인간의 활동 센서 클러스터링: 멀티-핫 인코딩과 버트 임베딩 방식의 텍스트 벡터화에 따른 활동 정확도 비교

이현서, 김지인, 황의석*

광주과학기술원

{lhyeonseo, jeeinkim}@gm.gist.ac.kr, *euiseokh@gist.ac.kr

Clustering Human Activity Sensors: Accuracy Comparison of Text Vectorization Using Multi-Hot Encoding and BERT Embedding

Lee Hyeonseo, Kim Jeein, Hwang Euiseok*

Gwangju Institute of Science and Technology

요 약

본 연구는 레이블이 제공되지 않는 활동 센서 데이터를 대상으로, 텍스트 기반 벡터화 기법과 클러스터링을 결합하여 활동을 예측한 결과를 제시한다. 센서 데이터는 텍스트 형태로 기록되며, 이를 벡터로 변환하기 위해 멀티-핫 인코딩과 BERT 임베딩을 적용하였다. UCI-ADL 데이터셋을 활용한 실험 결과, 멀티-핫 인코딩보다 BERT 임베딩이 의미적 유사성을 효과적으로 반영하여 더 높은 정확도를 달성하였다. 특히 단어 수준보다 문장 수준 임베딩이 센서 속성 간 관계를 풍부하게 표현함으로써 클러스터링 성능 향상에 기여하였다. 본 연구는 활동 센서 데이터 분석에서 텍스트 벡터화 방식의 선택이 성능에 미치는 영향을 실증적으로 검증하였다는 점에서 의의를 갖는다.

Ⅰ. 서 론

센서 데이터를 기반으로 한 인간 활동 인식은 스마트홈, 웨어러블 디바이스, 헬스케어 모니터링 등 다양한 응용 분야에서 필수적인 핵심 기술이다. 특히 스마트홈 환경에서는 사용자의 생활 패턴을 이해하고, 이를 기반으로 맞춤형 서비스를 제공하기 위해 일상생활 속에서 발생하는 활동을 정확하게 식별하는 것이 중요하다.

기존 연구들은 센서 데이터로부터 얻은 이벤트 로그를 전처리하여 기계학습 기법을 적용하는 방식으로 활동 인식을 시도해 왔다.[1] 그러나 센서 데이터는 시계열적 특성과 비정형적인 기록 구조를 동시에 가지므로, 단순한 벡터화 방식은 활동 간의 복잡한 의미적 관계를 충분히 반영하지 못하는 한계가 있다.

본 연구에서는 센서 기반 데이터를 활용하여 인간 활동을 클러스터링하고, 두 가지 벡터화 접근법인 멀티-핫 인코딩과 BERT 임베딩[2]을 비교한다. 멀티-핫 인코딩은 단순한 이벤트 발생 여부를 기반으로 벡터를 생성하는 방식이며, BERT 임베딩은 텍스트로 기록된 센서 정보를 의미적 벡터로 변환하여 유사성을 반영하는 방식이다. 본 연구의 목적은 두 벡터화 기법의 클러스터링 성능을 비교하고, 성능 차이가 발생하는 원인을 분석하며, 더 나은 클러스터링 정확도를 달성할 방안을 탐색하는 것이다.

Ⅱ. 데이터 및 방법론

2.1 데이터셋

본 연구에서 사용된 데이터는 무선 센서 네트워크를 통해 수집된 UCI-ADL 데이터셋이다.[3] 두 가정(집 A와 집 B)에서 기록되었으며, 본 연구에서는 집 A만의 데이터를 분석 대상으로 삼았다. 집 A의 데이터는 2011년 11월 28일부터 12월 12일까지 15일 동안 수집되었으며, 세면대, 의자, 냉장고, 침대 등의 12개의 센서 항목에 대한 활성화 이벤트가 포함되어 있다. 모든 데이터는 텍스트 형식으로 기록되어 있으며, 이는 클러스터링을 위해 벡터 표현으로 변환하였다.

2.2 벡터화 기법

2.2.1 멀티-핫 인코딩

센서 이벤트를 30분 단위로 구간화하고, 각 구간에서 활성화된 센서를 1, 비활성화된 센서는 0으로 표시하여 벡터를 구성하였다. 이때, 실제 라벨이 아침, 점심, 저녁 식사 활동을 구분하고 있어, 각 시간 구간을 아침, 오후, 저녁, 밤으로 나누어주었고, 자주 함께 활성화된 센서들은 나타낼 수 있도록 벡터를 생성하였다.

2.2.2 BERT 임베딩

BERT는 트랜스포머 기반의 사전 학습 언어 모델로, 좌우 문맥을 동시에 고려하여 단어와 문장의 의미를 포착한다. 입력된 텍스트는 고차원 벡터로 변환되며, 벡터 간 거리는 의미적 유사성을 반영한다. 본 연구에서는 사전 학습된 문장 트랜스포머 모델(all-MiniLM-L6-v2)을 사용하여 센서 데이터를 임베딩하였다.

임베딩은 단어 수준 임베딩과 문장 수준 임베딩으로 수행되었다. 단어 수준 임베딩은 센서 이름만을 대상으로 임베딩을 수행하였다. 각 센서 명은 384차원 벡터 공간에 매핑되며, 의미적으로 유사한 센서들은 벡터 공간에서 서로 가깝게, 관련성이 낮은 센서들은 멀리 배치되도록 하였다. 이 방법을 통해 센서 간 의미적 유사성을 반영한 클러스터링이 가능하다. 문장 수준 임베딩은 센서 이름뿐만 아니라, 센서 위치, 시작·종료 시간과 같은 맥락 정보를 자연어 문장으로 생성한 뒤, 이를 임베딩한다.

2.3 K-means 클러스터링

K-means 클러스터링은 원하는 클러스터 개수(k)를 지정한 후 각 관측값을 정확히 하나의 클러스터에 할당한다. 즉, 각 관측값을 반복적으로 클러스터 중심점과 비교하며 재할당하는 방식으로 클러스터 내 분산을 최소화한다.

Ⅲ. 실험 결과

3.1 멀티 - 핫 인코딩

생성한 벡터에 대해 K-means 클러스터링을 수행하였다. 이때, k 값은 원래 라벨에 존재하는 활동 범주의 수에 의해 임의로 결정되었다. 각 시간대별로 분할하여 클러스터링 성능을 평가하였는데, 멀티-핫 인코딩의 아침과 저녁 시간대의 클러스터링 결과는 표 1에서 확인할 수 있다. 저녁 시간대의 전체 클러스터 평균 정확도는 58.09%로 비교적 낮은 성능을 보였다. 아침 시간대의 클러스터링 성능은 저녁보다 낮았다. 특히, 행동 Sleeping은 전혀 예측하지 못하는 형태가 보였다. 이는 아침 시간대 활동이 더 짧은 구간 내에서 자주 변화하기 때문이며, 이러한 조건에서는 멀티-핫 인코딩 기반 클러스터링이 효과적이지 않음을 확인할 수 있었다.

표 1. 아침과 저녁 시간대의 클러스터링 정확도

Table 1. Classification accuracy in Morning and Evening

Cluster	Evening		Morning	
	Activity	Accuracy (%)	Activity	Accuracy (%)
Cluster 0	Spare_Time	90.4	Sleeping	88.0
Cluster 1	Grooming	45.5	Spare_Time	14.3
Cluster 2	Leaving	62.5	Sleeping	0.0
Cluster 3	Snack	33.3	Grooming	18.18
Cluster 4	Toileting	58.8	Showering	16.66

3.2 BERT 임베딩

BERT 임베딩은 단어와 문장 수준 임베딩으로 수행하였다. 표 2에서 확인할 수 있듯이, 문장 수준 임베딩 기반 클러스터링 결과는 센서 이름만을 사용한 단어 수준 임베딩보다 더 높은 클러스터링 정확도를 보여주었다. 특히 Toileting과 Grooming은 단어 수준 임베딩 대비 문장 수준 임베딩을 적용했을 때의 정확도가 약 10% 정도 향상된 것으로 나타났다. 센서 정보를 문장에 통합하는 문장 수준 임베딩이 단어 수준 임베딩 방식보다 우수한 클러스터링 성능을 보여줌을 표 2에서 확인할 수 있다. 문장 수준 임베딩 후 클러스터링을 진행한 결과를 PCA 차원 축소를 통해 3차원 공간에서 보여주는 그림 1의 결과를 보면, 같은 색으로 표시된 데이터 군집이 전반적으로 잘 모여 있는 모습을 보였다. 그러나 Toileting, Showering처럼 같은 센서 데이터를 공유하는 활동의 경우에는 다른 활동들처럼 명확하게 클러스터링 되지 않았다.

표 2. 단어와 문장 수준 BERT 임베딩 후 클러스터링 결과

Table 2. Result of clustering: word- and sentence-level BERT embedding

Cluster	Word-level		Sentence-level	
	Activity	Accuracy (%)	Activity	Accuracy (%)
Cluster 0	Toileting	58.62	Toileting	63.15
Cluster 1	Grooming	81.81	Grooming	92.53
Cluster 2	Showering	100	Showering	100
Cluster 3	Sleeping	100	Sleeping	100
Cluster 4	Eating	100	Eating	100
Cluster 5	Spare_Time	100	Spare_Time	100
Cluster 6	Leaving	96.3	Leaving	96.5

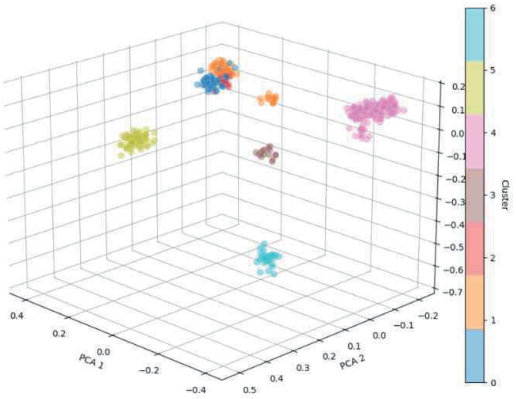

그림 1. 문장 수준 임베딩의 K-means 클러스터링 결과

Figure 1. K-means clustering using sentence-level embedding

Ⅳ. 토의 및 결론

본 연구에서는 UCI-ADL 데이터셋을 활용하여 두 가지 벡터화 기법, 즉 멀티-핫 인코딩과 BERT 임베딩을 비교하고, 클러스터링 성능 향상을 실증적으로 보여주었다. 멀티-핫 인코딩은 단순한 구현이 가능하지만, 활동을 인위적으로 30분 단위로 구간화해야 한다는 한계로 인해 낮은 클러스터링 성능을 보였다. 실제 활동은 균일한 30분 단위로 일어나지 않으므로 실제 라벨과 클러스터링 결과 간에 큰 불일치가 발생했고, 이는 결국 낮은 클러스터링 성능으로 이어졌다. 반면, BERT 임베딩은 원래의 타임스탬프와 센서 정보를 보존한 상태에서 의미적 유사성을 반영할 수 있어 더 높은 성능을 달성하였다. 특히, 문장 수준 임베딩은 센서 속성을 풍부하게 활용함으로써 정확도를 추가로 개선하였다.

향후 연구에서는 BERT 임베딩에 시간적 연속성을 반영할 수 있는 시계열 기반 딥러닝 모델과 결합하거나, 활동 간의 계층적 관계를 고려한 클러스터링 기법을 적용함으로써 성능을 더욱 향상할 수 있을 것이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2024-00349582).

이 논문은 과학기술정보통신부·광주광역시가 공동 지원한 '인공지능 중심 산업융합 집적단지 조성사업'으로 지원을 받아 수행된 연구 결과입니다.

참 고 문 헌

[1] P.-W. Chen et al., “Measuring activities of daily living in stroke patients with Motion Machine Learning Algorithms: A pilot study,” International Journal of Environmental Research and Public Health, vol. 18, no. 4, p. 1634, Feb. 2021. doi:10.3390/ijerph18041634

[2] Reimers, N., et al., “Classification and clustering of arguments with contextualized word embeddings,” *Proc. 57th Annu. Meet. Assoc. Comput. Linguistics*, 2019. (doi: 10.18653/v1/p19-1054).

[3] Ordez, F., “Activities of Daily Living (ADLs) Recognition Using Binary Sensors,” UCI Machine Learning Repository, 2013. (doi: 10.24432/C5J02M).