오픈 소스 H2o를 활용한 모델 자동 학습 및 자동 배포 프레임워크 개발

김종현, 박민호
숭실대학교 IT융합학과
kjhksh@gmail.com, mhp@ssu.ac.kr

Development of model auto-learning and auto-deployment framework using open source H2o

Kim Jong Hyun, Park Min ho
Dept. of IT Convergence, Soongsil Univ.

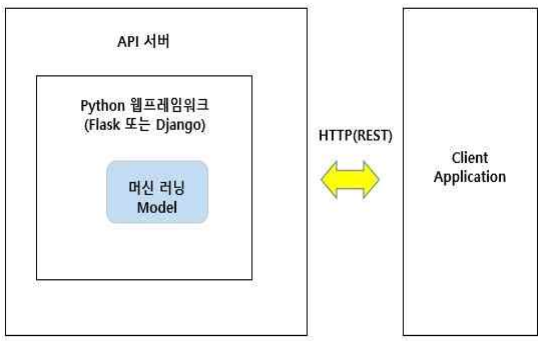
요 약

머신 러닝 모델 서빙이란 머신러닝 라이브러리를 이용하여 모델을 개발하고 다른 애플리케이션에서 모델을 사용할 수 있도록 모델을 배포하거나 모델 API를 제공하는 것을 말한다. 배포된 모델을 서비스 하기 위하서는 Flask나 Django와 같은 Python 웹프레임워크 사용하게 된다. Python은 인터프리터 언어로 개발 편의성이 높지만, 바이너리 파일만 실행시키면 되는 컴파일러와 달리, 변환과 실행을 동시에 진행해야 하므로 프로그램 자체 속도는 컴파일러보다 느리다. 본 논문에서는 오픈 소스 머신러닝 플랫폼 H2O ai를 이용하여 Java 애플리케이션에서 사용할있는 MOJO 형식의 모델을 생성하고 모델 서빙하는 Api Server를 구성하여 Python 웹프레임워크에서 서빙에 대한 문제점을 보완하고자 한다.

Ⅰ. 서 론

머신 러닝 모델 서빙이란 머신러닝 라이브러리를 이용하여 모델을 개발하고 다른 애플리케이션에서 모델을 사용할 수 있도록 모델을 배포하거나 모델 API를 제공하는 것을 말한다.
머신 러닝에서 사용하는 언어는 보통 Python 으로 개발한다.
머신러닝의 모델 서비스 절차는 데이터 수집, 데이터 전처리, 데이터 학습 모델 평가, 모델 배포의 단계로 이루어 지는데 마지막 단계인 배포된 모델을 서비스 하기 위하서는 Flask나 Django와 같은 Python웹 프레임워크 사용하게 된다.
Flask나 Django와 같은 웹 프레임워크는 Python 언어로 되어 있기 때문에 GIL(Global Interpretor Lock)로 인해 Multi-threading 성능에 제약이 있으며 GC로 인해 서버가 느려질 수도 있고 연산이 많은 resource가 필요한 모델을 서비스하는 것은 더 어려운 문제 발생 할수도 있다. 이유는 Python은 인터프리터 언어로 개발 편의성이 높지만, 바이너리 파일만 실행시키면 되는 컴파일러와 달리, 변환과 실행을 동시에 진행해야 하므로 프로그램 자체 속도는 느리다는데 있다.
본 논문에서는 머신 러닝 모델을 JAVA환경으로 모델 서빙을 하는 방법으로 Python 웹프레임워크에서 서빙에 대한 문제점을 보완하는 연구를 진행하고자 한다.

Ⅱ. Python웹 프레임워크를 이용한 모델 서빙



〈그림 1. Python 웹프레임워크 API Server 구성〉

〈그림 1〉은 Python 웹프레임워크을 이용한 구성에서 모델 서빙을 하는 API Server 구성을 나타낸다.
Client Application에서 모델 API Server로 요청 시 Python 기반의 웹 프레임워크으로 구성되기 때문에 인터프리터 언어의 문제인 변환과 실행을 동시에 진행해야 하기 때문에 프로그램 자체 속도는 느리다는 데 있다.

Ⅲ. 제안 시스템의 기반 기술

3.1. 스프링 프레임 워크

Spring Framework는 자바 플랫폼을 위한 오픈 소스 애플리케이션 프레임워크이다.
경량 컨테이너로서 자바 객체를 직접 관리한다. 각각의 객체 생성, 소멸과 같은 라이프 사이클을 관리하며 스프링으로부터 필요한 객체를 얻어올 수 있다.[4]

3.2. Spring Quartz

Job 스케줄링을 구현할 수 있는 오픈 소스 라이브러리이며 Java 어플리케이션에서 통합이 가능하다.
간단하거나 복잡한 스케줄까지 구현 가능하며 스케줄의 종료 시점부터 다음 실행 시점까지 시간 간격을 두는 인터벌(interval) 형식의 스케줄링이 가능하며 크론 표현식(cron expression) 방식을 이용한 복잡한 스케줄링도 지원한다.[4]

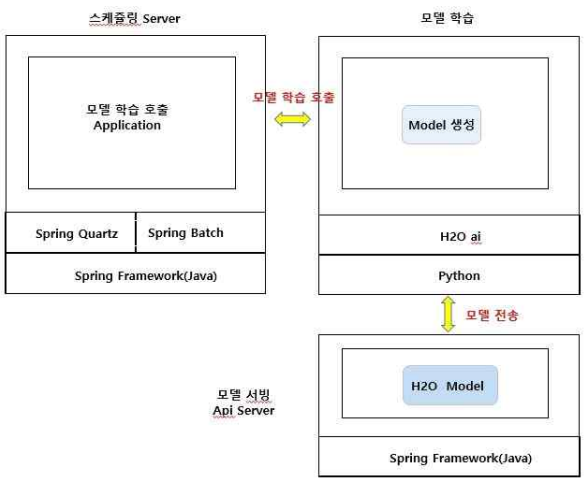
3.3. Spring Batch

엔터프라이즈 시스템의 일상적인 작업에 필수적인 강력한 배치 응용 프로그램을 개발할 수 있도록 설계된 가볍고 포괄적인 배치 프레임워크이다.[4]

3.4. H20 ai

H2O는 선형 확장성을 갖춘 완전 오픈 소스 머신러닝 플랫폼이다.
H2O는 gradient boosted machines, generalized linear models, deep learning등을 포함하여 가장 널리 사용되는 통계 및 머신 러닝 알고리즘 을 지원한다.
H2O에서 생성한 모델을 Java 애플리케이션에서 사용할 있는 MOJO 형식을 지원한다.[3]

III. 제안 시스템 구조



〈그림 2. H2O를 이용한 API Server 구성〉

본 논문에서는 〈그림 2〉와 같이 오픈 소스 머신러닝 플랫폼 H2O을 이용한 모델 서빙 API Server 구성을 제안한다
스케쥴링 Server에서 오픈 소스 머신러닝 플랫폼 H2O ai에서 모델 학습을 호출하고 학습된 모델을 모델 서빙을 담당하는 Api Server에 모델을 전송한다.
백엔드에서 모델을 서빙하는 Api Server는 JAVA 기반의 SpringFrameWork로 구성하여 머신러닝 모델을 Python과 같은 인터프리터가 서비스하는 것이 아닌 JAVA 컴파일러로 실행하여 인터프리터의 문제점을 보완할 수 있으거라 기대된다.

IV. 실험

〈그림 3. 모델 서빙 Api서버 접속하여 분류 값 확인〉

제안된 논문의 기능 확인을 위해 H20 ai Classification(분류)모델을 생성하여 Java환경에서 사용할 수있는 Mojo 형식으로 빌드했다.
모델을 서빙하는 Api Server는 SpringFrameWork로 구성하였고 Api Server처리에 대한 API 응답을 확인하였다.

IV. 결론

본 논문에서는 〈그림 2〉와 같이 오픈 소스 머신러닝 플랫폼 H2O ai와 자바 플랫폼인 Spring Framework을 이용한 시스템을 제안하였다.
Python 기반의 서비스가 아닌 Java 기반의 서비스로 Multi-threading 성능 향상 및 엔터프라이즈 서비스 환경에 적합한 시스템 구성할 수있다.
위 제시된 내용을 시스템 구성을 통해 검증을 추가할 계획이다.

ACKNOWLEDGMENT

Put sponsor acknowledgments.

참 고 문 헌

[1] Flask 의 모델 서빙을 이용한 웹 어플리케이션 구현 : 한국정보처리학

회 학술대회논문집 28권1호 454-456(3pages)

[2] 실시간 분석을 위한 도커 컨테이너 기반의 딥러닝 모델 관리 시스템 설계 및 성능 비교 : 한국통신학회논문지 제46권 제2호 (390 - 400 (11page))

[3] "H2O" https://h2o.ai/platform/ai-cloud/make/h2o/

[4] "spring" https://spring.io/projects/spring-framework