KIPS Trans. Softw. and Data Eng.
Vol.11, No.12 pp.489~498
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.12.489

The Development of Biodegradable Fiber Tensile Tenacity and Elongation Prediction Model Considering Data Imbalance and Measurement Error

Se-Chan Park† · Deok-Yeop Kim†† · Kang-Bok Seo†† · Woo-Jin Lee†††

ABSTRACT

Recently, the textile industry, which is labor-intensive, is attempting to reduce process costs and optimize quality through artificial intelligence. However, the fiber spinning process has a high cost for data collection and lacks a systematic data collection and processing system, so the amount of accumulated data is small. In addition, data imbalance occurs by preferentially collecting only data with changes in specific variables according to the purpose of fiber spinning, and there is an error even between samples collected under the same fiber spinning conditions due to difference in the measurement environment of physical properties. If these data characteristics are not taken into account and used for AI models, problems such as overfitting and performance degradation may occur. Therefore, in this paper, we propose an outlier handling technique and data augmentation technique considering the characteristics of the spinning process data. And, by comparing it with the existing outlier handling technique and data augmentation technique, it is shown that the proposed technique is more suitable for spinning process data. In addition, by comparing the original data and the data processed with the proposed method to various models, it is shown that the performance of the tensile tenacity and elongation prediction model is improved in the models using the proposed methods compared to the models not using the proposed methods.

Keywords : Data Imbalance, Outlier Handling, Data Augmentation, Tensile Tenacity and Tensile Elongation, Biodegradable Fiber(PLA)

데이터 불균형과 측정 오차를 고려한 생분해성 섬유 인장 강신도 예측 모델 개발

박 세 찬† · 김 덕 엽†† · 서 강 복†† · 이 우 진†††

요 약

최근 노동 집약적인 성격의 섬유 산업에서는 인공지능을 통해 섬유 방사 공정에 들어가는 비용을 줄이고 품질을 최적화하려고 시도 하고 있다. 그러나 섬유 방사 공정은 데이터 수집에 필요한 비용이 크고 체계적인 데이터 수집 및 처리 시스템이 부족하여 축적된 데이터양이 적다. 또 방사 목적에 따라 특정한 변수에만 변화를 준 데이터만을 우선으로 수집하여 데이터 불균형이 발생하며, 물성 측정 환경의 차이로 인해 동일 방사 조건에서 수집된 샘플 간에도 오차가 존재한다. 이러한 데이터 특성들을 고려하지 않고 인공지능 모델에 활용할 경우 과적합과 성능 저하 등의 문제가 발생할 수 있다. 따라서 본 논문에서는 방사 공정 데이터 특성을 고려한 이상치 처리 기법과 데이터 증강 기법을 제안한다. 그리고 이를 기존 이상치 처리 기법 및 데이터 증강 기법과 비교하여 제안한 기법이 방사 공정 데이터에 더 적합함을 보인다. 또 원본 데이터와 제안한 기법들로 처리된 데이터를 다양한 모델에 적용하여 비교함을 통해 제안한 기법들을 사용한 모델들이 그렇지 않은 모델들에 비해 인장 강신도 예측 모델의 성능이 개선됨을 보인다.

키워드 : 데이터 불균형, 이상치 처리, 데이터 증강, 인장 강신도, 생분해성 섬유(PLA)

※ 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. NRF-2018R1A6A1A03025109)이며 본 연구는 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구이며(P0022335) 또한 교육부 및 한국연구재단의 4단계 BK21 사업(경북대학교 컴퓨터학부 지능융합 소프트웨어 교육연구단)으로 지원된 연구임(4199990214394).
※ 이 논문은 2022년 한국정보처리학회 ASK 2022의 우수논문으로 "생분해성 섬유 방사 공정 데이터 특성을 고려한 물성 예측 모델 개발"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경북대학교 컴퓨터학부 석사과정
†† 준 회 원 : 경북대학교 컴퓨터학부 박사과정
††† 정 회 원 : 경북대학교 컴퓨터학부 교수
Manuscript Received : August 1, 2022
First Revision : September 13, 2022
Accepted : September 18, 2022
* Corresponding Author : Woo-Jin Lee(woojin@knu.ac.kr)

1. 서 론

최근 섬유 산업에서는 인공지능을 적용함으로써 자동화를 통해 비용을 줄이거나 품질을 최적화하려고 시도하고 있다[1]. 일반적으로 인공지능 모델의 성능은 데이터의 양과 질의 영향을 크게 받는다. 하지만 섬유 방사 분야에서는 인공 지능 모델에서 사용할 데이터의 양과 질을 확보하는 데 몇 가지 어려움이 있다. 먼저, 섬유 방사 공정 데이터를 수집하기 위해 소요되는 비용이 크기 때문에 충분한 양의 데이터 확보가 어렵고 체계적인 데이터 수집 및 관리 시스템이 부족하여 축적된 데이터의 양이 적다. 또 방사 목적에 관련된 특정한 변수에만 변화

를 주어 데이터를 수집하기 때문에 데이터 불균형 문제가 나타난다. 그리고 물성 측정 환경의 차이로 인해 동일한 방사 조건에서 수집된 데이터 샘플 간에도 측정값의 차이가 난다. 이러한 데이터 부족, 데이터 불균형, 샘플 간 오차 등의 데이터 특성을 고려하지 않고 인공지능 모델에 사용할 경우 모델에서 과적합 및 모델 성능 저하 등의 문제가 발생하기 쉽다.

본 논문에서는 이러한 문제를 해결하기 위해 방사 공정 데이터 특성을 고려한 이상치 처리 기법 및 데이터 증강 기법을 제안한다. 제안한 이상치 처리 기법은 공정관리한계 허용오차와 동일 방사 조건 내 샘플 개수를 고려하며 박스 플롯 기반 이상치 처리 기법 및 클러스터링 기반 이상치 처리 기법 적용 결과를 비교한다. 또 제안한 데이터 증강 기법은 공정변수별 예측 물성에 대한 상관계수와 불균형 비율을 고려하며 랜덤 오버샘플링 기법 및 SMOTE 기법 적용 결과와 비교한다. 이런 비교 결과 제안한 데이터 처리 기법들이 기존의 데이터 처리 기법과 비교하여 방사 공정 데이터에 더욱 적합함을 보인다. 또 원본 데이터와 제안한 기법들로 처리된 데이터를 다양한 모델에 적용하여 비교함을 통해 제안한 기법들을 사용한 모델이 그렇지 않은 모델에 비해 인장 강신도 예측 모델의 성능이 개선됨을 보인다.

본 논문의 2장에서는 관련 연구에 대하여 서술한다. 3장에서는 생분해성 섬유 방사 공정 데이터에 대해 구체적으로 서술하고 제안한 기법과 기존의 기법들을 비교하여 제안한 기법이 방사 공정 데이터에 더 적합함을 보인다. 4장에서는 원본 데이터와 제안한 기법들로 처리된 데이터를 다양한 모델에 적용하여 비교한다. 마지막으로 5장에서는 방사 공정 데이터 특성을 고려한 기법으로 모델 성능 개선 결과에 대해 서술하고 전체 연구 요약 및 결론에 대해 서술한다.

2. 관련 연구

2.1 기존 섬유 산업 분야에서의 모델 개발 연구

기존 섬유 산업 연구에서는 품질 개선을 위해 통계적으로 데이터를 분석하거나 수학적 모델링으로 데이터를 분석한다[2, 3]. 하지만 통계적 분석 방식은 단순한 선형 모델의 경우에는 유용하지만 복잡한 비선형 모델의 경우 적용하기 어렵다는 단점이 있다. 또 수학적 모델링의 경우 관련 수식이 알려져 있거나 구조화가 용이할 정도로 모델이 단순한 경우에는 유용하지만 그렇지 않은 경우 적용하기 어렵다는 단점이 있다. 이 외에도 30개의 매우 적은 데이터를 인공지능에 적용해 면방적사의 강도를 예측한 연구는 이른 시기에 인공지능을 섬유 산업에 적용하였다는 의의가 있으나 30개의 데이터가 모집단을 대표한다고 보기 어렵기 때문에 모델의 신뢰도가 떨어진다[4].

이처럼 기존 연구들에서 통계적 분석이나 수학적 모델링을 주로 선택하고 인공지능 적용 사례가 드문 것은 섬유 산업 분야에서 데이터 수집 비용이 크고 인공지능 적용에 충분한 데이터를 확보하기 어렵기 때문이다.

2.2 불균형한 데이터에서의 이상치 처리 기법

데이터가 불균형할 때 일반적인 이상치 처리 기법을 사용하면 데이터 분포가 적은 구간에 존재하는 정상 데이터들을 이상치로 잘못 탐지할 수 있다. 따라서 데이터가 불균형한 경우에는 이상치를 제대로 탐지하기 어렵다. 이를 해결하기 위해 기존 연구에서는 클러스터링 기법과 앙상블 기법을 기반으로 이상치를 탐지한다[5]. 해당 연구에서는 정상 데이터가 주어진다는 가정하에 기존 데이터를 먼저 클러스터링하고 새롭게 들어오는 데이터에 대해서 각 클러스터에 대해 이상치 판단을 수행하여 그 결과를 앙상블 기법으로 종합한다. 이는 충분한 양의 정상 데이터가 기존에 주어진 경우 유용하지만 기존 데이터의 정상 여부를 모르는 경우 클러스터링 기법을 적용할 수 없다는 단점이 있다.

2.3 회귀 문제에서의 데이터 증강 기법

데이터 불균형을 해결하기 위해 소수 클래스 데이터를 증강하는 것을 데이터 증강이라고 한다. 대부분의 기존 연구들은 분류 문제에서의 클래스 불균형을 해결하기 위한 데이터 증강 기법들을 다룬다[6, 7]. 그러나 회귀 문제에서도 변수의 구간별 데이터 분포 불균형에 따른 데이터 불균형은 존재한다. 대부분의 기존 기법들은 불연속적인 클래스의 불균형을 대상으로 하지만 회귀 문제인 섬유 물성 예측은 연속적인 공정 변수의 불균형을 대상으로 하기 때문에 그대로 적용하기 어렵다. 또 특정 변수의 데이터를 증강한 후에 오히려 다른 변수들의 데이터 불균형이 더 심해질 수 있다는 문제도 있다.

3. 다양한 데이터 처리 기법 적용 및 비교

3.1 생분해성 섬유 방사 공정 데이터

본 논문에서 활용하는 데이터는 816개의 생분해성 섬유 방사 공정 데이터이다. 데이터는 55개의 방사 공정 조건 변수와 2개의 예측 대상 물성인 인장 강도 측정값과 인장 신도 측정값으로 이루어진다. 생분해성 섬유 방사 공정 현장 전문가의 의견에 따르면 55개의 방사 공정 조건 변수 중 인장 강신도에 주로 영향을 주는 핵심 변수는 8개이다. 현재 수집한 데이터의 경우 제한된 시간과 비용으로 인해 핵심 변수 8개 중 2개 변수는 항상 단일 값이다. 따라서 모델에 영향을 주지 않는 단일 값을 제외한 변수인 스핀빔 온도, 롤러 속도, 롤러 온도, 권취 속도, 연신비 등의 6개 변수를 주요 공정 변수로 사용한다. 연신비의 경우 롤러1과 롤러2의 단순한 속도비이기 때문에 데이터 불균형 비율이나 상관 계수를 구할 때는 제외한다.

본 논문에서 활용하는 생분해성 섬유 방사 공정 데이터는 한 번의 방사 공정 구동 시 최대 4개의 샘플을 수집하여 동일 방사 조건에 대해 최대 4개의 샘플이 존재한다.

Table 1. Data Distribution and Ratio, Major/Minor Class by Data Section

	Data distribution by data section				
Spin beam temperature	250-253	254-257	258-261	262-265	266-268
	13.2%	22.9%	43.3%	10.7%	9.9%
	minor	major	major	minor	minor
Roller 1 speed	1000-1200		1201-1500		1501-3500
	3%		88.6%		8.4%
	minor		major		minor
Roller 2 speed	4039		4105		4140-4520
	6.5%		75.8%		17.7%
	minor		major		minor
Roller 2 temperature	95		100		105
	6.9%		92%		1.1%
	minor		major		minor
F/R speed	4000	4100	4200	4300	4400
	82.3%	1.5%	6.4%	4.9%	4.9%
	major	minor	minor	minor	minor

본 논문에서는 데이터를 크게 다수 구간 데이터와 소수 구간 데이터로 구분한다. 주요 공정 변수의 구체적인 구간과 그 비율, 다수/소수 구분은 Table 1에 나타낸다.

3.2절과 3.3절에서는 주 예측 대상 물성인 인장 강도를 기준으로 다양한 데이터 처리 기법을 적용하고 제안하는 데이터 처리 기법이 방사 공정 데이터에 가장 적합함을 보인다. 3.4절에서는 부 예측 대상 물성인 인장 신도를 기준으로 제안하는 데이터 처리 기법을 적용한다.

3.2 다양한 이상치 처리 기법 적용 및 비교

이상치 처리는 기본적으로 다른 관측값과 상이한 관측값을 찾는 것이기 때문에 데이터 분포가 불균형한 경우 정상적인 소수 구간 데이터들이 이상치로 탐지될 수 있다는 문제가 있다. 따라서 데이터 불균형이 나타나는 경우 일반적인 이상치 처리 기법으로는 실제 이상치를 탐지하기 어렵다. 방사 공정 데이터의 주요 공정 변수에 대한 다수/소수 구간 데이터 비율은 Table 2에 나타낸다.

특히, 주요 공정 변수 중 롤러2 온도는 다수 구간 데이터의 비율이 92%로 나타날 정도로 불균형이 심각하기 때문에 일반적인 이상치 처리 기법을 적용하기에 적합하지 않다.

Table 2. Key Process Variable Data Unbalanced Ratio

	Spin beam temperature	Roller 1 speed	Roller 2 speed	Roller 2 temperature	F/R speed
Major Ratio	65%	88%	76%	92%	82%
Minor Ratio	35%	12%	24%	8%	18%

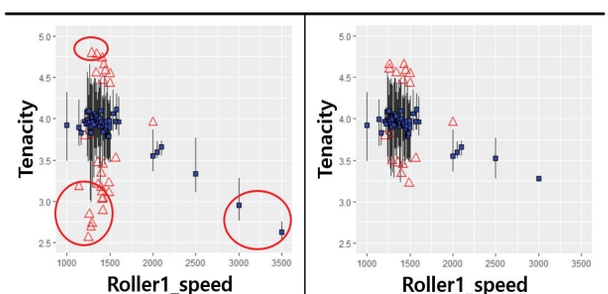
1) 박스 플롯 기반 이상치 처리 기법

박스 플롯 기반 이상치 처리 기법은 데이터를 사분위수를 통해 계산한 일정 범위 밖에 있는 값들을 모두 이상치로 판단하여 제거하는 기법이다. 박스 플롯 기반 이상치 처리를 적용 전후를 비교하는 박스 플롯들은 Fig. 1에 나타낸다. 박스 플롯 기반 이상치 처리 기법 적용 결과로 빨간 동그라미로 표시된 인장 강도 약 3.2 이하의 데이터와 약 4.7 이상의 데이터들이 모두 제거된 것을 확인할 수 있다. 제거된 데이터에는 강도 소수 구간에 속하는 정상적인 데이터들이 포함되어 있기 때문에 박스 플롯 기반 이상치 처리 기법은 불균형 데이터인 방사 공정 데이터에 적합하지 않다.

2) 클러스터링 기반 이상치 처리

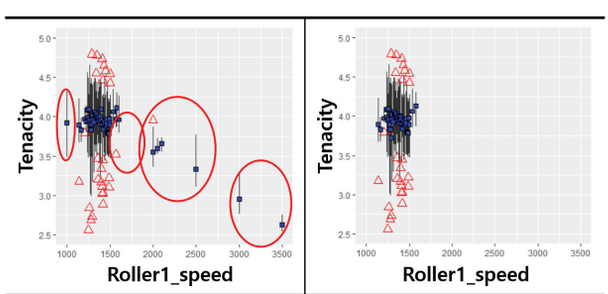
CBLOF(Clustering Based Local Outlier Factor) 기법은 다양한 클러스터링 기반 이상치 처리 기법 중 하나다[8]. CBLOF 기법은 클러스터링을 통해 입력 데이터를 작은 클러스터와 큰 클러스터로 나눈다. 그리고 작은 클러스터에 속하지 않으면서 큰 클러스터와의 거리가 일정 거리 이상인 데이터를 이상치로 분류한다. CBLOF 기법 적용 전후를 비교하는 박스 플롯들은 Fig. 2에 나타낸다.

빨간 동그라미로 표시된 것은 클러스터에 속하지 못하여 제거된 데이터들이다. 정상적으로 공정 변수 소수 구간에서 수집된 데이터들이 CBLOF 기법 적용 결과 모두 이상치로 분



(A) Original data | (B) Data after box plot outlier handling

Fig. 1. Comparison of Box Plots after Box Plot Outlier Handling



(a) Original data | (b) Data after CBLOF outlier handling

Fig. 2. Comparison of Box Plots after CBLOF Outlier Handling

류되어 제거된 것을 확인할 수 있다. 클러스터링 기반 이상치 처리 기법은 박스 플롯 기반 이상치 처리와는 달리 강도뿐만이 아니라 공정 변수까지 고려하여 이상치를 탐지한다. 그러나 두 기법 모두 공통으로 정상적인 소수 구간 데이터들을 제거한다는 문제가 있다.

3) 방사 공정데이터 특성을 고려한 거리 기준 이상치 처리

생분해성 섬유 방사 공정 데이터가 동일 방사 조건에 대해 최대 4개의 샘플을 가지고 있기 때문에 전체 데이터를 동일한 방사 조건을 가진 클러스터로 나눈다. 그리고 각 클러스터의 평균을 계산하고 각 데이터의 평균까지 거리가 이상치 판단 거리 기준 이상이면 이상치로 판단한다[9]. 제안하는 군집 평균 대비 거리 기준에 따른 이상치 탐지 기법은 Fig. 3에 나타낸다. 제안한 기법은 전체 데이터에 대해 이상치를 탐지하는 것이 아니라 데이터 특성을 고려하여 나뉜 각 데이터 군집 내에서 이상치를 판단한다. 이는 전체 데이터의 불균형에 영향을 받지 않고 소수 구간 데이터에서도 방사 및 측정 오류로 인한 이상치 탐지를 가능하게 한다.

이상치 판단 거리 기준 실험은 각 거리 기준에서 10번의 서로 다른 무작위 추출로 테스트 셋을 추출하고 각 테스트 셋에서 10번의 테스트를 실시하여 거리 기준 당100개에 달하는 MAE 값들의 평균을 각 거리 기준의 MAE로 표기하였다. 다양한 이상치 판단 거리 기준에 따라 분류된 이상치 수와 비중, 해당 이상치가 제거된 데이터의 강도 예측 모델 적용 결과 평균절대오차는 Table 3에 나타낸다.

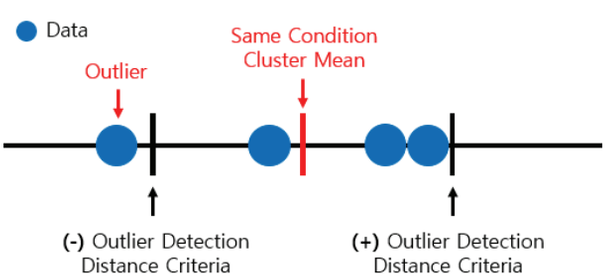
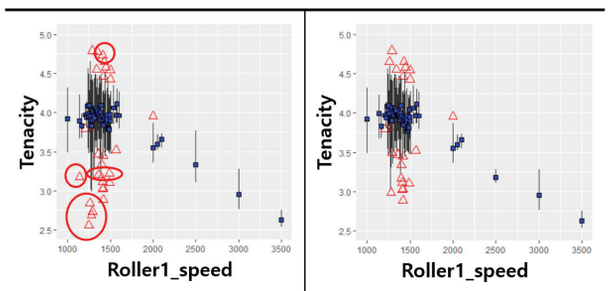


Fig. 3. Outlier Detection based on Cluster Mean-to-Distance Criteria

Table 3. Results by Each Distance Criteria

Distance criteria	# of data	# of outlier	Outlier ratio	Mean Absolute Error
-	816	0	0.0%	0.165
0.4 or more	796	20	2.5%	0.148
0.5 or more	806	10	1.2%	0.157
0.6 or more	811	5	0.6%	0.160
0.7 or more	813	3	0.4%	0.163
0.8 or more	814	2	0.2%	0.166
0.9 or more	815	1	0.1%	0.165



(a) Original data | (b) Data after 0.4 or more outlier handling

Fig. 4. Comparison of Box Plots after Outlier Handling based on Mean-to-Distance Criteria

섬유 방사 산업 분야에서 적용하는 공정관리한계 허용오차의 절대치가 0.3이고, 강도의 범위가 2.5에서 5.0이기 때문에 허용오차인 0.3을 기준으로 0.1씩 증가시키며 이상치를 판단 기준을 변화시킨다. 그러나 0.3 기준으로 분류되는 데이터들은 동일 방사 조건의 다른 샘플들과 차이가 크지 않은 경우가 많으며 이상치 개수가 43개로 전체 데이터 대비 비중이 5%라서 큰 편이다. 방사 조건별로 축적 및 확보된 데이터가 매우 적음을 고려했을 때 0.3 기준으로 분류된 데이터들은 추가 데이터 확보에 따라 이상치로 분류되지 않을 가능성이 있다고 판단하여 0.3은 이상치 거리 기준에서 제외한다. 제외한 이상치 거리 기준 중 가장 높은 성능을 보이는 것은 0.4이다. 따라서 이상치 처리를 위한 거리 기준은 0.4가 가장 적합하다고 판단된다.

원본 데이터와 0.4 기준 이상치 처리를 적용한 박스 플롯은 Fig. 4에 나타낸다. 세모는 박스 플롯 기반 이상치 처리에서 이상치로 의심되는 데이터이며 빨간 동그라미로 표시된 일부 데이터가 제안한 이상치 처리 기법 적용 후 제거되었음을 확인할 수 있다. 이상치 처리 후에도 남아 있는 이상치 의심 데이터들은 소수 구간의 데이터이거나 0.3 수준의 차이로 인해 이상치로 오판단된 데이터들이다. 이상치 처리 전후 결과를 비교해보면 데이터 분포에서 많이 동떨어진 데이터들은 제거된 것을 확인할 수 있다.

4) 이상치 처리 기법 비교

각 이상치 처리 기법 적용 결과들을 Fig. 5에서 나타낸다. 박스 플롯 기반 이상치 처리 기법은 특정 강도 범위를 벗어난 정상적인 강도 소수 구간 데이터를 모두 제거하여 데이터 소실을 보인다. CBLOF 기법도 비슷하게 정상적인 공정 변수 소수 구간 데이터들을 클러스터에 속하지 않다는 이유로 모두 제거하여 데이터 소실을 보인다. 하지만 제안한 방사 공정 데이터 특성을 고려한 이상치 처리 기법이 전체적인 데이터 양상을 유지하면서 심각한 데이터 소실을 보이지 않고 이상치를 적절히 제거하는 것을 확인할 수 있다. 따라서 방사 공정 데이터에는 방사 공정 데이터 특성을 고려한 이상치 처리 기법이 가장 적합하다고 판단된다.

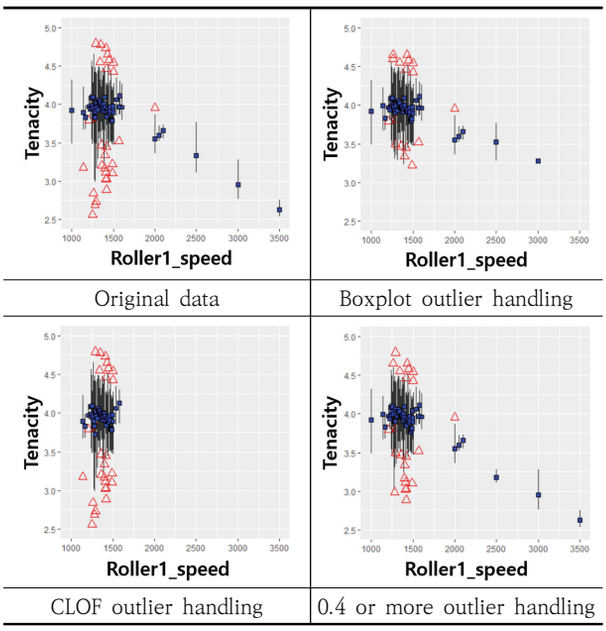


Fig. 5. Comparison of Box Plots after Each Outlier Handling

3.3 다양한 데이터 증강 기법 적용 및 비교

3.2절에서 최종적으로 20개의 이상치를 처리하여 796개로 줄어든 데이터는 여전히 데이터 불균형 문제와 데이터 부족 문제에 대한 해결이 필요하다. 이런 문제들은 데이터 증강을 통해 완화될 수 있다. 그러나 여러 변수에서 여러 구간의 데이터 불균형이 복합적으로 나타나는 경우 고려해야 할 점들이 있다. 특정 변수의 소수 구간 데이터가 다른 변수에서의 다수 구간에 속하는 데이터인 경우 데이터 증강 시 다른 변수에서 데이터 불균형이 심해질 수 있다는 점과 소수 구간 내에 존재하는 데이터 불균형을 고려해야 한다는 점이다. 일반적인 데이터 증강 기법은 이러한 점들을 고려하여 균형 있게 데이터를 증강하기 어렵다.

따라서 각 주요 공정 변수의 데이터 불균형 정도와 각 변수의 인장 강신도에 대한 상관 계수를 고려해 균형 있는 데이터 증강 기법을 적용할 필요가 있다.

1) 랜덤 오버샘플링 기법

랜덤 오버샘플링 기법(Random OverSampling; ROS)은 소수 클래스 데이터를 단순 복제하여 데이터를 증강하는 기법이다. 방사 공정 데이터 중 가장 데이터 불균형이 심한 롤러2 온도를 기준으로 랜덤 오버샘플링을 적용한 결과에 대한 주요 공정 변수들의 불균형 완화 정도는 Table 4에 나타낸다. ROS 기법 적용 결과 전체적으로 데이터 불균형이 완화되었지만 스핀 빔 온도에서는 다수 구간 비율이 82%로 데이터 불균형이 심해졌다. 이는 앞서 언급하였듯이 생분해성 섬유 방사 공정 데이터는 여러 변수에서 복합적으로 데이터 불균형이 나타나기 때문에 특정 변수에서 소수 구간 데이터를 증강하여도 다른 변수에서 불균형이 심해질 수 있기 때문이다.

Table 4. Before and After Comparison with ROS

	Spin beam temperature	Roller 1 speed	Roller 2 speed	Roller 2 temperature	F/R speed
Major ratio (before)	67%	89%	76%	92%	82%
Major ratio (after)	82%	58%	41%	50%	75%
Minor ratio (before)	33%	11%	24%	8%	18%
Minor ratio (after)	18%	42%	59%	50%	25%

Table 5. Before and after Comparison with SMOTE

	Spin beam temperature	Roller 1 speed	Roller 2 speed	Roller 2 temperature	F/R speed
Major ratio (before)	67%	89%	76%	92%	82%
Major ratio (after)	82%	58%	41%	50%	75%
Minor ratio (before)	33%	11%	24%	8%	18%
Minor ratio (after)	18%	42%	59%	50%	25%

2) SMOTE 데이터 증강 기법

SMOTE(Synthetic Minority OverSampling Technique) 기법은 임의의 소수 클래스 데이터에서 가장 가까운 이웃 사이에 새로운 데이터를 생성하여 데이터를 증강하는 기법이다[10]. 방사 공정 데이터 중 가장 데이터 불균형이 심한 롤러2 온도를 기준으로 SMOTE 기법을 적용한 결과 주요 공정 변수들의 불균형 완화 정도는 Table 5에 나타낸다. SMOTE 기법 적용 결과는 ROS 기법 적용 결과와 동일하게 나타난다.

3) ROS 기법 및 SMOTE 기법의 취약점

앞선 두 가지 데이터 증강 기법 적용 결과에서 다수 구간 데이터 비율과 소수 구간 데이터 비율을 보아 전체적으로 데이터 불균형 문제가 해소된 것으로 보인다. 그러나 방사 공정 데이터가 실제로는 다수 구간과 소수 구간으로만 나뉘지 않는다. 실제로는 각 변수에서 여러 구간이 존재한다. 표에서는 데이터 불균형 비율을 알아보기 쉽도록 다수 구간 데이터와 소수 구간 데이터로 오직 두 가지로만 나누었지만 소수 구간 안에 존재하는 여러 구간별 데이터 불균형이 나타난다. 이를 고려하지 않고 데이터 증강을 할 경우 소수 구간 안에 존재하는 데이터 불균형은 그대로 유지된다. 두 기법에서 데이터 증강 기준으로 한 롤러2 온도에 대한 소수 구간에서의 데이터 불균형 정도 비교는 Table 6에 나타낸다.

95°C와 105°C에 속하는 데이터 비율이 ROS와 SMOTE 모두 각각 44.3%와 5.7%로 소수 구간 내 약 7:1의 데이터

Table 6. Data Imbalance Comparison in Minor Data Among Original Data, ROS and SMOTE

Data section in minor data for Roller2_temperature	Original data	Data after ROS	Data after SMOTE
95°C	6.9%	44.3%	44.3%
105°C	1.1%	5.7%	5.7%

불균형이 남아있음을 확인할 수 있다. 이처럼 여러 가지 변수에서 동시에 여러 구간의 데이터 불균형이 나타나는 경우, 일반적인 데이터 증강 기법을 적용하면 소수 구간에서의 데이터 불균형이 해소되지 않는 문제가 발생한다.

4) 데이터 특성을 고려한 복합 데이터 증강 기법 및 비교

여러 변수의 여러 구간에서 나타나는 데이터 불균형을 해소하기 위해 각 변수의 강도에 대한 상관계수와 데이터 불균형 정도를 고려하여 복합적으로 데이터를 증강한다[9]. Table 7은 주요 공정 변수와 인장 강신도 간 상관분석 결과인 각 변수의 상관계수와 불균형 정도, 우선순위, 증강 비율을 나타낸다.

상관계수의 크기와 데이터 불균형 정도를 고려해서 우선순위를 결정하고 우선순위에 따라 각각 다른 증강 비율로 소수 구간 데이터를 증강한다. 데이터 증강 결과 1,193개로 늘어났으며 주요 공정 변수들의 불균형 완화 정도는 Table 8에 나타낸다. 또 롤러2 온도에 대한 소수 구간에서의 데이터 불균형 정도는 Table 9에 나타낸다.

Table 7. Priority and Augmentation Ratio Considering Correlation Coefficient and Imbalance for Tenacity

	Spin beam temperature	Roller 1 speed	Roller 2 speed	Roller 2 temperature	F/R speed
Corr.	-0.51	-0.42	0.13	0.33	0.09
Imbalance (major, minor)	67, 33	89, 11	76, 24	92, 8	82, 18
Priority	3rd	1st	2nd	1st	2nd
Augmen-tation ratio	-	3	2	3	2

Table 8. Before and After Comparison with Complex Data Augmentation

	Spin beam temperature	Roller 1 speed	Roller 2 speed	Roller 2 temperature	F/R speed
Major ratio (before)	67%	89%	76%	92%	82%
Major ratio (after)	77%	68%	50%	68%	68%
Minor ratio (before)	33%	11%	24%	8%	18%
Minor ratio (after)	23%	32%	50%	32%	32%

Table 9. Data Imbalance Comparison in Minor Data Among Original Data, Complex Data Augmentation

Data section in minor data for Roller2_temperature	Original data	Data after complex data augmentation
95°C	6.9%	19.2%
105°C	1.1%	12.8%

전체적으로 데이터 불균형이 감소하고 95°C와 105°C의 데이터 비율이 각각 19.2%와 12.8%로 약 3:2로 소수 구간 내 데이터 불균형이 크게 완화되었음을 확인할 수 있다.

ROS 기법과 SMOTE 기법은 둘 다 불균형이 제일 심한 변수에 대해 적용한 결과 전체적으로 데이터 불균형이 완화되었지만 가장 강한 상관계수를 가진 스핀빔 온도의 다수 구간 비율이 82%로 심해졌다. 반면에 각 변수의 상관 계수와 데이터 불균형 정도를 고려한 복합 데이터 증강 기법의 경우 스핀빔 온도의 다수 구간 비율이 77%로 비교적 낮고 전체적으로 데이터 불균형이 완화되었다. 또한 소수 구간에서의 데이터 불균형을 비교해보면 ROS 기법과 SMOTE 기법 적용 결과의 불균형 비율인 7:1에 비해 복합 데이터 증강의 경우 3:2로 크게 개선된 것을 확인할 수 있다. 따라서 데이터 특성을 고려한 복합 데이터 증강 기법이 방사 공정 데이터에 가장 적합하다고 판단된다.

3.4 제안하는 데이터 처리 기법 : 인장 신도 기준

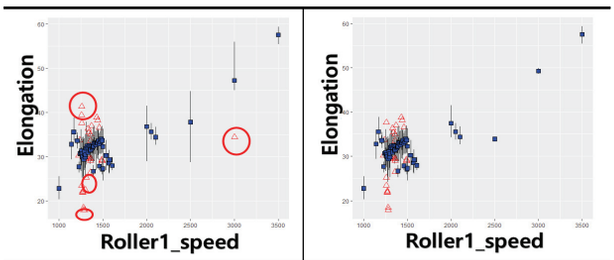
3.2절과 3.3절에서는 인장 강도를 기준으로 다양한 데이터 처리 기법을 적용해보고 제안한 기법이 가장 적합함을 보였다. 이번 절에서는 가장 적합하다고 판단되는 제안하는 기법을 인장 신도를 기준으로 적용하고 그 결과를 보인다.

Table 10은 인장 신도를 기준으로 제안하는 이상치 처리 기법을 적용한 결과들을 나타낸다.

섬유 방사 산업 분야에서 적용하는 인장 신도의 공정관리 한계 허용오차 절대치가 3.5이고, 신도의 범위가 대략 18부터 60까지 이기 때문에 허용오차인 3.5를 기준으로 1씩 증가시키며 이상치 판단 기준을 변화시킨다. 그러나 3.5 기준으

Table 10. Results by Each Distance Criteria

Distance criteria	# of data	# of outlier	Outlier ratio	Mean Absolute Error
-	816	0	0.0%	1.860
4.5 or more	796	20	2.5%	1.681
5.5 or more	803	13	1.6%	1.707
6.5 or more	808	8	1.0%	1.754
7.5 or more	812	4	0.5%	1.789
8.5 or more				
9.5 or more	815	1	0.1%	1.865



(a) Original data | (b) Data after 4.5 or more outlier handling

Fig. 6. Comparison of Box Plots after Outlier Handling based on Mean-to-Distance Criteria

로 분류되는 데이터들은 동일 방사 조건의 다른 샘플들과 차이가 크지 않은 경우가 많으며 이상치 개수가 37개로 전체 데이터 대비 비중이 5%라서 큰 편이다. 방사 조건별로 축적 및 확보된 데이터가 매우 적음을 고려했을 때 3.5 기준으로 분류된 데이터들은 추가 데이터 확보에 따라 이상치로 분류되지 않을 가능성이 있다고 판단하여 3.5는 이상치 거리 기준에서 제외한다. 제외한 이상치 거리 기준 중 가장 높은 성능을 보이는 것은 4.5이다. 따라서 이상치 처리를 위한 거리 기준은 4.5가 가장 적합하다고 판단된다. 원본 데이터와 4.5 기준 이상치 처리를 적용한 데이터의 박스플롯 비교는 Fig. 6에 나타낸다.

세모는 박스 플롯 기반 이상치 처리에서 이상치로 의심되는 데이터이며 빨간 동그라미로 표시된 일부 데이터가 제안한 이상치 처리 기법 적용 후 제거되었음을 확인할 수 있다. 이상치 처리 후에도 남아있는 이상치 의심 데이터들은 소수 구간의 데이터이거나 3.5 수준의 차이로 인해 이상치로 오판단된 데이터들이다. 이상치 처리 전후 결과를 비교해보면 전체적인 데이터 양상을 유지하면서 심각한 데이터 소실을 보이지 않고 데이터 분포에서 많이 동떨어진 데이터들은 제거된 것을 확인할 수 있다.

앞선 이상치 처리로 인해 20개의 이상치가 제거되어 796개의 데이터가 존재한다. 이 데이터를 대상으로 제안하는 복합 데이터 증강 기법을 적용한다.

Table 11은 주요 공정 변수와 인장 신도 간 상관분석 결과인 각 변수의 상관계수와 불균형 정도, 우선순위, 증강 비율을 나타낸다.

상관계수의 크기와 데이터 불균형 정도를 고려해서 우선순위를 결정하고 우선순위에 따라 각각 다른 증강 비율로 소수 구간 데이터를 증강한다. 데이터 증강 결과 1,159개로 늘어났으며 주요 공정 변수들의 불균형 완화 정도는 Table 12에 나타낸다.

전체적으로 데이터 불균형이 감소하였음을 확인할 수 있다. 이렇듯 인장 강도뿐만 아니라 인장 신도를 대상으로 제안하는 데이터 처리 기법을 적용하였을 때도 적절한 이상치 제거와 데이터 증강이 이루어짐을 확인할 수 있다.

Table 11. Priority and Augmentation Ratio Considering Correlation Coefficient and Imbalance for Elongation

	Spin beam temperature	Roller 1 speed	Roller 2 speed	Roller 2 temperature	F/R speed
Corr.	-0.08	0.59	-0.25	-0.20	-0.24
Imbalance (major, minor)	67, 33	89, 11	76, 24	92, 8	82, 18
Priority	3rd	1st	2nd	1st	2nd
Augmentation ratio	-	3	2	3	2

Table 12. Before and After Comparison with Complex Data Augmentation

	Spin beam temperature	Roller 1 speed	Roller 2 speed	Roller 2 temperature	F/R speed
Major ratio (before)	67%	89%	76%	92%	82%
Major ratio (after)	77%	70%	53%	71%	67%
Minor ratio (before)	33%	11%	24%	8%	18%
Minor ratio (after)	23%	30%	47%	29%	33%

4. 다양한 모델 적용 및 비교

제안한 기법들을 적용한 결과 인장 강신도 예측 모델의 성능 개선 정도를 확인하고 어떤 모델이 가장 적합한지 알기 위해 기본 데이터, 이상치 처리 데이터, 이상치 처리 후 데이터 증강을 적용한 데이터로 인장 강신도 예측 모델을 학습시키고 모델 성능을 비교한다. 학습용 데이터와 테스트 데이터의 비율은 9:1이며 모델 학습을 위해 정규화 처리를 한다. 또 과적합을 방지하고 더 많은 데이터를 학습 및 검증에 활용하기 위해 학습용 데이터에 5-fold 교차검증을 수행한다. 모델 성능 비교를 위한 성능 지표는 조정된 결정 계수와 평균절대오차, 평균제곱오차, 허용오차 내 예측치 개수와 비율을 사용한다. 조정된 결정계수는 피팅된 모델 기준이며 평균절대오차 및 평균제곱오차의 경우 테스트 데이터에 대한 예측 성능이다.

4.1 Polynomial 모델

Polynomial 모델은 종속 변수와 독립 변수 간 비선형 관계를 분석하기 위한 모델이다. 예측 대상 물성인 인장 강신도가 종속 변수에 해당하며 주요 공정 변수는 독립변수에 해당한다. 각 데이터에 대한 인장 강신도 Polynomial 예측 모델의 성능 비교 결과는 Table 13에 나타낸다.

4.2 KNN 모델

KNN 모델은 일반적으로 클러스터링으로 사용되지만 주어진 입력과 가장 가까운 N개의 평균을 내는 방식으로 회귀

모델을 구현할 수 있다. K값을 변경하며 테스트한 결과 K가 1일 때 가장 좋은 성능을 보인다. 각 데이터에 대한 인장 강신도 KNN 예측 모델의 성능 비교 결과는 Table 14에 나타낸다.

Table 13. Polynomial Model Performance Comparison

Prediction target	Data processed for prediction target	Adjusted R-square	Mean Absolute Error	Mean Square Error	# of data within allowable error (ratio to test data)
Tenacity	Original data	0.511	0.161	0.050	72 (87.5%)
	After outlier handling	0.571	0.148	0.038	71 (89.2%)
	After data augmentation	0.760	0.139	0.040	109 (89.7%)
	After outlier handling & data augmentation	0.811	0.128	0.030	109 (91.1%)
Elongation	Original data	0.469	2.016	7.620	70 (84.9%)
	After outlier handling	0.491	1.870	6.169	70 (87.0%)
	After data augmentation	0.772	2.086	8.727	103 (84.4%)
	After outlier handling & data augmentation	0.814	1.810	5.750	102 (88.2%)

Table 14. KNN Model Performance Comparison

Prediction target	Data processed for prediction target	Adjusted R-square	Mean Absolute Error	Mean Square Error	# of data within allowable error (ratio to test data)
Tenacity	Original data	0.584	0.190	0.071	67 (81.7%)
	After outlier handling	0.610	0.169	0.047	65 (81.3%)
	After data augmentation	0.806	0.137	0.037	107 (87.7%)
	After outlier handling & data augmentation	0.811	0.117	0.023	114 (95.0%)
Elongation	Original data	0.730	1.969	7.230	71 (86.6%)
	After outlier handling	0.851	1.654	4.727	73 (91.3%)
	After data augmentation	0.836	1.452	3.654	110 (90.2%)
	After outlier handling & data augmentation	0.864	1.238	2.434	112 (96.6%)

4.3 MLP 모델

MLP 모델은 퍼셉트론들을 여러 층으로 중첩하여 입력데이터로부터 결괏값을 추출하는 모델이다. 사용한 MLP 모델의 구체적인 구성 정보는 Table 15에 나타낸다. 각 데이터에 대한 인장 강신도 MLP 예측 모델의 성능 비교 결과는 Table 16에 나타낸다.

각 데이터 처리 여부 별 데이터에 대한 공정관리한계 허용오차 이내 개수와 비율을 보면 허용오차 이내에 있는 데이터의 개수와 비율이 점점 증가함을 확인할 수 있다. 이를 통해 모델의 성능이 방사 공정에서 요구되는 성능에 적합하게 개선되었음을 확인할 수 있다.

4.4 모델 성능 비교

모델들의 평가 지표들을 비교한 결과 인장 강도의 경우 모델 간 큰 성능 차이는 나타나지 않는다. 이는 현재 수집된 데이터 내에서 데이터가 선형이거나 단순하게 분포되어 있기 때문으로 판단된다. 그러나 인장 신도의 경우 KNN 모델에서 다른 모델보다 명확하게 더 좋은 성능을 보인다. 이는 데이터가 일반화하기에 어려울 정도로 복잡하게 분포되어 있어 K가 1인 KNN 모델이 과적합으로 인해 더 높은 성능을 내는 것으

Table 15. MLP Model Information

# of Hidden layer	# of layer unit	Activation function	Optimizer	Cross Validation
2	64 / 64	ReLU	Adam	5-Fold

Table 16. MLP Model Performance Comparison

Prediction target	Data processed for prediction target	Adjusted R-square	Mean Absolute Error	Mean Square Error	# of data within allowable error (ratio to test data)
Tenacity	Original data	0.479	0.165	0.047	70 (85.4%)
	After outlier handling	0.496	0.148	0.045	72 (89.4%)
	After data augmentation	0.782	0.132	0.031	110 (91.6%)
	After outlier handling & data augmentation	0.789	0.120	0.027	101 (92.4%)
Elongation	Original data	0.494	1.860	6.686	73 (88.8%)
	After outlier handling	0.519	1.681	4.836	73 (91.4%)
	After data augmentation	0.796	1.813	6.921	110 (89.8%)
	After outlier handling & data augmentation	0.824	1.677	5.329	106 (91.0%)

로 보인다. MLP 모델 기준 평균절대오차 기준으로 인장 강도는 약 0.1 수준 인장 신도는 약 1.6 수준의 오차를 보인다. 이는 인장 강도와 인장 신도의 단위 및 범위를 고려할 때 측정값들의 10% 미만의 오차 수준이다.

모든 모델에서 공통으로 데이터 증강을 하였을 때 조정된 결정계수의 개선을 보인다. 또 이상치 제거 시 MAE와 MSE에서 약간의 개선을 보인다. 그러나 이상치 처리와 함께 데이터 증강까지 적용한 데이터의 경우 유의미한 오차 개선과 결정계수 개선을 동시에 보인다. 이런 개선 결과를 보아 방사 공정 데이터 특성을 고려한 데이터 처리 기법들을 통해 인장 강신도 예측 모델의 성능이 개선됨을 확인할 수 있다.

5. 결 론

본 논문에서는 생분해성 섬유 방사 공정 데이터에서 나타나는 데이터 부족, 데이터 불균형, 샘플 간 오차 등의 특성에서 인해 발생할 수 있는 과적합 및 모델 성능 저하 문제를 해결하기 위해 이상치 처리 기법 및 데이터 증강 기법을 제안한다. 또 제안한 이상치 처리 기법과 데이터 증강 기법을 기존 기법들과 비교하여 제안한 기법이 방사 공정 데이터에 더욱 적합함을 보인다. 제안한 기법으로 처리된 데이터를 여러 가지 모델에 적용한 결과, 제안한 데이터 처리 기법에 의해 모델 성능이 개선되었음을 보인다. 구체적으로는 MLP 모델에서 제안한 데이터 처리에 의해 주 예측 물성인 인장 강도를 기준으로 평균절대오차는 약 27% 줄고 평균제곱오차는 약 43% 줄어드는 등의 큰 개선을 보인다. 또 0.5 미만이었던 결정계수가 1에 가까운 약 0.8로 크게 개선된다. 또 예측치와 실제치의 절대오차가 공정관리한계 허용오차 내에 들어오는 데이터의 비율도 85.4%에서 92.4%로 증가하였다. 이는 인장 강신도 예측 오차가 감소하고 예측 모델이 데이터에 더 적합해졌으며 공정에서 요구하는 정확도에 부합하는 예측치들의 비율이 증가했음을 의미한다. 본 논문에서 제안한 데이터 처리 기법들과 같이 인공지능 모델 개발 시에 데이터 특성이 고려된 적절한 데이터 처리 기법을 적용한다면 모델 성능 개선에 도움이 될 것이다.

References

[1] Z. Zhou, W. Deng, Z. Zhu, Y. Wang, J. Du, and X. Liu, “Fabric defect detection based on feature fusion of a convolutional neural network and optimized extreme learning machine,” *Textile Research Journal*, Vol.92, No.7-8, pp.1161-1182, 2022.

[2] K. H. Cho and S. H. Jeong, “The statistical approach for determining the parallel-bundle strength from single-filament data of PET,” *Proceeding of Korean Fiber Society*, pp.291-29, 2003.

[3] Y. Huh and J. S. Kim, “Experimental modeling of spinning tension behavior during the cops building on the ring spinner,” *Textile Science and Engineering*, Vol.39, No.2, pp.209-216, 2002.

[4] B. S. Jeon and C. G. Yang, “Prediction of cotton yarn strength using regression analysis and neural networks,” *Textile Science and Engineering*, Vol.34, No.11, pp.731-738, 1997.

[5] C. H. Park, T. G. Kim, J. Kim, S. Choi, and K. H. Lee, “Outlier detection by clustering-based ensemble model construction,” *KIPS Transactions on Software and Data Engineering*, Vol.7, No.11, pp.435-442, 2018.

[6] J. Park, G. Ahn, and S. Hur, “Oversampling based on k-NN and GAN for effective classification of class imbalance dataset,” *Journal of the Korean Institute of Industrial Engineers*, Vol.46, No.4, pp.365-371, 2020.

[7] S. Kim and S. Kim, “Recursive oversampling method for improving classification performance of class unbalanced data in patent document automatic classification,” *Journal of the Institute of Electronics and Information Engineers*, Vol.58, No.4, pp.43-49, 2021.

[8] Z. He, X. Xu, and S. Deng, “Discovering cluster-based local outliers,” *Pattern Recognition Letters*, Vol.24, No.9-10, pp.1641-1650, 2003.

[9] S. C. Park, D. Y. Kim, K. B. Seo, and W. J. Lee, “The development of property prediction model in consideration of biodegradable fiber spinning process data characteristics,” *Proceedings of the Annual Spring Conference of Korea Information Processing Society Conference (KIPS) 2022*, Vol.29, No.1, pp.362-365, 2022.

[10] 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.

박 세 찬

https://orcid.org/0000-0001-6614-7994
e-mail : hasmi5452@gmail.com
2022년 경북대학교 컴퓨터학부(학사)
2022년 ~ 현 재 경북대학교 컴퓨터학부 석사과정
관심분야 : Game AI (Game Tree, Reinforce Learning) & Simulation

김 덕 엽

https://orcid.org/0000-0003-1680-1278
e-mail : ejrduq77@naver.com
2016년 경북대학교 컴퓨터학부(학사)
2018년 경북대학교 컴퓨터학부(석사)
2018년 ~ 현 재 경북대학교 컴퓨터학부 박사과정
관심분야 : Software Testing & Data Science & Computer Science Education

서 강 복

https://orcid.org/0000-0003-3716-700X
e-mail : dating1227@gmail.com
2014년 국가평생교육진흥원 학점은행제 컴퓨터공학(학사)
2017년 경북대학교 컴퓨터학부(석사)
2018년 ~ 현 재 경북대학교 컴퓨터학부 박사과정
관심분야 : Software Testing & Software Testing Automation

이 우 진

https://orcid.org/0000-0002-8075-5248
e-mail : woojin@knu.ac.kr
1992년 경북대학교 컴퓨터학부(학사)
1994년 KAIST 전산학과(석사)
1999년 KAIST 전산학과(박사)
1999년~2002년 ETRI 선임연구원
2002년 ~ 현 재 경북대학교 컴퓨터학부 교수
관심분야 : Software Testing & Requirements Engineering & Embedded Systems