

데이터 드리프트 검출 및 추적 시스템의 설계 및 개발

박종빈, 김경원, 정종진

한국전자기술연구원

jpark@keti.re.kr

Design and Development of Data Drift Detection and Tracking System

Jongbin Park, Kyung-Won Kim, Jong-Jin Jung

Korea Electronics Technology Institute

요약

본 논문은 기계학습이나 인공지능 모델을 개발하고 운용 시 발생하는 데이터 드리프트 현상을 검출하고 추적하는 시스템을 설계하고 개발한 내용을 소개한다. 통상적으로 데이터 드리프트 문제는 데이터 간 통계 분포의 불일치, 외부 환경 및 요구사항의 변화 등이 분석 모델의 성능 저하로 이어진다. 구체적인 데이터 드리프트 발생 원인은 다양하며, 서비스나 데이터 특성에 따라 다양한 감시 및 추적 기술이 요구된다. 이러한 데이터 드리프트 문제를 조기에 검출하고, 지속적으로 추적하며 상황에 맞는 해결 방안을 제시하는 것은 날로 중요해지고 있다. 따라서 본 논문에서는 데이터 드리프트 검출 및 추적을 위한 벡터공간에서의 데이터 이상치 검출, 시계열 데이터에서의 드리프트 검출, 확률 분포 함수의 차이를 사용한 드리프트 검출, 텍스트 데이터에서의 시간에 따른 키워드 변화 추적을 통한 드리프트 검출 기술을 설계하고 최소 기능 구현 제품으로 구현한다.

I. 서론

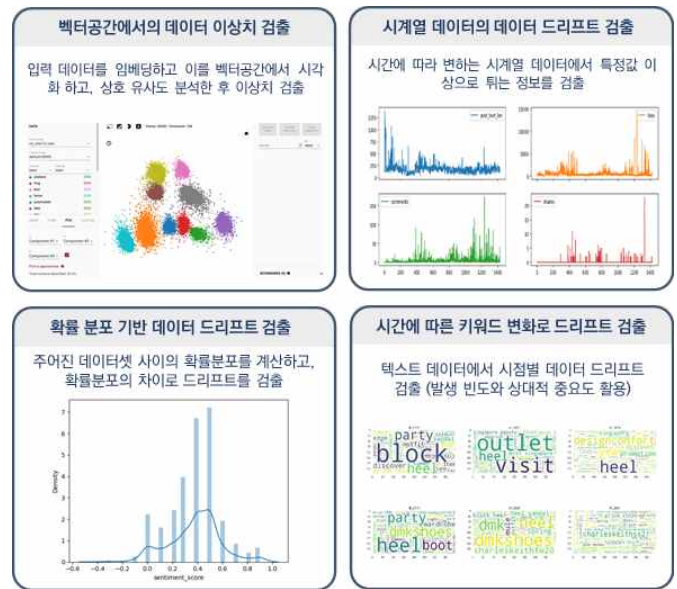
데이터 드리프트 문제는 분석을 위해 사용하는 소정의 모델이 실증 및 현장의 서비스 단계에서 성능이 감소하는 현상으로 해석할 수 있다[1]. 이를 해결하기 위해서 데이터 드리프트 현상이 언제, 어떻게, 얼마나 발생했는지를 파악하는 기술이 중요하며, 여러 가지 데이터 종류(예: 텍스트, 이미지, 비디오)에 대해서도 유기적으로 대응할 수 있어야 한다. 이런 문제를 해결하고자 본 논문에서는 다양한 데이터 종류에 대응하는 데이터 드리프트 현상 검출 및 추적 기술의 구현 내용을 소개한다.

II. 본론

그림 1은 개발한 데이터 드리프트 검출 및 추적 기술 관련 요소기술을 나타낸다. 이들은 마이크로 서비스 아키텍처[2] 방식으로 구현하며 통합 프레임워크를 구성한다. 그림 1의 왼쪽 상단은 Cifar10 이미지 데이터를 비전 트랜스포머(ViT, Vision Transformer)로 임베딩 후에 고차원 벡터 정보를 3차원 영역에 사영(projection)한 결과이다. 군집에서 벗어난 데이터를 이상치로 판단하거나 새롭게 입력된 이미지가 기존 군집과 얼마나 떨어져 있는지를 계산할 수도 있다. 그림 1의 오른쪽 상단은 시계열 데이터에서 데이터 변화를 검출하며 푸리에 변환, 웨이블릿 변환 등을 사용하여 데이터 주기성과 이에 따른 급격한 데이터 변화(Spike) 사건을 찾는 예시이다. 그림 1의 왼쪽 하단은 주어진 데이터에서 확률 분포 함수를 계산 후 Earth Mover distance나 Kullback-Leibler divergence를 계산하여 데이터셋 사이의 드리프트를 정량적으로 계산하는 예시이다. 그림 1의 오른쪽 하단은 키워드 분석을 통해 시간에 따른 드리프트 현상을 검출하고 추적하는 기술을 보여준다.

III. 결론

본 논문에서는 기계학습 및 인공지능 영역에서 활용할 수 있는 데이터 드리프트 검출 및 추적 기술을 최소 기능 제품으로 구현한 내용을 소개했다.



<그림 1> 데이터 드리프트 검출 및 추적 기술 예시

ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2024-00337489, 분석 모델의 성능저하 극복을 위한 데이터 드리프트 관리 기술 개발).

참고 문헌

- [1] A. Mallick, K. Hsieh, B. Arzani, G. Joshi, "Matchmaker: Data drift mitigation in machine learning for large-scale systems," Proceedings of Machine Learning and Systems, vol. 4, pp. 77-94, 2022.
- [2] N. Alshuqayran, N. Ali, and R. Evans, "A systematic mapping study in microservice architecture," IEEE 9th International Conference on Service-Oriented Computing and Applications (SOCA), pp. 44-51, 2016.