Isolation Forest 알고리즘을 활용한 증권 데이터 모니터링 시스템 개발

안우용[1], 김홍집[1], 김정연[1], 서승현[2]
[1]한양대학교 ERICA 전자공학부 학부생
[2]한양대학교 ERICA 전자공학부 교수

annavision21@hanyang.ac.kr, redhouse33@hanyang.ac.kr,
kjykjy6990@hanyang.ac.kr , seosh77@hanyang.ac.kr

Development of a Stock Data Monitoring System Using the Isolation Forest Algorithm

Woo-Yong An[1], Hong-Jip Kim[1], Jung-Yeon Kim[1], Seung-Hyun Seo[2]
[1]Dept. of Electrical Engineering, Hanyang University ERICA
[2]Dept. of Electrical Engineering, Hanyang University ERICA

요 약

변동성이 심한 증권 데이터의 특성 상 데이터의 다양한 요소에서 장애 상황이 발생한다. 따라서 실시간 대용량 데이터 처리 과정에서 발생할 수 있는 다양한 서비스 장애 요인들을 식별하고, 이를 신속하게 대응하기 위한 효율적인 실시간 모니터링 시스템 구축이 필요하다. 본 연구는 국내 증권사로 송신되는 해외 선물옵션 및 주식 데이터를 이상치 탐지 알고리즘인 Isolation Forest 를 통해 데이터의 이상치를 판단하고 알림 신호를 발생시키는 시스템을 제안한다.

1. 서론

ChatGPT 의 등장과 함께 인공지능 관련 주식의 급등세가 지속되면서, 연평균 투자 금액과 건수가 크게 증가하는 추세를 보이고 있다[1]. 이러한 흐름은 증권 데이터의 실시간 송수신 트래픽 증가로 이어지며, 다양한 원인에 의해 예기치 않은 서비스 장애가 발생하게 된다. 특히, 최근 주식 투자에 대한 관심이 높아지는 가운데, 증권 데이터 서비스의 장애는 투자자와 증권사 양측에 심각한 손실[2]을 초래하며, 이러한 손실은 장애 대응이 지연될수록 증가하는 경향을 보인다. 따라서, 실시간 데이터 모니터링 시스템을 구축하여 신속한 대처를 수행할 수 있는 환경을 구성할 수 있는 방안 마련이 필요하다.

현재 대형 증권사에서는 벤더의 이중화를 통해 실시간 데이터를 비교하여 이상치를 탐지한다. 그러나 이러한 방식은 중소형 증권사 측면에서 벤더를 이중으로 계약해야 한다는 금전적 부담이 존재한다.

이에 본 논문은 단일 벤더의 증권 데이터 내에서 정상적인 범위와 다른 패턴을 감지할 수 있는 Isolation Forest 을 활용한 시스템을 개발하였다. Isolation Forest 는 의사결정나무 기반의 비지도 학습 알고리즘으로, 다 변량 데이터에서 효율적으로 이상치를 탐지하여 신속한 장애 대응을 가능하게 하는 이상 탐지 알고리즘이다. 이에 증권 데이터의 다 변량 특성과 복잡성을 고려하여, Isolation Forest 알고리즘을 활용한 모니터링 시스템을 제안하였다.

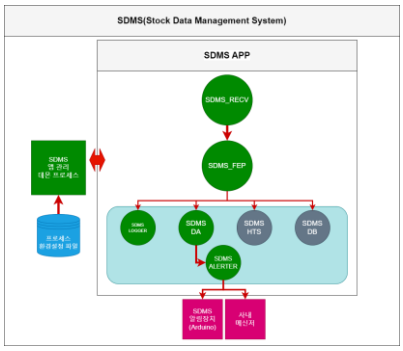
2. SDMS(Stock Data Monitoring System)

본 연구의 작업 환경은 아래와 같다.

<표 1> SDMS 작업환경

OS	Redhat Linux 7.9
Server	Dell PowerEdge R420
Language	Python 3.10.13

(그림 1)은 본 연구에서 개발된 증권 데이터 모니터링 시스템(Stock Data Monitoring System, SDMS)의 아키텍처를 설명하는 전체 흐름도를 보여준다. 시스템은 데이터 수신(SDMS_RECV), 데이터 가공(SDMS_FEP), 로깅(SDMS_LOGGER), 데이터베이스 저장(SDMS_DB), 데이터 분석 및 장애 탐지(SDMS_DA), Alert 신호 전달 (SDMS_ALERTER)을 담당하는 여러 프로세스(이하 'APP')로 구성되어 있다. 또한, 모든 APP 의 시작, 종료 및 인터럽트를 관리하는 데몬 프로세스와, 장애 발생 시 알림을 제공하는 Arduino 기반의 알림 장치로 이루어져 있다. 이 구조는 실시간 증권 데이터 처리의 효율성과 신뢰성을 극대화하기 위해 설계되었다.



(그림 1) SDMS 전체 구조도

2.1. SDMS_RECV

이는 증권 데이터의 실시간 수신을 담당하는 데이터 수신 전용 APP 이다. Multicast 방식을 통해 데이터를 수신하며, 이후 SDMS_FEP 로 데이터를 전송하는 기능을 갖추고 있다.

2.2. SDMS_FEP

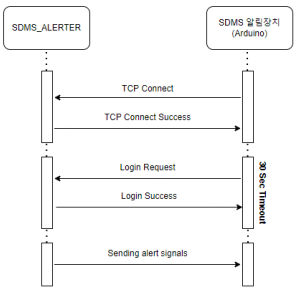
APP 간의 데이터 송수신 역할을 수행하는 APP 로, UNIX Domain Socket 을 활용하여 SDMS_RECV 에서 받은 데이터를 다른 모든 APP 에게 분배한다.

2.3. SDMS_DA

데이터의 이상 패턴을 식별하여 알림 신호를 발생시키는 기능을 하는 APP 이다. 이상치 탐지는 SDMS_DA 앱 내에서 여러 개의 Thread 로 운영되어 처리된다. 감지된 이상치는 정수화 된 Feature 를 다시 문자열로 변환한 후 JSON 형태로 SDMS_ALERTER 로 이상치 데이터를 전달한다.

2.4. SDMS_ALERTER

이 APP 는 SDMS_DA 로부터 전달받은 JSON 형태의 이상치 데이터를 외부 알림 장치로 전송하는 역할을 담당한다. 비동기 프로그래밍을 지원하는 asyncio 라는 표준 라이브러리를 활용하여 TCP 서버를 개발하였다. (그림 3)은 Arduino 기반의 외부 알림 장치와 SDMS_ALERTER 애플리케이션 간의 상호작용 방식을 보여준다.



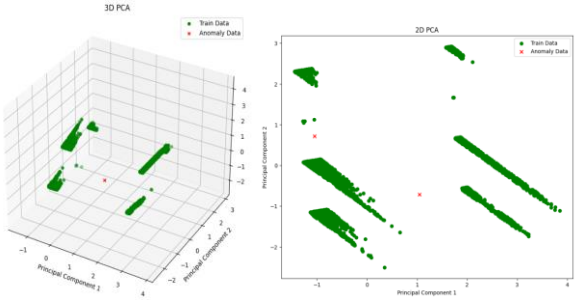
(그림 3) SDMS_ALERTER 상호작용 방식

3. 데이터 수신 이상치 탐지

본 연구에서는 우선 데이터 이상치 중 정규 장 내 수신 장애에 대해 탐지하는 프로그램을 개발하였다.

이상치를 탐지하기 위해 Isolation Forest 모델을 사용하며, 데이터 처리를 위한 첫 단계로 Raw 데이터의 전처리 과정을 진행한다. 이 과정에서 거래소 이름, 종목 코드, 데이터 유형, 2 분 간격 할당 시간 클래스, 요일, 그리고 2 분 동안의 데이터 수신 횟수를 추출한다. 모델 입력을 위해 문자열 데이터는 ord() 함수를 통해 유니코드 정수로 변환된다.

후속 단계에서는 각 종목 코드별로 3 주 간의 Train 데이터를 수집하고, 이를 바탕으로 모델 학습을 진행한다. Real 데이터에서 이상치 탐지를 통해 정상으로 분류된 데이터는 모델에 학습시킨다. Isolation Forest 모델은 의사 결정 나무 100 개로 구성되며, 이상치 비율은 44%, max_features 는 전체 특성의 수로 설정하여 최적화한다[4]. 모델의 학습과 이상치 도출은 2 분 간격으로 수행되며, 이 과정에서 joblib 과 pickle 라이브러리를 활용하여 모델과 데이터셋의 저장 및 로딩이 이루어진다. (그림 2)는 Receive Checker 에 의해 탐지된 이상치와 Train 데이터셋을 PCA 를 활용해 3 차원 및 2 차원으로 축소한 분포를 나타낸다.



(그림 2) Receive Checker Outlier Detection

4. 결론

본 연구를 통해, Isolation Forest 알고리즘을 활용한 자동화된 데이터 모니터링 플랫폼을 개발하였다. 앞으로의 연구에서는 추가적인 실험과 검증 작업을 진행하여, 이상치 탐지의 정확도를 더욱 향상시키고 모니터링 시스템의 전반적인 신뢰도를 향상하고자 한다.

참고문헌

[1] 리서치센터 글로벌주식팀(2024), 생성 AI, 영상 혁명의 시작, 삼성증권

[2] 장순환, “코스콤 시스템 오류..전 증권사 한때 주식 서비스 장애, 2023.07.31.

[3] 고영호,김승환,이학태,박지훈,and 박헌진. "군집분석과 Isolation Forest 를 이용한 접근 이상 항적 탐지." 한국지능시스템학회논문지 34.1 (2024): 25-35.

[4] 시종욱(Jongwook Si),and 김성영(Sungyoung Kim). "Isolation Forest 기반의 타일 패턴 불량 감지를 위한 특징 맵에 따른 성능 비교." Proceedings of KIIT Conference 2023.6 (2023): 127-128.