통신 메타데이터의 엔트로피 분석과 하이브리드 딥러닝 기반 청년 고립 탐지

손병훈, 장여진, 이상금*
*국립한밭대학교

20217135@edu.hanbat.ac.kr, jyeoj251@gmail.com, *sangkeum@hanbat.ac.kr

Youth Isolation Detection Based on Communication Metadata Entropy Analysis and Hybrid Deep Learning

Byounghoon Son, Yeojin Jang, Sangkeum Lee*
*Hanbat National University

요 약

청년층의 사회적 고립은 심각한 사회문제로 대두되고 있으나, 기존의 설문 기반 조사나 고비용 센서 데이터 수집 방식은 확장성과 효율성에 한계가 있다. 본 연구는 스마트폰 통신 메타데이터만을 활용하여 고립 위험군을 조기에 탐지하는 HA-XGB(Hybrid LSTM-Autoencoder embedded XGBoost) 모델을 제안한다. 제안 모델은 통신 상대의 다양성을 정량화하는 섀넌 엔트로피(Shannon Entropy)를 핵심 특징으로 도입하고, LSTM-Autoencoder 를 통해 정상 통신 패턴의 시계열 특징을 비지도 학습한 후, 추출된 잠재 벡터와 재구성 오차를 XGBoost 분류기에 결합하는 2 단계 하이브리드 구조를 채택하였다. 서울시 시민생활 데이터(2022-2024)를 대상으로 한 실험 결과, 제안 모델은 F1-Score 0.89, 재현율 0.94 를 달성하여 Logistic Regression(F1: 0.62) 및 단일 XGBoost(F1: 0.87) 대비 우수한 성능을 보였다. 본 연구는 프라이버시 침해를 최소화하면서도 복지 사각지대의 고립 청년을 선제적으로 발굴할 수 있는 실용적 방법론을 제시한다.

Ⅰ. 서 론

현대 사회에서 청년층의 사회적 고립은 우울증, 자살 충동 등 심각한 정신건강 문제로 이어질 수 있어 조기 발견과 개입이 중요한 사회적 과제로 부상하고 있다. 특히 개인주의 심화와 비대면 생활의 장기화로 은둔형 외톨이, 고립 청년 문제가 심화되고 있으나, 기존의 인적 발굴 체계는 비용과 효율성 측면에서 한계가 존재한다[1].

이에 스마트폰 센서를 활용한 디지털 표현형(Digital Phenotyping) 연구가 주목받고 있다. Lee 등[2]은 GPS, 가속도계 등 다중 센서 데이터를 통해 외로움과 사회적 고립의 행동 지표를 수동적으로 감지하는 방법을 제안하였다. 그러나 이러한 접근은 지속적인 센서 데이터 수집에 따른 배터리 소모, 사생활 침해, 높은 리소스 비용 등의 문제가 제기된다. 또한 대부분의 선행연구는 통화 빈도나 문자 횟수 등 단순 통계량에 의존하여 소통의 질적 특성을 반영하지 못하는 한계가 있다.

본 연구는 이러한 한계를 극복하기 위해 고비용 센서 데이터 대신 기지국 접속 로그 등 경량화된 통신 메타데이터를 활용한다. 핵심 아이디어는 단순 통화량을 넘어 통신 상대의 다양성을 섀넌 엔트로피(Shannon Entropy)로 정량화하고, 시계열 이상 탐지 기법[3]을 접목하는 것이다. 구체적으로, LSTM-Autoencoder를 통해 정상 통신 패턴의 잠재 표현을 비지도 학습하고, 추출된 심층 특징을 XGBoost 분류기와 결합하는 HA-XGB(Hybrid LSTM-Autoencoder embedded XGBoost) 모델을 제안한다.

본 연구의 주요 기여는 다음과 같다. 첫째, 센서 데이터 없이 통신 메타데이터의 엔트로피 분석만으로 고립 위험을 탐지하는 프라이버시 보존형 방법론을 제시한다. 둘째, 시계열 특징 추출과 앙상블 분류를 결합한 하이브리드 모델을 통해 데이터 불균형 환경에서도 높은 재현율을 달성한다. 셋째, 서울시 실제 생활 데이터를 활용한 실증 분석을 통해 제안 방법의 실용적 적용 가능성을 검증한다.

Ⅱ. 본론

2.1 데이터 구성 및 전처리

본 연구는 서울시 '시민생활 데이터'와 '청년 고립 실태조사' 융합 데이터(2022.01-2024.12)를 활용하였으며, 총 12,450명(일반군 11,205명, 고립 위험군 1,245명)을 분석 대상으로 하였다. 전처리 과정에서 결측치는 중앙값으로 대체하고, 이상치는 상하위 1%를 윈저라이징 처리하였다. 핵심 변수인 통신 엔트로피는 섀넌 엔트로피 공식을 적용하여 다음과 같이 정의한다.

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

여기서 $P(x_i)$는 전체 통신 횟수 중 특정 대상 i와의 통신 비율을 의미한다. 엔트로피 값이 낮을수록 소수 대상에게 통신이 편중된 고립 상태를 나타내며, 높을수록 다양한 대상과 소통함을 나타낸다.

2.2 HA-XGB 모델 아키텍처

본 연구가 제안하는 HA-XGB 모델은 2단계 하이브리드 구조를 가진다. 첫째, 비지도 특징 추출(Unsupervised Feature Extraction) 단계에서는 시계열 데이터의 시간적 의존성을 학습할 수 있는 LSTM-Autoencoder를 사용하여 일반 청년층의 정상

통신 시퀀스를 학습한다. 이때 모델이 예측한 값과 실제 값의 차이인 재구성 오차(Reconstruction Error)는 고립으로 인한 패턴 이질성을 나타내는 척도로 활용되며, 잠재 벡터(Latent Vector)와 함께 추출된다. 둘째, 지도 분류(Supervised Classification) 단계에서는 추출된 심층 특징(Deep Features)과 정적 통계 변수를 결합하여 XGBoost 모델의 입력값으로 사용한다. 클래스 불균형 해결을 위해 비용 민감 학습(Cost-sensitive Learning)을 적용하고 scale_pos_weight를 8.0으로 설정하였다. LSTM-Autoencoder는 2층 구조(hidden units: 64-32)로 구성하고, 시퀀스 길이는 30일로 설정하였다. 이를 통해 고립 위험군과 같이 상대적으로 표본 수가 적은 클래스에 대한 오분류 비용을 증가시켜 재현율 저하 문제를 완화한다.

2.3 실험 결과 및 분석

그림 1의 탐색적 데이터 분석(EDA) 결과, 고립 위험군의 평균 통신 엔트로피는 2.15로 일반군(3.91) 대비 유의미하게 낮았다($p<.001$). 고립 위험군이 전반적으로 낮은 엔트로피 구간에 밀집되어 있는 반면, 일반군은 상대적으로 높은 엔트로피 구간에 넓게 분포하는 경향을 보였다.

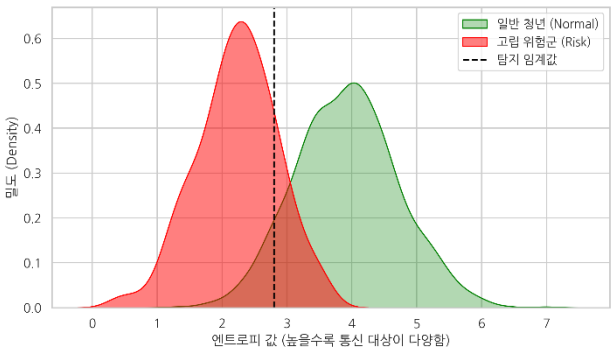

그림 1. 통신 엔트로피(Shannon Entropy) 분포 비교

이러한 결과는 고립 청년의 소통 구조가 소수의 특정 대상에게 강하게 편중되어 있거나 소통 자체가 극도로 제한적임을 시사한다. 즉, 통신 상대의 다양성이 급격히 감소함에 따라 외부와의 상호작용 패턴이 단조로워지는 특성이 엔트로피 저하로 나타난 것이다.

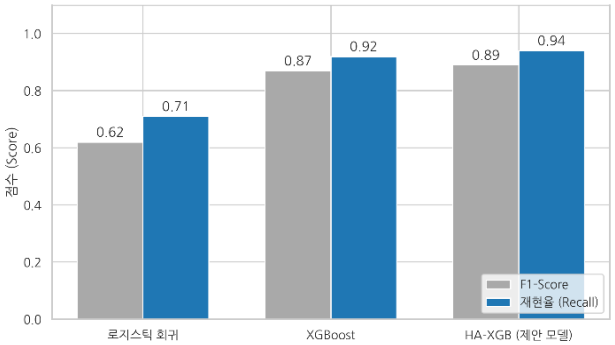

그림 2. 고립 탐지 모델별 성능 비교

실험은 8:2 비율로 학습/테스트 데이터를 분할하고 5-fold 교차검증을 수행하였다. 그림 2의 제안 모델(HA-XGB)과 베이스라인 모델들의 성능 비교 결과, Logistic Regression은 F1-Score 0.62, 단일 XGBoost는 0.87을 기록한 반면, 제안된 HA-XGB 모델은 가장 높은 F1-Score(0.89)와 재현율(0.94)을 기록하였다. 왜냐하면 기존 머신러닝 모델이 포착하기 어려운 시간적 패턴 정보를 LSTM 기반 특징 추출 단계가 효과적으로 반영했기 때문이다.

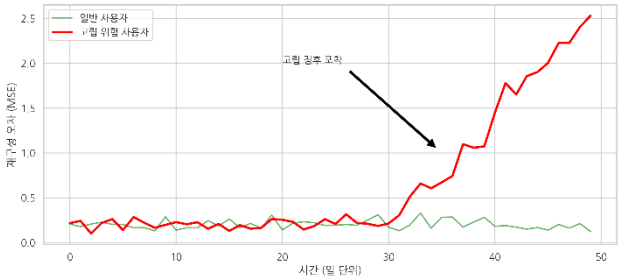

그림 3. 시간에 따른 재구성 오차(이상 징후) 변화

또한 그림 3에 제시된 재구성 오차의 시계열 변화 결과를 통해 일반 사용자에 비해 고립 위험 사용자의 오차가 특정 시점 이후 급격히 증가하는 경향을 확인할 수 있다. 이는 고립이 단일 시점의 이상 상태가 아니라 시간에 따라 점진적으로 누적·심화되는 과정임을 시사하며 LSTM-Autoencoder가 이러한 변화 양상을 효과적으로 반영할 수 있음을 보여준다.

일부 선행 연구에서 인간 행동의 예측 불가능성을 지적했으나, 본 연구는 딥러닝 기반의 특징 추출이 이러한 불확실성을 완화하고 예측 정확도를 높일 수 있음을 입증하였다.

Ⅲ. 결론

본 연구는 통신 메타데이터의 엔트로피 분석과 하이브리드 딥러닝 기법을 결합하여 청년 사회적 고립을 선제적으로 탐지하는 HA-XGB 모델을 제안하였다. 실험 결과, 제안 모델은 F1-Score 0.89, 재현율 0.94를 달성하여 기존 모델 대비 우수한 성능을 보였으며, 통신 다양성의 결여와 시계열 패턴의 이질성이 고립을 설명하는 핵심 요인임을 규명하였다. 이는 고비용 센서나 대면 조사 없이 복지 사각지대를 해소할 수 있는 실무적 도구를 제공한다는 점에서 의의가 있다. 다만 본 연구는 서울시 데이터에 한정되어 지역적 일반화에 한계가 있으며, 통신 메타데이터가 사회적 관계를 완전히 대변하지 못하므로 결과는 상관관계 수준에서 해석되어야 한다. 향후 연구에서는 다지역 데이터 검증과 마이데이터(MyData) 기반 실시간 위기 경보 시스템으로의 고도화가 요구된다.

참 고 문 헌

[1] Seoul Institute, "Analysis of social isolation among young adults in Seoul and policy suggestions," The Seoul Institute, Report No. 2023-12, 2023.

[2] H. Lee, A. Mensa, and S. Moon, "Identifying behavioral phenotypes of loneliness and social isolation with passive sensing," JMIR mHealth and uHealth, vol. 7, no. 7, e13237, 2019.

[3] Kwon, Woohyeon, et al. "FFT-Guided Multi-Window USAD with DTW-Isolation Forest for Reliable Anomaly Detection in Industrial Power Time-Series." *Energies* 18.24 (2025): 6584.