리 확률변수로써 빈도 불균형한 다항분포의 결합 분포로 정의한다.
- 수치형 판별자는 생성자로부터 생성된 Tabular Data를 입력받아 마스킹 레이어를 통해 수치형 데이터만 분리하여 입력 처리한다. (3.4-A)
- 수치형 판별자는 샘플러로부터 받은 Tabular Data를, 목표 데이터로 입력받아, 마스킹 레이어를 통해 수치형 데이터만 분리하여 입력 처리한다. (3.4-B)
- 다중 피드백 레이어는 두 데이터의 정규분포 간의 상호 의존성을 상호 정보로 측정한다. (3.4-C)

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \qquad (2)$$

수식 (2):
$P(x,y)$는 x,y의 결합확률분포함수
$P(x), P(y)$는 x,y의 주변부 확률분포함수
합계는 x,y의 모든 가능한 확률변수 값

- 다중 피드백 레이어는 두 데이터의 상호정보량 차이가 최소가 되는 방향으로 생성자와 판별자의 학습을 진행한다. (3.4-D)
- 최종적으로, 수치형 판별자는 결합 분포상의 정규분포 간 관계를 학습하게 되어, 그 관계에 해당하는 사기 거래 데이터를 판별하게 된다. (3.4-E)
- 범주형 판별자는 생성자로부터 생성된 Tabular Data를 입력받아, 마스킹 레이어로 범주형 데이터만 분리하여 입력 처리한다. (3.4-F)
- 범주형 판별자는 샘플러로부터 받은 Tabular Data를 목표 데이터로 입력받아, 마스킹 레이어로, 범주형 데이터만 분리하여 입력 처리한다. (3.4-G)
- 다중 피드백 레이어는 두 데이터의 다항분포 간의 상호 의존성을 상호정보량으로 측정한다. (3.4-H)
- 다중 피드백 레이어는 두 데이터의 상호정보량 차이가 최소가 되는 방향으로 생성자와 판별자의 학습을 진행한다. (3.4-I)
- 최종적으로, 범주형 판별자는, 상호 의존성의 기반한, 다항분포 간 관계를 학습하게 되어, 그 관계에 해당하는 사기 거래 데이터를 판별하게 된다. (3.4-J)

3.5 이상 탐지기의 구성

최종적으로 사기 거래 검출을 위하여 학습을 마친 수치형 판별자(3.5-A)와 범주형 판별자(3.5-B)로 Figure 3과 같이 이상 탐지기(3.5-C)를 구성한다. 두 판별자에 각각 0.6의 가중치를 주고 합산한 결괏값을 반올림하여 이상에 좀 더 민감하게 반응하도록 앙상블 처리하여 사기 거

래를 탐지하도록 한다.

4. 실 험

4.1 학습 데이터 셋

학습 데이터셋은 Vesta에서 제공한 IEEE-CIS Fraud Detection Dataset을 사용하였다. IEEE-CIS Fraud Detection Dataset은 실제 상거래에 사용된 카드 거래 데이터로, 라벨 데이터를 외의 432개의 피쳐를 가지고 있으며, 이중 범주형 데이터는 20개를 가지고 있다. 총거래 건수 590,539건 중 부정 거래가 11,318건이 존재한다 [19].

4.2 모델 학습

학습, 검증, 평가 데이터로, 학습 데이터를 6:2:2의 비율로 분할하고, 옵티마이저는 Adam [20]를 사용하여, 반복 학습을 수행하여, 학습률 0.0002, Beta1 0.5, epoch 8000회, batch size 8000을 최적의 하이퍼파라미터로 설정하였다. 데이터 불균형 문제를 처리하기 위해서 언더샘플링과 오버샘플링을 혼합하여 데이터의 불균형을 최소화하고 가중치 부여와 Focal Loss를 보조 손실함수로 사용하여 데이터 불균형 문제를 해소하였다. 성능지표로는 정밀도(Precision, PRE), 재현율(Recall, REC), 성능 효율성(F1 score, F1), 곡선 아래 면적(AUC), 네 가지를 측정하였다. 그리고 불균형 데이터의 경우, 정확도(Accuracy)는 너무 편향된 결과를 가지기 때문에 평가지표에서 제외하였다.

학습 완료 후 1차로 GAN 모델들과 성능을 비교하였다. 비교한 GAN 모델들은 다음과 같다.

- 손실함수가 없는 생성자와 손실함수가 Binary Cross Entropy인 단일 판별자를 가진 GAN 모델(G-LF/1D+BCE)
- 손실함수가 없는 생성자와 손실함수가 Binary Cross Entropy인 이중 판별자를 가진 GAN 모델(G-LF/2D+BCE)
- 손실함수가 가워계수인 생성자와 손실함수가 Binary Cross Entropy인 단일 판별자를 가진 GAN 모델(G+G/1D+BCE)
- 손실함수가 가워계수인 생성자와 손실함수가 Binary Cross Entropy인 이중 판별자를 가진 GAN 모델(G+G/2D+BCE)

2차는 샘플러가 없는 RGGAN(-SP)과 샘플러가 있는 RGGAN 성능을 비교하고 3차는 생성형 모델들인 신용카드 사기 거래 탐지 오토인코더 기반 모델(AE) [21], 신용카드 사기 거래 탐지 TabNet 기반 모델(TabNet) [22][23], 신용카드 사기 거래 탐지 GAN 기반 모델(UAAD-FDNet) [19]과 성능을 비교하였다. 비교 모델의 요약 정보는 Table 1과 같다.

(표 1) 생성형 모델 요약
(Table 1) Generative Model Summary

Model name	Base model	Calculation of characteristics
AE	Autoencoder	Denosing Compression / Restoration
TabNet	Transformers	Attention mechanism
UAAD-FDNet	Autoencoder + GAN	Attention mechanism

5. 결 과

학습을 완료 후 이상 탐지기를 구성하여 비교 모델들과 사기 거래 탐지 성능을 비교 평가한 결과는 Table 2, Table 3, Figure 4, Figure 5, Figure 6과 같다.

(표 2) GAN 모델 성능 평가 지표
(Table 2) GAN Model Performance Evaluation Indicators

Model	PR	RC	F1	AUC
G-LF/1D+BCE	0.741	0.587	0.722	0.812
G-LF/2D+BCE	0.793	0.509	0.682	0.824
G+G/1D+BCE	0.834	0.633	0.751	0.832
G+G/2D+BCE	0.876	0.653	0.781	0.849
RGGAN	**0.946**	**0.697**	**0.802**	**0.892**

G-LF/1D+BCE, G-LF/2D+BCE, G+G/1D+BCE, G+G/2D+BCE 등의 GAN 모델들과 RGGAN을 비교한 결과, Table 2와 같이 RGGAN이 비교 모델들보다 AUC에서 4~8%의 비교 우위를 보였다.

(표 3) 샘플러 유무 성능 평가 지표
(Table 3) Sampler presence or absence Performance Evaluation Indicators

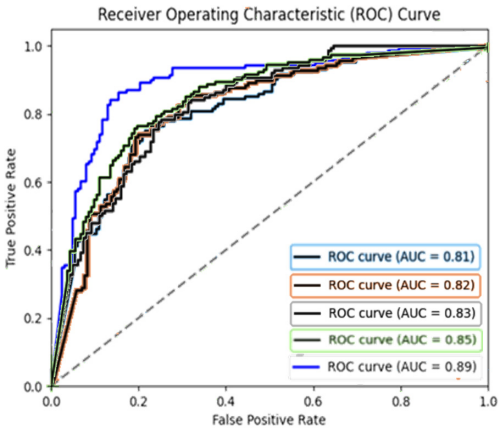
Model	PR	RC	F1	AUC
RGGAN(-SP)	0.842	0.642	0.761	0.852
RGGAN	**0.946**	**0.697**	**0.802**	**0.892**

RGGAN(-SP)과 RGGAN을 비교한 결과, Table 3과 같이 RGGAN이 RGGAN(-SP) 보다 AUC에서 4%의 비교 우위를 보였다.

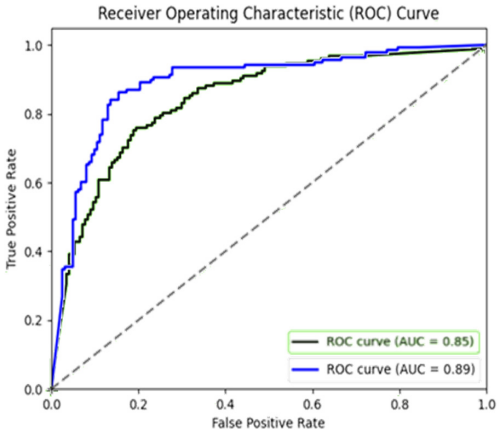
(표 4) 생성형 모델 성능 평가 지표
(Table 4) Generative Model Performance Evaluation Indicators

Model	PR	RC	F1	AUC
AE	0.941	0.587	0.712	0.818
TabNet	0.776	0.509	0.615	0.884
UAAD-FDNet	0.934	0.628	0.751	0.856
RGGAN	**0.946**	**0.697**	**0.802**	**0.892**

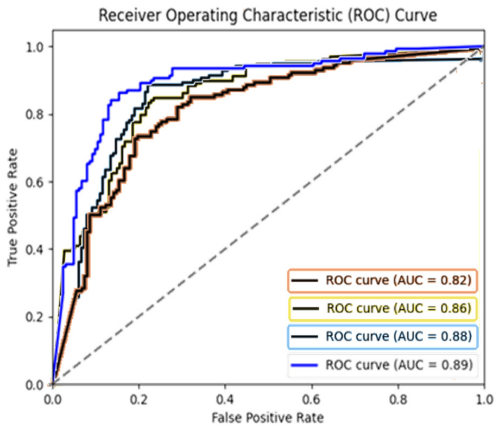
AE, TabNet, UAAD-FDNet 등의 생성형 모델들과 RGGAN 비교한 결과, Table 4와 같이 RGGAN이 비교 모델들보다 AUC에서 1~7%의 비교 우위를 보였다.



(그림 4) GAN 모델 ROC 곡선
(Figure 4) GAN Model ROC Curve



(그림 5) 샘플러 유무 ROC 곡선
(Figure 5) Sampler presence or absence ROC Curve



(그림 6) 생성형 모델 ROC 곡선
(Figure 6) Generative Model ROC Curve

G-LF/1D+BCE, G-LF/2D+BCE, G+G/1D+BCE, G+G/2D+BCE 등의 GAN 모델들과 RGGAN을 비교한 결과, Figure 4와 같이 ROC Curve 면적이 비교 모델들보다 4~8% 넓음을 확인할 수 있다. RGGAN(-SP) 과 RGGAN을 비교한 결과, Figure 5와 같이, RGGAN이 RGGAN(-SP)보다 ROC Curve 면적이 4% 넓음을 확인할 수 있다. AE, TabNet, UAAD-FDNet 등의 생성형 모델들과 RGGAN을 비교한 결과, Figure 6과 같이 RGGAN이 비교 모델들보다 AUC ROC Curve 면적이 비교 모델들보다 1~7% 넓음을 확인할 수 있다.

6. 결 론

본 논문에서는 Tabular Data인 카드거래 데이터를 효과적으로 학습하여 신용카드 사기 거래를 탐지하는 생성형 모델로서 RGGAN을 제안하였다. 1차, 2차, 3차에 걸친 비교 모델들과의 성능을 비교한 결과, 다음 사항들이 입증되었다. G-LF/1D+BCE, G-LF/2D+BCE, G+G/1D+BCE, G+G/1D+BCE와 RGGAN을 비교한 결과, AUC에서 4~8%의 비교 우위를 보임으로써, 가워계수와 상호정보량의 손실함수 적용과 이중 판별망 구성이 Tabular Data 학습에 효과적임을 입증하였다. 샘플러가 없는 RGGAN(-SP)과 샘플러가 있는 RGGAN의 성능을 비교한 결과, 샘플러가 있는 RGGAN이 AUC에서 4%의 비교 우위를 보임으로써, RGGAN의 샘플러가 Tabular Data의 올바른 학습 방향을 제시함이 입증되었다. 최종적으로 생성형 모델들인 AE, TabNet, UAAD-FDNet과 RGGAN을 비교한 결과, RGGAN

이 비교 모델들 보다, AUC에서 1~7%의 비교 우위를 보임으로써, Tabular Data의 패턴 학습을 위해, 별도의 특성 추출 또는 특성 계산 알고리즘 없이도, 가워계수와 상호정보량의 손실함수만으로도 Tabular Data의 패턴 학습에 효과적임을 입증하였다.

본 연구에서는 가워계수와 상호정보량의 손실함수 적용이 Tabular Data의 학습에 효과적임을 실험으로 입증하였다. 그러나 이를 수학적 귀납법으로 증명해 내지는 못하였다. 향후 이를 수학적으로 증명하고, Tabular Data 학습에 적합한 방식으로 수식을 개선한다면, 더 큰 학습효과를 효과를 얻을 수 있을 것이다. 또한 군집 별 균등 샘플링을 통하여 학습 데이터를 제공하는 샘플러가 Tabular Data의 올바른 학습 방향을 제시함을 실험으로 입증하였다. 그러나 이 역시 이를 수학적 귀납법으로 증명해 내지는 못하였다. 향후 이를 수학적으로 증명하고 학습 단계별로 피쳐 중요도를 반영하여 샘플링 알고리즘을 개선한다면, 더 효과적으로 올바른 학습 방향을 제시할 수 있을 것이다. 끝으로 이중 판별망 구성이 단일 판별자 구성보다. Tabular Data 판별에 효과적임을 실험으로 입증하였으나 이 역시 이를 수학적 귀납법으로 증명해 내지는 못하였다. 향후 이를 수학적으로 증명하고 이중 판별망 모델을 정교화한다면 사기 거래 판별의 정확성을 보다 개선할 수 있을 것이다.

참고문헌(Reference)

[1] Kyungeun Lee, Ye Seul Sim, HyeSeung Cho, Moon jung Eo, Suhee Yoon, Sanghyu Yoon, Woohyung Lim, “Binning as a Pretext Task: Improving Self-Supervised Learning in Tabular Domains,” NeurIPS, 2023. https://doi.org/10.48550/arXiv.2405.07414

[2] Xin Huang, Ashish Khetan, Milan Cvitkovic, Zohar Karnin, “TabTransformer: Tabular Data Modeling Using Contextual Embeddings,” NeurIPS, 2020. https://doi.org/10.48550/arXiv.2012.06678

[3] Lei Xu, Kalyan Veeramachaneni, “Synthesizing Tabular Data using Generative Adversarial Networks,” arXiv 27 November 2018 Computer Science, 2018. https://doi.org/10.48550/arXiv.1811.11264

[4] Sercan O. Arık, Tomas Pfister, “TabNet: Attentive Interpretable Tabular Learning,” Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 8, pp. 6679-6687, 2021. https://doi.org/10.1609/aaai.v35i8.16826

[5] Lei Xu, Kalyan Veeramachaneni, “Synthesizing Tabular Data using Generative Adversarial Networks,” arXiv 27 November 2018 Computer Sience, 2018. https://doi.org/10.48550/arXiv.1811.11264

[6] Lei Xu, Maria Skoularidou, Alfredo CuestaInfante, Kalyan Veeramachaneni, “Modeling Tabular Data using Conditional GAN,” Advances in Neural Information Processing Systems 32, 2019. https://doi.org/10.48550/arXiv.1907.00503

[7] HL Nakayiza, LAC Ahakonye, DS Kim, JM Lee, “Machine Learning Algorithms for Detecting Intra-Vehicular Data Falsification,” ResearchGate, 2024. https://www.researchgate.net/publication/382330545_Machine_Learning_Algorithms_for_Detecting_Intra-Vehicular_Data_Falsification

[8] V Borisov, Leemann, K Seßler, J Haug, “Deep Neural Networks and Tabular Data: A Survey,” IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 7499-7519, June 2024,. https://doi.org/10.1109/TNNLS.2022.3229161

[9] Agresti, “Categorical Data Analysis,” WileyInterscience, 2002. https://onlinelibrary.wiley.com/doi/book/10.1002/0471249688

[10] Mehdi Mirzaet, Simon Osindero, “Conditional Generative Adversarial Nets,” arXiv.org, 6 November 2014. https://doi.org/10.48550/arXiv.1411.1784

[11] Augustus Odena, “Semi-Supervised Learning with Generative Adversarial Networks,” arXiv:16-06.01583 Statistics, 2017. https://doi.org/10.48550/arXiv.1606.01583

[12] Minjung Kyung and Jeff Gill, George Casella, “Estimation in Dirichlet random effects models,” Ann. Statist, 38(2), pp. 979-1009. April 2010. https://doi.org/10.1214/09-AOS731

[13] Marcello D’Orazio, “Distances with Mixed-Type Variables, some Modified Gower’s Coefficients,” arXiv:2101.02481, 2021. https://arxiv.org/abs/2101.02481

[14] Monia Ranalli, Roberto Rocci, “Applying Gower distance as a dissimilarity measure for mixed type data in a clustering problem,” Springer Nature Switzerland

AG, 2021.
https://arxiv.org/pdf/2101.02481

[15] Z. Zhang, “Generalized Mutual Information,” MDPI, 3(2), pp. 158-165, 2020.
https://doi.org/10.3390/stats3020013

[16] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, Pieter Abbeel, “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets,” Advances in Neural Information Processing Systems 29, 2016.
https://doi.org/10.48550/arXiv.1606.03657

[17] Dilan G¨or¨ur1, Carl Edward Rasmussen, “Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution,” Journal of Computer Science and Technology, Vol. 25, pp. 615-626, 2010.
https://doi.org/10.1007/s11390-010-9355-8

[18] Minjung Kyung and Jeff Gill, George Casella, “Estimation in Dirichlet random effects models,” Ann. Statist., 38(2), pp. 979-1009. April 2010.
https://doi.org/10.1214/09-AOS731

[19] Shanshan Jiang, Ruiting Dong, Jie Wang, Min Xia, “Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network,” Systems 2023, MDPI, 11(6), 305, 2023.
https://doi.org/10.3390/systems11060305

[20] Kingma, D. P., Ba, J, “Adam: A Method for Stochastic Optimization,” arXiv preprint arXi-v: 1412.6980, 2014.
https://doi.org/10.48550/arXiv.1412.6980

[21] J. Zou, J. Zhang, P. Jiang, “Credit Card Fraud Detection Using Autoencoder Neural Network,” arXiv: 1908. 11553, 2019.
https://arxiv.org/abs/1908.11553

[22] Lei Zhang, Fang Yuan, KaiFeng Ma, We-nJun Fang, “A Tabnet based Card Fraud detetion Algorithm with Feature Engineering,” 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, pp. 911-914, 2022.
https://doi.org/10.1109/ICCECE54139.2022.9712822

[23] Chew Chee Meng, Kian Ming Lim, Chin Poo Lee, Jit Yan Lim, “Credit Card Fraud Detection using TabNet,” 23 11th International Conference on Information and Communication Technology (ICoICT), Melaka, Malaysia, pp. 394-399, 2023.
https://doi.org/10.1109/ICoICT58202.2023.10262711

◐ 저 자 소 개 ◑

성 찬 식(Chan-sik Sung)
2011년 한국방송통신대학교 컴퓨터과학과(공학사)
2015년 세종대학교 대학원 디지털정보학과(공학석사)
2023년 가천대학교 일반대학원 IT융합학과(공학박사 수료)
관심분야 : 인공지능, etc.
E-mail : mob2000@gachon.ac.kr

임 준 식(Joon-sik Lim)
1986년 인하대학교 전자계산학과(공학사)
1989년 (미)ALABAMA, UNIV. OF Computer Science(공학석사)
1994년 (미)LOUISIANA STATE UNIV. Computer Science(공학박사)
1995년~현재 가천대학교 일반대학원 컴퓨터공학부 교수
관심분야 : 인공지능, etc.
E-mail : jslim@gachon.ac.kr