데이터셋 특성에 기반한 특징 선택을 통한 머신러닝 향상 기법

박나은 [1], 김연진 [2], 이일구 [3]
[1] 성신여자대학교 미래융합기술공학과 박사과정
[2] 성신여자대학교 융합보안공학과 석사과정
[3] 성신여자대학교 미래융합기술공학과, 융합보안공학과 교수

nepark.cse@gmail.com, 220246046@sungshin.ac.kr, iglee@sungshin.ac.kr

Enhancing Machine Learning via Dataset Characterization-Driven Feature Selection

Na-Eun Park[1], Yeon-Jin Kim[2], Il-Gu Lee[1,2]
[1] Dept. of Convergence Security Engineering, Sungshin Women's University
[2] Dept. of Future Convergence Technology Engineering, Sungshin Women's University

요 약

본 연구는 경량 디바이스의 데이터 처리를 위한 Dataset Characterization-Driven Feature Selection 방식을 제안한다. 실험 결과에 따르면, DC-DFS 가 선택한 특징 중요도 기반 특징 선택에서 종래의 정적 피처 선택 방식보다 약 10.32% 더 높은 정확도를 보였다.

1. 서론

사물인터넷(Internet of Things, IoT) 기술이 발전하면서, 경량화 된 데이터 처리 방식이 주목받고 있다. 고밀집 네트워크 환경에서 이상 트래픽을 탐지하기 위해서는 방대한 트래픽을 효과적으로 처리해야 하지만, 데이터 전처리와 모델 학습에 소요되는 시간 및 비용으로 인한 오버헤드 문제가 발생한다[1]. 특히, 경량 장치의 적은 데이터는 모델 학습에 어려움을 야기하므로, 모델 성능과 학습 비용 간의 트레이드 오프를 완화할 수 있는 효과적인 데이터 처리 방식이 필요하다.

이에 따라, 모델 성능을 유지하면서 학습 비용을 절감하기 위한 다양한 연구가 활발히 진행되고 있다. 고차원 데이터의 중요한 특징들은 유지하면서 불필요한 정보나 변수를 제거하여 저차원으로 압축하거나, 데이터에 포함된 노이즈를 제거하는 방식들이 주목받고 있다[2]. 특히, 각 피처가 모델의 학습에 얼마나 기여하는 지 분석하는 피처 중요도를 기준으로, 피처를 선별하여 학습하는 피처 선택 방식은 모델 과적합 문제를 해결하고 성능 개선과 비용 효율성을 개선하는 모델로 주목받고 있다[3].

다양한 머신 러닝 연구에서 효율적인 데이터 전처리를 위해 피처 선택 방식을 활용하고 있지만, 동일한 모델과 데이터셋에서 피처 중요도 측정 방법에 따라 피처의 중요도 순위가 달라지는 문제가 있다. 즉, 데이터 특성에 맞지 않은 피처 중요도 방식을 사용하면, 피처 선택 방식이 최적의 성능을 보장하지 못한다. 따라서 데이터의 특성을 분석하고, 데이터셋에 적절한 피처 중요도 방식을 선정, 적용하는 연구가 필요하다.

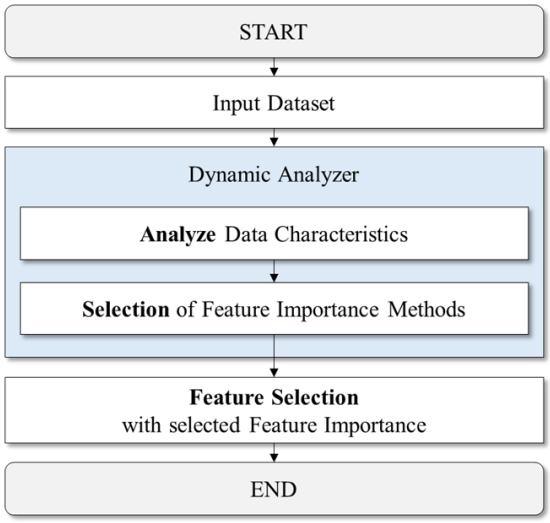
본 논문에서는 종래 피처 중요도 방식의 한계점을 극복하기 위한 DC-DFS (Dataset Characterization-Driven Feature Selection)을 제안한다. 데이터셋의 특성을 분석하고 종합적으로 평가하여 최적의 피처 중요도 측정 방식을 도출하여 데이터에 적합한 피처 선택 방식을 사용할 수 있다. 본 연구의 기여점은 다음과 같다.

- 제안하는 DC-DFS 을 이용하여 데이터 특성을 분석하고, 모델의 평가 결과에 대한 설명 가능성을 제공할 수 있는 프레임워크를 제안했다.
- 학습 데이터의 특성에 따라 동적으로 피처 중요도 방식을 선택하는 제안 모델과 고정된 피처 중요도 방식을 이용하는 정적인 종래 모델의 피처 선택 성능 평가 결과에 따르면 제안 모델이 약 10.32% 더 높은 정확도를 보였다.

본 논문은 다음과 같이 구성된다. 2 절에서는 제안하는 DC-DFS 모델을 설명하고, 3 절에서는 실험 환경과 평과 결과를 제시한다. 이후 4 절에서는 연구의 결론과 향후 연구 방향을 논의한다.

2. Dataset Characterization-Driven Feature Selection

데이터셋의 특성에 따라 피처 중요도 방식을 선택하는 DC-DFS 의 전체 흐름은 그림 1 과 같다.



(그림 1) Flowchart of proposed DC-DFS

학습 데이터가 입력되면, DC-DFS 모델의 동적 분석기는 데이터 크기, 고유 값, 상관관계, 클래스 분포, 엔트로피의 5 개 특성 값을 기반으로 데이터에 적합한 피처 중요도 방식을 선정한다.

본 연구에서는 선행 연구 분석을 기반으로 정의한 표 1 과 같은 특성 매트릭을 정의하여 활용하였으며, 피처 중요도 방식으로는 MDI importance, permutation importance 를 사용했다.

<표 1> Characteristics metrics [4,5]
(○: suitable, △: limited suitability, ×: not suitable)

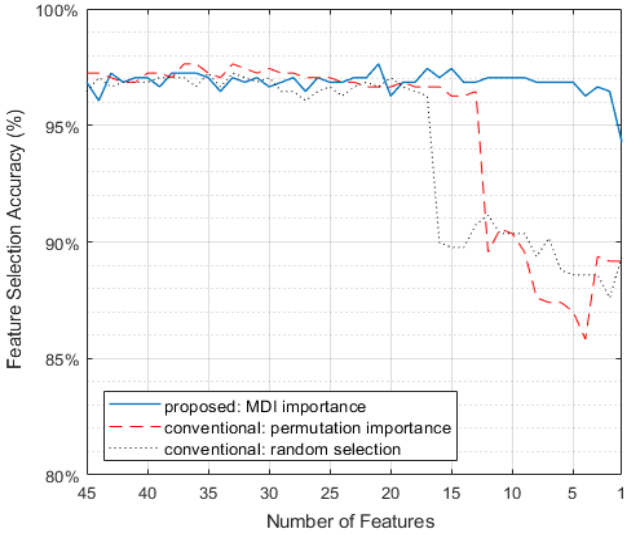
Properties	MDI	Permutation
Data size	○	△
Unique value	×	○
Correlation	○	×
Distribution	△	○
Entropy	○	△

3. 성능 평가 및 분석

본 논문에서 제안하는 DC-DFS 의 성능 검증을 위해, 피처 중요도 기반의 피처 선택 방식을 통해 모델 성능을 비교했다. 데이터셋의 특성에 따라 피처 중요도 방식을 동적으로 선택하는 DC-DFS 과 고정된 피처 중요도 방식을 선택하는 정적 피처 선택 방식, 그리고 무작위로 피처를 선택하는 무작위 선택 방식을 파이썬으로 구현하고 정확도를 비교했다.

본 논문에서는 UNSW-NB15 데이터셋을 사용하고, 경량 장치의 희소 데이터 환경을 시뮬레이션하기 위해 전체 데이터셋의 0.1%에 해당하는 1,150 개 샘플을 무작위로 추출하여 사용하였다.

데이터셋 분석 결과에 따르면 고유 값의 비율은 12.86%, 상관관계는 8%, 클래스 분포는 8.28%로 낮은 비율을 보였고, 엔트로피는 76%으로 높은 값을 보였다. 표 1 에 따르면, 높은 엔트로피는 permutation importance 의 판별 성능을 저하시킬 수 있으므로, 제안하는 DC-DFS 은 MDI importance 를 선택하여 피처 선택 방식을 수행한다. 각 피처 중요도 방식에 대한 실험 결과 비교는 그림 2 와 같다.



(그림 2) Comparison of accuracy by methods

DC-DFS 에 의해 선택된 MDI importance 를 활용한 피처 선택 결과 대부분의 경우에서 95% 이상의 정확도를 보였다. 그러나 permutation importance 를 이용한 피처 선택 방식은 피처가 12 개 남았을 때 정확도가 90% 미만으로 저하되었으며, 무작위 선택 방식은 피처가 16 개 남았을 때부터 90% 미만의 정확도를 보였다.

즉, DC-DFS 이 선택한 MDI importance 가 적절한 피처 중요도 방식임을 실험적으로 검증했으며, 실험 결과에 따르면 5 개의 피처가 남아있는 환경에서도 제안 모델이 평균 10.32% 더 높은 정확도를 보였다.

4. 결론

본 연구는 경량 디바이스에서 소규모 데이터에 대한 효율적인 데이터 처리와 이상 탐지를 위해 데이터셋 특성 기반의 피처 중요도 선택 방법인 DC-DFS 을 제안했다. 입력된 데이터셋의 특성에 따라 적절한 피처 중요도 방식을 선택하여 탐지 성능을 향상시켰다. 실험 결과에 따르면, 제안하는 모델이 선택한 피처

중요도 방식을 이용한 피처 선택 기반의 학습에서 평균 10.32% 더 높은 정확도를 보였다. 향후 연구에서는 이 연구는 데이터셋의 특성 매트릭에 대한 더 상세한 기준과 근거를 이용하고, 다양한 데이터 및 환경을 대상으로 한 성능 검증을 수행한다.

ACKNOWLEDGEMENT

본 논문은 2024 년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Ye-Seul Kil, Yeon-Ji Lee, So-Eun Jeon, Ye-Sol Oh, Il-Gu Lee, "Optimization of Privacy-Utility Trade-off for Efficient Feature Selection of Secure Internet of Things," IEEE Access, vol. 12, pp. 142582-142591, 2024.

[2] Uthayakumar Jayasankar, Vengattaraman Thirumal, Dhavachelvan Ponnurangam, "A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications," Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 2, pp. 119-140, 2021.

[3] Utkarsh Mahadeo Khaire, R. Dhanalakshmi, "Stability of feature selection algorithm: A review," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 4, pp. 1060-1073, 2022.

[4] Tousi Ashkan, Luján Mikel, "Comparative analysis of machine learning models for performance prediction of the spec benchmarks," IEEE Access, vol. 10, pp. 11994-12011, 2022.

[5] Zhe Li, Yahui Cui, Longlong Li, Runlin Chen, Liang Dong, Juan Du, "Hierarchical amplitude-aware permutation entropy-based fault feature extraction method for rolling bearings," Entropy, vol. 24, 2022.