JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 1: 64~69, Jan. 2022

공공기술 사업화를 위한 CTGAN 기반 데이터 불균형 해소

황철현*

Resolving CTGAN-based data imbalance for commercialization of public technology

Chul-Hyun Hwang*

*Associate Professor, Department of Software Fusion, Kyoung-Bok University, Gyeonggi, 12051 Korea

요 약

공공기술 사업화는 정부가 주도하는 과학기술의 혁신과 R&D 성과를 민간에 이전하는 것으로 경제 성장을 주도하는 핵심 성과로 인식되고 있다. 따라서 기술 이전을 활성화시키기 위해 성공 요인을 식별하거나 사업화 가능성이 높은 공공기술과 수요기업을 매칭하는 다양한 기계학습의 방법들이 연구되고 있다. 하지만 공공기술 사업화 데이터는 표 형태로 구성되어 있고, 성공-실패 비율이 큰 차이를 보이는 불균형 상태이기 때문에 기계학습 성능이 높지 않는 문제점을 가지고 있다. 이 논문에서는 표 형태로 구성된 공공기술 데이터에서 불균형을 해소하기 위해 CTGAN을 활용하는 방법을 제시한다. 또한 제시된 방법의 효과를 검증하기 위해 실제 공공기술 사업화 데이터를 활용하여 통계적 접근방법인 SMOTE와 비교 실험을 수행하였다. 다수의 실험 사례에서 CTGAN은 공공기술 사업화 성공사례를 안정적으로 예측하는 것을 확인하였다.

ABSTRACT

Commercialization of public technology is the transfer of government-led scientific and technological innovation and R&D results to the private sector, and is recognized as a key achievement driving economic growth. Therefore, in order to activate technology transfer, various machine learning methods are being studied to identify success factors or to match public technology with high commercialization potential and demanding companies. However, public technology commercialization data is in the form of a table and has a problem that machine learning performance is not high because it is in an imbalanced state with a large difference in success-failure ratio. In this paper, we present a method of utilizing CTGAN to resolve imbalances in public technology data in tabular form. In addition, to verify the effectiveness of the proposed method, a comparative experiment with SMOTE, a statistical approach, was performed using actual public technology commercialization data. In many experimental cases, it was confirmed that CTGAN reliably predicts public technology commercialization success cases.

키워드 : 공공기술 사업화, 데이터 불균형, 데이터 증폭, 표 형태 데이터, CTGAN

Keywords : Public technology commercialization, Data Imbalance, Data Amplification, Tabular Data, CTGAN

Received 22 November 2021, **Revised** 24 November 2021, **Accepted** 3 December 2021

*** Corresponding Author** Chul-Hyun Hwang(E-mail:chhwang@kbu.ac.kr, Tel:+82-31-570-9608)
Professor, Department of Software Fusion, KyoungBok University, Gyronggi-do, 12051 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.1.64 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

과학기술의 혁신과 연구개발이 국가의 경제성장을 주도하는 요인으로 부각되면서, 주요선진국에서는 정부가 주도하는 과학기술에 대한 연구개발 투자를 매우 공격적으로 수행하고 있다[1]. 이와 함께 연구개발이 완료된 과학기술을 민간에 이전하고, 사업화하는 실적을 국가의 과학기술에 대한 효율성을 측정하는 지표로 삼고 관리하고 있다[2].

이와 같은 사업화 지표를 높이기 위한 기존의 연구 방향은 설문조사 등을 통해 성공요인을 찾고자하는 정책적 연구가 주를 이루고 있다[1][2]. 이와 달리 정부는 공공기술을 민간의 수요기업과 연결하고, 산업에서 필요한 기술을 선정하는 빅데이터 기반의 기계학습 시스템을 구축하는 등 구체적으로 접근하고 있다[3][4]. 빅데이터 기반의 시스템 구축 초기 단계는 기계학습 예측모델에 대한 연구와 함께 예측의 기반이 되는 학습 데이터에 대한 연구도 병행되어야 하므로, 이 논문에서는 데이터 측면에서 접근하고자 한다.

공공기술 사업화와 수요기업 예측과정에서 발생하는 데이터 문제 중 하나는 데이터 불균형(data imbalance)이다. 지난 20년간의 데이터에서 정부의 주도로 개발된 공공기술이 기업에 이전된 사례가 전체 개발 실적에 비해 현격히 낮은 데이터 불균형 문제로 인해 기계학습의 예측 성능이 낮게 산출되는 것이다. 이는 기계학습 과정에서 다수 분류로 예측을 편향시키거나 소수 분류를 비정상 데이터로 취급하기 때문이다.

데이터 불균형에 의해 편향학습이 발생할 경우 다수 분류에 대한 예측 정확도는 높지만 소수 분류의 예측 정확도는 낮은 현상을 보인다. 이 경우 다수 분류의 사례로 인해 전체 예측 정확도는 높게 나타날 수 있지만 소수 분류를 찾는 것이 예측 목표이기 때문에 실질적인 예측 정확도는 낮은 문제점이 발생한다. 공공기술 사업화 영역에서도 분석 목표는 소수 사례인 공공기술의 성공사례를 예측하는 것으로 실패사례를 예측하는 것보다 더욱 중요한 의미를 가진다.

이러한 불균형 문제를 해결하기 위해서 많이 사용하는 기법은 오버샘플링(over sampling)과 언더샘플링(under sampling)이다. 두 기법은 데이터를 생성하거나 삭제를 통해 소수 분류와 다수 분류의 데이터 비율을 비슷하게 유지시켜준다. 최근 불균형 문제를 해결하기 위해서 오버샘플링에 더욱 주목하고 있다. 이것은 언더샘플링과 같이 소수 사례의 데이터 수에 다수 사례의 데이터 수를 맞추게 될 경우, 공공기술 사업화와 같이 소수 사례의 데이터 수가 절대적으로 적은 경우에 전체 학습 데이터의 수가 줄어들기 때문이다.

반면, 오버샘플링은 소수 사례의 데이터 수를 다수 사례의 데이터 수만큼 생성하는 방법이기 때문에 얼마나 실제 데이터와 유사하게 만들어 예측모델의 성능에 긍정적인 영향을 미치는 지가 중요하다. 주요 알고리즘으로는 통계적 기반의 접근방법인 SMOTE와 딥러닝 기반의 접근방법인 GAN이다[5].

이 논문에서는 공공기술에 대한 사업화 예측을 위해 오버샘플링을 수행하는 방법과 절차를 제시하고, SMOTE와 GAN 알고리즘을 적용한 예측성능을 비교하여 제시한다. 특히 표 형태로 구성된 공공기술 관련 데이터 특성을 반영하고, 사업화 성공여부에 따른 데이터를 구분 생성하기 위해 CTGAN(Conditional Tabular GAN)을 적용한다.

실제 공공기술사업화 데이터를 샘플링하여 실험하였으며, 반복 실험을 위해 k-Folder를 적용하였다. 또한 증폭 데이터에 의한 예측모델의 성능 변화를 파악하기 위해 대표적인 분류 알고리즘인 Random Forest를 통해 성능을 산출하였다.

실험에서 SMOTE에 비해 CTGAN에 의해 증폭된 데이터를 사용한 분류 성능에서 일관되게 높은 분류 정확도를 보임으로써 제시된 적용 방법에 따른 CTGAN이 최적방법임을 입증하였다.

Ⅱ. 관련 연구

공공기술 사업화 관련 연구와 데이터 현황, 데이터 불균형을 해소하기 위한 데이터 증폭기술에 대한 연구 현황에 대해 제시한다.

2.1. 공공기술 사업화 관련 연구

정부 주도 연구개발의 목적 중의 하나는 공공 연구기관의 연구개발 결과로 발생한 기술이 기업으로 이전되어 산업에 기여하는 것이다[6]. 이를 위해 공공기술을 기업에 성공적으로 이전하는데 미치는 영향 요인과 효율성에 관한 연구가 지속되고 있다[7].

하지만 공공기술 이전 및 사업화와 관련된 많은 연구에도 불구하고 기존의 연구는 문헌이나 설문 조사를 통한 정책연구에 치중되어 왔었다. 좀 더 적극적으로 사업화 성과를 달성하기 위해서는 정책적 연구에서 벗어나 개별 수요기업의 니즈를 만족시킬 수 있는 구체적인 서비스와 매칭 방법에 대한 연구가 필요하다.

이러한 필요성으로 인해 정부는 지난 20년간 축적된 공공기술사업화와 기술 이전 관련된 데이터를 데이터 플랫폼으로 구축하고, 수요기업과 공공기술을 매칭하는 맞춤형 서비스 구축을 추진하고 있다[3][4]. 맞춤형 서비스는 기존의 성공 요인을 찾는 정책 연구과정과 달리 데이터 기반의 구체적이고 실증 기반의 연구와 기술 개발이 필요하다.

본 논문은 이러한 요구에 부응하기 위해 공공기술과 수요기업을 매칭하는데 필요한 데이터 확보기술과 적용 방법을 제시하고 실험을 통해 효과를 검증한다.

2.2. 데이터 불균형 관련 연구

공공기술 사업화 데이터는 개발된 전체 공공기술에서 수요기업에 이전된 사례가 소수를 차지하는 전형적인 불균형 데이터이다.

데이터 불균형은 기계학습에서 분류기법을 활용할 때, 기계학습 모델이 다수분류에 편향 학습되어 분류성능 중 하나인 재현율(recall)에 악영향을 미치게 된다. 즉 소수 분류인 사업화 성공사례를 찾지 못하고, 다수 분류인 사업화 실패사례로 분류하게 되어 낮은 재현율이 산출되는 것이다. 데이터의 불균형은 실제 현실세계에서 자주 발생하는데다 소수 분류가 업무에서 차지하는 비중이 크기 때문에 다양한 분야에서 데이터 분류를 해결하기 위한 노력이 시도되었다.

이한수 등은 대기관측을 위한 레이더 전파의 이상굴절의 발생 신호인 이상전파에코(anomalous propa-gation echo)를 식별하기 위해 베이즈분류기 기반의 SMOTE를 활용하는 방법과 성능을 제시하였다[8]. SMOTE는 Chawla에 의해 개발된 방법으로, 소수 분류에 포함된 데이터와 이웃 데이터를 KNN (k-Nearest neighbor)을 활용하여 추출한 후, 보간하여 신규 데이터를 생성하는 방법이다.

SMOTE는 인접한 다수 분류의 데이터를 고려하지 않아 다른 분류와 겹쳐서 데이터가 생성되기 때문에 고차원의 데이터에는 활용할 수 없는 문제가 있다.

이러한 문제로 인해 최근에는 인공신경망 기반의 딥러닝을 활용하는 시도가 증가하였다. 특히 적대적 신경망인 GAN은 데이터 생성을 통해 불균형 문제를 해결하는 방법으로 주목받고 있다. 김희수 등의 연구에 의하면 인공신경망이 목적함수에만 집중하면서 실제 데이터 범주에는 발생할 수 없는 데이터를 생성하는 모델 붕괴(model collapsing) 문제를 제기하고 있다.

이를 해소하기 위해 GAN 기반의 오버 샘플링 방법을 제안함으로써 생성 데이터 값이 실제 데이터와 유사한 분포를 가지도록 하였다[9].

위와 같은 연구에도 불구하고 GAN은 이미지 분류를 위한 CNN에서 발전되었기 때문에 표 형태의 데이터를 증폭하는 데에는 다음과 같은 한계가 있었다[6].

- 연속·불연속 데이터와 같은 다양한 데이터 형태
- 가우시안 분포를 따르지 않는 데이터 분포
- 텍스트 등을 포함한 멀티 모달 데이터 형태
- one-hot encoding에 의해 발생된 희소 행렬
- 높은 불균형을 가진 범주형 변수

위의 문제를 해결하여 GAN을 표 형태 데이터에서도 활용할 수 있도록 하는 여러 연구가 진행되었는데 그 가운데 하나가 바로 CTGAN이다. CTGAN(conditional tabular GAN)은 데이터를 증폭할 때 앞서 제시된 표 형태의 특성 문제를 종속변수의 조건별 확률 밀도를 활용하는 방법이다. 이를 통해 CTGAN은 종속변수 불균형을 가진 지도학습 데이터에 유용하게 활용될 수 있다. 다음 그림 1과 수식(1)은 CTGAN이 각 조건별 데이터를 생성하기 위해 실제 데이터의 분포를 재구성하는 수식과 방법을 제시하였다[5].

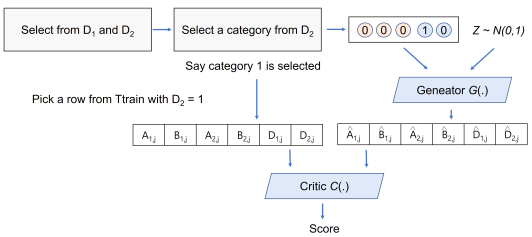


Fig. 1 CTGAN model process

$$P(row) = \sum_{k \in D_{i*}} P_{\zeta}(row | D_{i*} = k^{*}) P(D_{i*} = k) \qquad (1)$$

Ⅲ. 불균형 해소를 위한 증폭 방법

3.1. 공공기술 사업화 데이터 구조

공공기술사업화 데이터에 대한 핵심 구조는 2개의 마스터 데이터(공공기술과제, 수요기업)가 기술사업화 이력 데이터를 통해 상호 연결되어 있는 데이터 구조(data model)를 가지고 있다. 다음 그림 2는 공공기술 사업화 데이터의 개념적 구조(conceptual data model)를 설명한다.

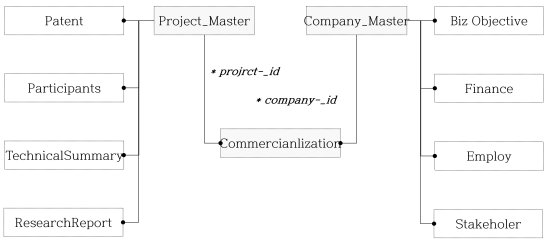


Fig. 2 Data structure of public technology commercialization

3.2. 공공기술 사업화 데이터에 대한 증폭 절차

공공기술 사업화 데이터에 대한 증폭 효과를 측정하기 위해 실험 데이터 확보와 분류 성능 비교의 두 단계를 거친다. 먼저 실험 데이터를 확보하는 방법은 다음과 같다.

- 실험을 효율적으로 수행할 수 있도록 전체 데이터에서 random sampling을 수행한다. (1,000:100)
- 전처리(one-hot encoding, min-max scaler)를 수행하여 기계학습이 가능한 데이터를 준비한다.
- 데이터 증폭에 의한 효과를 반복 측정하기 위해 x-folder를 수행한다. (x = 5)
- 수행 시 마다 훈련과 시험구간의 데이터를 구분하여 실험 데이터를 구축한다.
- 훈련 데이터를 활용하여 SMOTE, CTGAN을 수행하여 데이터를 증폭한다. 증폭된 데이터를 학습 데이터에 통합하여 최종 훈련 데이터를 구축한다.

확보된 데이터(원본, 증폭)를 활용하여 분류 성능을 측정하는 실험 절차는 다음과 같다.

- 각 증폭 사례의 데이터를 활용하여 동일한 환경의 random forest를 수행하여 분류 결과를 생성한다.
- 분류 결과를 활용하여 각 증폭 사례의 분류 성능을 비교 분석한다.

앞서 제시한 실험 절차를 통해 확보된 데이터의 분포는 다음 그림 3과 같다. 원본 데이터(a)는 불균형을 포함한 상태인 것을 알 수 있고, SMOTE와 CTGAN을 통해 데이터 불균형이 증폭을 통해 해결(b,c)되었음을 보여준다.

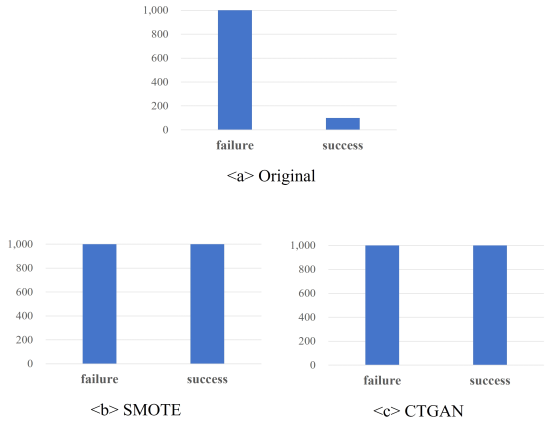


Fig. 3 Data structure of public technology commercialization

다음 그림 4는 실험에 대한 전 과정을 도식화하여 제시하였다. 실험 데이터는 x-folder에 의해 증폭 방법별로 5회 생성되고, 매회 정확도를 증폭 방법별로 비교하여 평가한다.

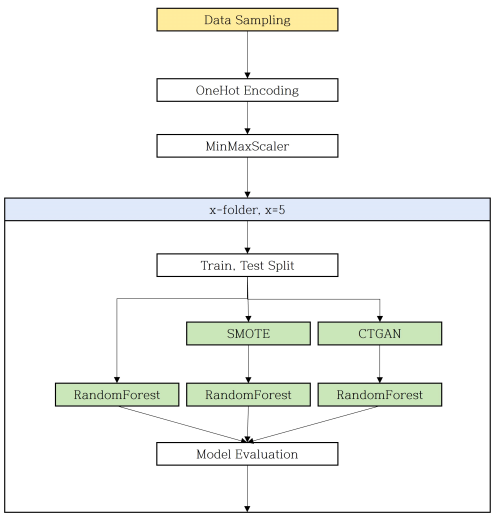


Fig. 4 Data structure of public technology commercialization

Ⅳ. 실험결과 및 비교분석

4.1. 실험 결과

3장에서 제시한 절차에 따라 원본 데이터를 그대로 활용한 방법, SMOTE를 통한 증폭된 데이터를 활용하는 방법, CTGAN을 활용하여 증폭한 데이터를 활용하는 방법의 3가지 방법에 의한 분류 성능을 상호 비교한다.

그림 5에 의하면 CTGAN이 불균형을 포함한 데이터 원본, SMOTE에 의한 방법에 비해 precision, recall, f1-score의 모든 지표에서 높은 성능을 보이고 있다.

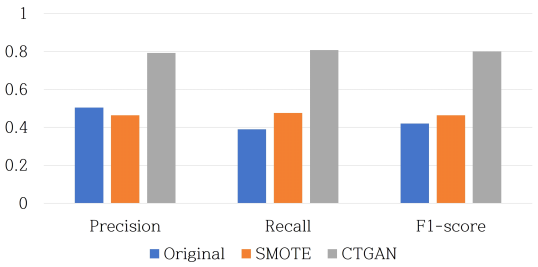


Fig. 5 Comparison of classification performance

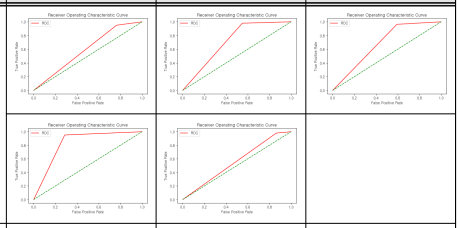
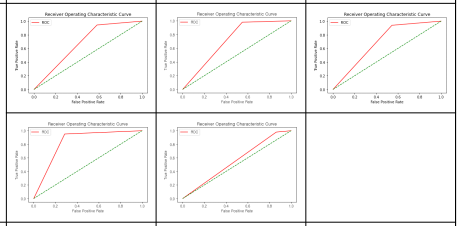
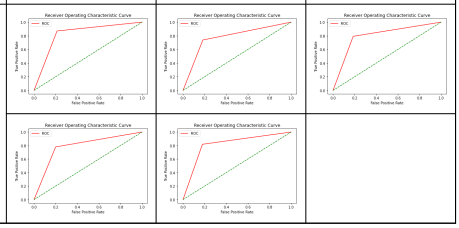
다음 표 1에서 각 실험 case별로 소수 분류(사업화 성공)에 대한 분류 성능을 상세하게 제시하였다.

Table. 1 Ratio of Commercialized projects

division	seq	precision	recall	f1-score	support
original	1	0.29	0.24	0.26	17
	2	0.69	0.45	0.55	20
	3	0.61	0.41	0.49	27
	4	0.50	0.71	0.59	14
	5	0.43	0.14	0.21	22
SMOTE	1	0.48	0.42	0.44	24
	2	0.56	0.43	0.49	21
	3	0.45	0.45	0.45	27
	4	0.50	0.71	0.59	14
	5	0.33	0.37	0.35	14
CTGAN	1	0.88	0.79	0.83	216
	2	0.72	0.82	0.77	180
	3	0.76	0.81	0.79	177
	4	0.76	0.80	0.78	185
	5	0.84	0.82	0.83	212

다음 표 2는 각 실험 case에서 소수 분류(사업화, 기술이전 성공)의 ROC curve를 제시하였다.

Table. 2 Ratio of Commercialized projects

division	case 1/case 4	case 2/case 5	case 3
Original			
SMOTE			
CTGAN			

위 표 1과 표 2를 통해 CTGAN의 성능이 다른 실험 case에서보다 높은 분류 성능을 보여주는 것과 함께, 각 실험 case의 결과가 유사하게 산출되는 안정된 결과를 보여준다는 것을 알 수 있다. 이러한 결과는 향후 실제 서비스에 적용하더라도 데이터 환경의 변화에 덜 민감하게 기계학습이 반응하여 안정된 성능을 제공할 수 있다는 것을 추정할 수 있다.

다음 표 3에서는 각 실험에서 분류 성능의 종합적인 성능을 나타내는 f1-score에 대한 평균을 산출하여 데이터 증폭 방법별로 구분하여 제시하였다. 공공기술 사업화 데이터가 차원이 많기 때문에 SMOTE는 극단적으로 높은 효과를 보지 못했지만 CTGAN의 경우 원본에 비해 90% 정도의 성능 향상 효과를 본 것으로 나타났다.

Table. 3 Ratio of Commercialized projects

dividion	average of F1-score
Original	0.42
SMOTE	0.46
CTGAN	0.8

Ⅴ. 결 론

공공기술 사업화의 성과는 국가 경쟁력을 나타내는 지표로 활용되는 매우 중요한 문제이다. 또한 정부는 기존의 정책적 연구에서 벗어나 빅데이터 기반의 기계학습을 통한 공공기술과 수요기업을 매칭하고 추천 서비스를 제공하는 구체적인 시도를 하고 있다.

본 논문은 데이터 불균형의 문제를 지적하고 이를 극복하기 위한 CTGAN 기반의 데이터 증폭방법에 대한 적용 방법을 제시하였다. 또한 실제 공공기술 데이터를 활용하여 실증 실험을 수행하여 제시된 방법에 대한 효과를 비교 분석하였다. 실험에서 CTGAN을 활용한 데이터 증폭 방법의 효과가 원본 데이터보다 1.9배의 성능 향상 효과가 있음을 확인하였고, 반복적인 실험환경에서도 안정적으로 성능을 보여준다는 것을 확인할 수 있었다.

향후 연구 분야는 수요기업에 대한 데이터 증폭 효과 검증과 함께 딥러닝 기반의 분류 알고리즘에 대한 효과를 검증하는 것이다.

References

[1] G. M. Grossman and E. Helpman, “Innovation and growth in the global economy,” *MIT Press*, 1991.

[2] T. H. Kwon, “What makes Korean firms transfer public technology and commercialize well? : An empirical study on public technology licensee firms,” Ph. D. dissertation, Hanyang University, Seoul, Korea, 2020.

[3] KISTI Institutional Repository. Finding a way for innovative growth of SMEs in data-based technology commercialization [Internet]. Available: https://repository.kisti.re.kr/handle/10580/15035.

[4] KISTI Institutional Repository. KISTI's technology commercialization platform continues to spread [Internet]. Available: https://repository.kisti.re.kr/handle/10580/15243.

[5] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling Tabular Data using Conditional GAN,” in *Proceeding of the 33th Conference on Neural Information Processing Systems(NeurIPS2019)*, Vancouver: NY, 2019.

[6] K. S. Hwang, “R&S accountability and dilemma within the Korean science and technology context,” *Korean Public Administration Review*, vol. 50, no. 2, pp. 189-213, Jun. 2016.

[7] D. H. Jo, S. H. Choi, S. K. Kim, and H. J. Lee, “The Effect of Public Technology Value on Technology Transfer Performance,” *Journal of Digital Convergence*, vol. 16, no. 3, pp. 189-199, Mar. 2018.

[8] H. S. Lee and S. S. Kim, “Naive Bayes Clasifier based Anomalous Propagation Echo Identification using Class Imbalanced Data,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 20, no. 6, pp. 1063-1068, Jun. 2016.

[9] H. S. Kim and H. S. Lee, “Generative Adversarial Networks based Data Generation Framework for Overcoming Imbalanced Manufacturing Process Data,” *Journal of Korean Institute of Intelligent Systems*, vol. 29, no. 1, pp. 1-8, Feb. 2019.

황철현(Chul-Hyun Hwang)

1991년 금오공과대학교 전자공학과(공학사)
1995년 경남대학교 컴퓨터공학과(공학석사)
2015년 배재대학교 컴퓨터공학과(공학박사)
2019~현재 경복대학교 소프트융합과 부교수
※관심분야 : 빅데이터, 인공지능, 기계학습, 딥러닝, IoT, Data Architecture