

생분해성 섬유 방사 공정 데이터 특성을 고려한 물성 예측 모델 개발

박세찬⁰, 김덕엽, 서강복, 이우진
경북대학교 컴퓨터학부

hasmi5452@gmail.com⁰, ejrduq77@naver.com, dating1227@gmail.com, woojin@knu.ac.kr

The Development of Property Prediction Model in Consideration of Biodegradable Fiber Spinning Process Data Characteristics

SeChan Park⁰, Deok Yeop Kim, Kang Bok Seo, Woo Jin Lee
Kyungpook National University

요 약

최근 노동 집약적인 성격의 섬유 산업에서는 AI를 통해 공정에 들어가는 시간과 비용을 줄이고 품질을 최적화 하려는 시도를 하고 있다. 그러나 섬유 방사 공정은 데이터 수집에 필요한 비용이 크고 체계적인 데이터 처리 시스템이 부족하여 축적된 데이터양이 적다. 또 방사 목적에 따라 특정 변수 위주의 조합에 대한 데이터만을 우선적으로 수집하여 데이터 불균형이 발생하며, 물성 측정환경 차이로 인해 동일 방사조건에서 수집된 샘플 간에도 오차가 존재한다. 이러한 데이터 특성들을 고려하지 않고 AI 모델에 활용할 경우 과적합과 성능 저하 등의 문제가 발생할 수 있다. 따라서 본 논문에서는 물성 단위 및 허용오차를 고려한 이상치 처리 기법과 데이터 불균형 정도 및 물성과의 상관성을 고려한 오버샘플링 기법을 물성 예측 모델에 적용한다. 두 기법들을 모델에 적용한 결과 그렇지 않은 모델에 비해 물성 예측 오차와 방사 공정 데이터에 대한 모델의 적합도가 개선됨을 보인다.

1. 서론

최근 노동 집약적인 성격의 섬유 산업은 AI를 적용하려는 시도[1]를 하고 있다. AI를 적용하면 시간과 비용을 줄일 수 있고 자동화 및 최적화를 통해 제품 품질을 향상시킬 수 있기 때문이다. 일반적으로 AI 모델의 성능은 데이터의 양과 질의 영향을 받는다. 하지만 섬유 방사 분야에서는 AI 모델에 활용할 수 있는 충분한 데이터의 양과 질을 확보하는데 문제가 있다. 먼저 섬유 방사 공정 데이터를 수집하기 위해 소요되는 시간과 비용이 크기 때문에 다양한 조합의 방사 공정 데이터들을 확보하기 어렵다. 그리고 체계적인 데이터 수집 및 관리 시스템이 부족하여 축적된 데이터의 양이 적다. 또 방사 목적에 따라 특정 방사 공정 변수들 위주의 데이터를 수집하기 때문에 축적된 데이터의 분포가 불균형하다. 이 외에도 물성 측정환경의 차이 및 공정관리한계로 인해 동일 방사 조건에서 수집된 데이터 샘플 간에도 물성 값의 차이가 나타난다. 이러한 적은 데이터 양, 데이터 불균형, 샘플 간 오차 등의 데이터 특성들을 고려하지 않을 경우 섬유 물성 예측 모델에서 과

적합, 모델 성능 저하 등의 문제가 발생할 수 있다.

따라서 본 논문에서는 섬유 방사 공정 데이터 특성을 반영하여 예측 대상 물성의 단위와 범위 및 허용오차를 고려한 이상치 처리 기법과 데이터 불균형 정도와 물성과의 상관성을 고려한 오버샘플링 기법을 제안한다. 또 제안한 데이터 처리 기법들을 섬유 물성 예측 모델에 적용하여 물성 예측 오차와 데이터에 대한 모델 적합도가 개선되는 것을 보인다.

2. 관련 연구

2.1 기존 섬유 산업의 모델 개발 연구

기존 섬유 산업 연구에서는 품질 개선을 위해 통계적으로 데이터를 분석하거나 수학적 모델링으로 데이터를 분석하였다. 하지만 통계적 분석 방식[2]은 단순한 선형 모델의 경우에는 유용하지만 복잡한 비선형 모델의 경우 적용하기 어렵다는 단점이 있다. 또 수학적 모델링[3]의 경우 관련 수식이 있거나 구조화가 용이할 정도로 모델이 단순한 경우는 유용하지만 그렇지 않은 경우 수학적 모델링 자체를 할 수 없다는 단점이 있다. 이 외에도 30개 정도의 매우

적은 데이터를 AI에 적용시켜 먼방적사의 강도를 예측한 연구[4]도 있다. 이는 이른 시기에 AI를 섬유 산업에 적용시켰다는 의의가 있으나 30개 데이터가 모집단을 대표한다고 보기 어렵기 때문에 이를 통해 만들어진 모델의 신뢰도가 떨어진다.

기존 섬유 산업 연구들이 통계적 분석이나 수학적 모델링이 주가 되고 AI 적용에 대한 사례가 드문 것은 시간과 비용의 문제로 AI 적용에 충분한 데이터양을 확보하기 어렵기 때문이다.

2.2 불균형한 데이터에서의 이상치 처리 기법

데이터 분포가 불균형한 경우에는 일반적인 방법을 사용하면 데이터 분포가 적은 소수 구간 데이터들을 이상치로 잘못 판단할 수 있다. 따라서 측정 오류 및 실험 오류로 인한 이상치를 제대로 처리하기 어렵다. 이를 해결하기 위해 군집화와 앙상블을 기반으로 이상치를 탐지하는 방법[5]이 제안되었다. 제안하는 방법은 정상 데이터가 주어진다 가정 하에 기존 데이터를 먼저 군집화하고 새로운 데이터를 각 군집에 대해서 이상치 탐지를 수행하여 그 결과들을 앙상블 기법으로 종합한다. 이는 충분한 양의 정상 데이터가 기존에 주어진 경우 유용할 수 있지만 기존 데이터의 정상여부를 모르는 경우 군집들로 제대로 나눌 수 없다는 단점이 있다.

2.3 회귀 문제에서의 오버샘플링 기법

데이터 불균형을 해결하기 위해 소수 클래스 데이터를 증강하는 것을 오버샘플링이라고 한다. 대부분의 연구들은 분류 문제에서의 클래스 불균형을 해결하기 위한 오버샘플링 기법들[6-7]을 다룬다. 그러나 회귀 문제에서도 변수의 구간별 데이터 분포의 불균형에 따른 데이터 불균형은 존재한다.

대부분의 오버샘플링 기법들은 불연속적인 클래스의 불균형을 대상으로 하지만 회귀 문제인 섬유 물성 예측은 연속적인 공정 변수의 불균형을 대상으로 하기 때문에 그대로 적용시키긴 어렵다. 또 특정 변수를 오버샘플링 한 후에 오히려 다른 변수들에서의 데이터 불균형이 더 심해질 수 있다는 문제도 있다.

3. 데이터 특성을 고려한 데이터 처리 기법

3.1 생분해성 섬유 방사 공정 데이터

본 논문에서 활용하는 데이터는 816개의 생분해성 섬유 방사 공정 데이터이다. 데이터는 55개의 방사 공정 조건 변수와 1개의 예측 대상 물성인 인장강도

측정값으로 이루어진다. 55개의 방사 공정 조건 변수 중 강도에 영향을 주는 변수 8개 중 단일 값이 아닌 변수인 스핀빔 온도, 롤러1 속도, 롤러2 속도, 롤러2 온도, 권취 속도, 연신비 6개 변수를 주요 공정 변수로 사용한다. 연신비의 경우 롤러1과 롤러2의 속도비이기 때문에 데이터 불균형 비율이나 상관 계수를 구할 때는 제외한다.

3.2 예측 물성 및 허용오차를 고려한 이상치 처리

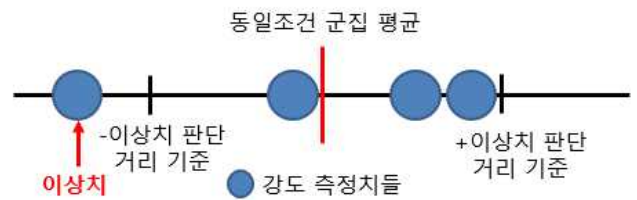
이상치 탐지는 기본적으로 다른 관측값과 상이한 관측값을 찾는 것이기 때문에 데이터 분포가 불균형하면 소수 구간 데이터들이 이상치로 분류될 수 있다. 따라서 데이터양이 적고 불균형한 경우에는 일반적인 방법으로 이상치를 처리하기 어렵다.

<표 1> 주요 공정 변수 데이터 불균형 비율

| 구분 | 스핀빔 온도 | 롤러1 속도 | 롤러2 속도 | 롤러2 온도 | 권취 속도 |
|----|--------|--------|--------|--------|-------|
| 다수 | 65% | 88% | 76% | 92% | 82% |
| 소수 | 35% | 12% | 24% | 8% | 18% |

표 1은 주요 공정 변수에 대한 다수 구간 데이터와 소수 구간 데이터 비율을 나타낸다. 롤러2 온도는 다수 구간 데이터의 비율이 92%로 나타날 정도로 불균형이 심각하기 때문에 일반적인 이상치 처리 기법이 아닌 생분해성 섬유 방사 공정 데이터 특성을 고려한 이상치 처리 기법의 적용이 필요하다.

생분해성 섬유 방사 공정 데이터가 동일 방사 조건에 대해 대부분 4개의 샘플을 가지고 있기 때문에 전체 데이터를 동일한 방사 조건을 가진 군집으로 나누었다. 그리고 각 군집에 대한 평균을 계산하고 각 데이터의 평균까지 거리가 이상치 판단 거리 기준을 초과하면 이상치로 분류하였다.



(그림 1) 평균 대비 거리 기준에 따른 이상치 분류

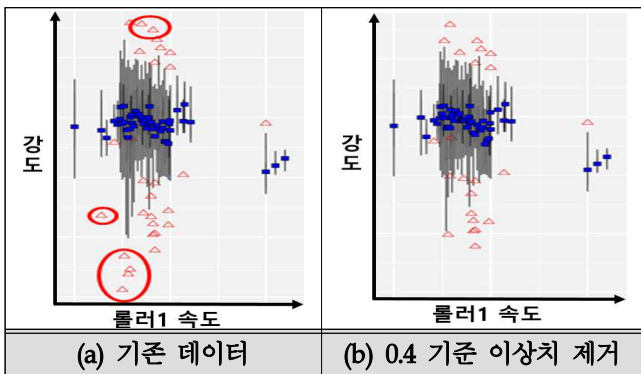
그림 1은 제안한 군집 평균 대비 거리 기준에 따른 이상치 처리 기법을 나타낸다. 이 기법은 전체 데이터에 대해 이상치를 분류하는 것이 아니라 데이터 특성을 고려하여 나뉜 각 데이터 군집 내에서 이

상치를 분류한다. 따라서 전체 데이터의 불균형에 영향을 받지 않고 각 데이터 군집 내에서 방사 및 측정 오류로 인한 이상치를 제대로 분류할 수 있다.

<표 2> 이상치 판단 거리 기준별 적용 결과

| 이상치 판단 거리 기준 | 이상치 수 | 전체 데이터 대비 비중 | 평균절대오차 |
|--------------|-------|--------------|--------|
| - | 0개 | 0.0% | 0.165 |
| 0.3 이상 | 43개 | 5.3% | 0.144 |
| 0.4 이상 | 20개 | 2.5% | 0.148 |
| 0.5 이상 | 10개 | 1.2% | 0.157 |
| 0.6 이상 | 5개 | 0.6% | - |
| 0.7 이상 | 3개 | 0.4% | |
| 0.8 이상 | 2개 | 0.2% | |
| 0.9 이상 | 1개 | 0.1% | |

표 2는 다양한 이상치 판단 거리 기준에 따라 분류된 이상치 수와 비중, 해당 이상치가 제거된 데이터의 강도 예측 모델 적용 결과 평균절대오차를 나타낸다. 섬유 방사 산업 분야에서 적용하는 공정관리한계 허용오차가 0.3이고, 강도가 2.5에서 5.0 범위 내에서 나타나기 때문에 0.3을 기준으로 0.1씩 증가시키며 이상치를 분류해보았다. 그 결과 0.5 이상으로 분류되는 데이터들은 확실하게 이상치로 분류할 수 있었다. 그러나 0.3의 경우 전체 데이터 대비 비중이 큰 편이고 방사 조건별로 확보된 데이터 샘플이 매우 적음을 고려했을 때 추가 데이터 확보에 따라 이상치로 분류되지 않을 가능성이 있기 때문에 이상치 제거를 위한 거리 기준을 0.4로 정했다.



(그림 2) 이상치 처리 후 박스플롯 비교

그림 2(a)는 기존 816개 데이터, 그림 2(b)는 이상치 제거 후 796개 데이터에 대한 강도와 롤러1 속도 간 박스플롯 결과를 나타낸다. 세모는 이상치로 의심되는 데이터이며 그 중 일부가 이상치 제거 후 없어졌음을 확인할 수 있다. 이상치 제거 후에도 남아 있는 이상치 의심 데이터들은 해당 구간의 분포 데

이터가 적거나 0.3 수준의 차이로 인해 이상치로 오 판단된 데이터들이다. 이상치 처리 전후 결과를 비교해보면 데이터 분포에서 많이 동떨어진 데이터들은 제거된 것을 확인할 수 있다.

3.3 불균형정도와 상관계수를 고려한 오버 샘플링

이전 절에서 이상치 처리 문제는 해결하였으나 여전히 데이터 불균형 문제와 데이터양 부족 문제는 해결되지 않았다. 이런 문제들은 오버샘플링을 통해 해결될 수 있다. 그러나 데이터 불균형이 여러 변수에서 나타나는 경우 특정 변수의 소수구간 데이터가 다른 변수에서의 다수구간에 속하는 데이터라 오버샘플링 시 다른 변수에서 데이터 불균형이 심해질 수 있다는 점을 고려해야 한다. 따라서 각 주요 공정 변수들의 데이터 불균형 정도와 예측 물성 대상인 강도에 어느 정도 영향력이 있는지를 알 수 있는 상관성을 활용해 균형있는 오버샘플링 기법을 적용할 필요가 있다.

<표 3> 강도와 주요 공정 변수 간 상관계수

| 구분 | 스핀빔 온도 | 롤러1 속도 | 롤러2 속도 | 롤러2 온도 | 권취 속도 |
|----|--------|--------|--------|--------|-------|
| 강도 | -0.51 | -0.42 | 0.13 | 0.33 | 0.09 |

표 3은 주요 공정 변수와 강도 간 상관분석 결과 상관계수를 나타낸다. 상관계수의 크기와 데이터 불균형 정도를 고려해서 소수 데이터를 복사하는 방식으로 데이터를 증강하였다. 상관계수의 크기가 비교적 크고 불균형 정도가 심한 롤러1 속도와 롤러2 온도의 경우 소수 구간 데이터를 3배씩 증강하였으며 상관계수의 크기가 비교적 작고 불균형 정도가 덜한 롤러2 속도와 권취 속도의 경우 소수 구간 데이터를 2배만큼 증강하였다. 또 스핀빔 온도의 경우 상관계수의 크기가 가장 크지만 데이터 불균형이 심하지 않기 때문에 데이터를 증강하지 않았다. 오버샘플링 결과 1193개로 늘어났으며 주요 공정 변수들의 불균형 완화 정도는 표 4에서 확인할 수 있다.

<표 4> 오버샘플링 전후 데이터 불균형 비율 비교

| 구분 | 스핀빔 온도 | 롤러1 속도 | 롤러2 속도 | 롤러2 온도 | 권취 속도 |
|-------|--------|--------|--------|--------|-------|
| 다수(전) | 65% | 88% | 76% | 92% | 82% |
| 다수(후) | 77% | 68% | 50% | 68% | 68% |
| 소수(전) | 35% | 12% | 24% | 8% | 18% |
| 소수(후) | 23% | 32% | 50% | 32% | 32% |

4. 물성 예측 모델 성능 비교

제안한 기법들을 적용한 결과 강도 예측 모델 성능의 개선 정도를 확인하기 위해 기본 데이터, 이상치 처리 데이터, 이상치 처리 후 오버샘플링 데이터로 강도 예측 모델을 학습시키고 모델 성능을 비교하였다. 학습 및 검증용 데이터와 테스트 데이터의 비율은 9:1이며 모델 학습을 위해 정규화 처리를 했다. 모델은 데이터 샘플링 등의 요인에 영향을 덜 받아 성능 편차가 적게 나타나는 MLP(Multi-Layer Perceptron)를 사용하였다. 모델에 대한 상세 정보는 표 5에 나타난다.

<표 5> MLP 모델 정보

| 은닉층수 | 레이어 | 활성화함수 | 옵티마이저 | 교차검증 |
|------|---------|-------|-------|--------|
| 2 | 64 / 64 | ReLU | Adam | 5-Fold |

모델 성능 비교를 위한 성능 지표는 조정된 결정계수와 평균절대오차, 평균제곱오차를 사용했다. 조정된 결정계수는 피팅이 완료된 모델 기준이며 평균절대오차 및 평균제곱오차의 경우 테스트 데이터에 대한 예측 성능이다.

<표 6> 각 데이터에 대한 모델 성능 평가 결과

| 구분 | 조정된 결정계수 | 평균절대 오차 | 평균제곱 오차 |
|------------|----------|---------|---------|
| 기본 데이터 | 0.479 | 0.165 | 0.047 |
| 이상치 처리 데이터 | 0.496 | 0.148 | 0.045 |
| 오버샘플링 데이터 | 0.789 | 0.120 | 0.027 |

표 6은 각 데이터들에 대한 모델 성능 지표 비교 결과를 나타낸다. 모든 평가지표에서 이상치 처리 후 오버샘플링 데이터가 가장 좋은 결과를 보인다. 조정된 결정계수는 모델이 종속변수를 얼마나 잘 설명하는지를 나타내는 지표로 1에 가까울수록 모델이 데이터에 잘 적합했다고 할 수 있다. 오버샘플링을 적용하지 않은 데이터들에서 0.5 미만으로 나타나던 결정계수가 오버샘플링 데이터에서는 0.8 수준으로 높아졌다. 이것은 모델이 데이터 분포가 적은 구간에서도 데이터에 대한 적합도가 높아졌다는 것이다. 또 평균절대오차 및 평균제곱오차 결과를 통해 강도 예측 오차가 이상치 처리 후 0.1정도, 오버샘플링 후 0.4정도 줄어든 것을 확인할 수 있다. 즉 제안한 기법들을 적용하여 기존 0.15에서 0.2 정도의 예측 오차를 0.1에서 0.15 정도로 개선했다. 그리고 두 기법 중 오버샘플링 적용에 대한 모델 성능 향상이 이상치 제거에 따른 성능 향상보다 좋게 나타났다.

5. 결론

본 논문에서는 섬유 방사 공정 데이터를 활용한 물성 예측 모델의 성능 개선을 위해 데이터 불균형, 적은 데이터양 및 샘플 간 오차 등의 데이터 특성을 고려한 이상치 처리와 오버샘플링 기법을 제안한다. 제안한 데이터 처리 기법들을 적용한 결과 기존 모델에 비해 평균절대오차는 약 27% 줄고 평균제곱오차는 약 43% 줄었다. 또 0.5 미만이었던 결정계수가 1에 가까운 약 0.8로 크게 개선되었다. 이는 강도 예측 오차가 감소하고 예측 모델이 데이터에 더 적합해졌음을 의미한다. 이렇듯 섬유 방사 공정 데이터 특성을 고려한 데이터 처리 기법으로 물성 예측 모델의 성능을 개선시킨 것처럼, AI 모델 개발 시에 데이터 특성이 고려된 적절한 데이터 처리 기법을 적용한다면 모델 성능 개선에 도움이 될 것이다.

※ 본 연구는 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구이며 (1415176833) 또한 교육부 및 한국연구재단의 4단계 BK21 사업(경북대학교 컴퓨터학부 지능융합 소프트웨어 교육연구단)으로 지원된 연구임 (4199990214394).

참고문헌

- [1] Zhou, Z. et al, "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] 조기환, 정성훈, "P-71 : PET single filament 데이터로부터의 번들강도 결정을 위한 통계적 접근", 한국섬유공학회 학술발표논문집, Vol. 2003, pp. 291-292, 2006.
- [3] 허유, 김종성, "콤팩트 형성에 따른 링 정방장력 거동의 실험적 모델링", 한국섬유공학회지, Vol. 39, pp. 209-216, 2002.
- [4] 전봉수, 양철곤, "회귀분석과 신경망을 이용한 면방적사의 강도 예측", 한국섬유공학회지, Vol. 34, pp. 731-738, 1997.
- [5] 박정희, 김태공, 김지일, 최세목, 이경훈, "클러스터링 기반 양상불 모델 구성을 이용한 이상치 탐지", 정보처리학회논문지, 소프트웨어 및 데이터 공학 7, No.11 pp. 435-442, 2018.
- [6] 박지수, 안길승, 허선, "클래스 불균형 데이터의 효과적인 분류를 위한 k-NN과 생성적 적대 신경망 기반의 오버샘플링", 대한산업공학회지, Vol. 46, No. 4, pp. 365-371, 2020.
- [7] 김성훈, 김승천, "클래스 불균형 문제가 있는 특이분류 데이터의 자동분류 성능 개선을 위한 모델 재귀적 오버샘플링 방법", 전자공학학회논문지, Vol. 58, No. 4, pp. 43-49, 2021.