# 고성능 자율 기계학습을 위한 인텔리전트 데이터베이스 플랫폼 설계

## Design of an Intelligent Database Platform for High-Performance Autonomic Machine Learning

임 종 태*, 김 민 수*, 최 도 진*, 복 경 수*, 유 재 수*+
충북대학교 정보통신공학부*

Jongtae Lim*, Minsoo Kim*, Dojin Choi*,
Kyoungsoo Bok*, Jaesoo Yoo*+
School of Information and Communication Engineering,
Chungbuk National University, Korea*

요약

최근 기계학습에 대한 연구들이 사회적으로 이슈가 되고 있다. 하지만 기계학습은 기계학습 모델을 만들고 세밀히 조정해야하는 복잡한 작업을 수행할 수 있는 전문 지식을 가진 사용자가 필요하다. 따라서 기계학습 과정에서 사용자가 수행하여야 하는 다양한 작업을 자동으로 수행할 수 있는 자율 기계학습이 연구되고 있다. 본 논문에서는 고성능 자율 기계학습을 위한 인텔리전트 데이터베이스 플랫폼을 제안한다.

## Ⅰ. 서론

최근 기계학습(Machine Learning)에 대한 연구들이 사회적으로 이슈가 되고 있다[1-3]. 기계학습은 인공 지능의 한 분야로, 컴퓨터가 학습할 수 있도록 알고리즘과 기술을 개발하는 연구 분야를 의미한다. 기계학습은 수신한 이메일이 스팸 메일인지 아닌지를 구분하는 간단한 응용으로부터 인공지능 비서와 같은 복잡한 응용까지 다양한 분야에 활용되고 있다. 이처럼 기계학습은 다양한 응용에 활용될 수 있는 강력한 힘을 가지고 있는 기술이지만 이를 제대로 활용하려면 기술과 도구, 컴퓨팅 파워, 데이터 외에도 기계학습 모델을 만들고 세밀히 조정해야하는 복잡한 작업을 수행할 수 있는 전문 지식을 가진 사용자가 필요하다. 기계학습 기반의 응용 개발에 대한 필요성이 증가하며 텐서플로우(Tensorflow)[1], 카페(Caffe)[2], 케라스(Keras)[3]와 같은 다양한 기계학습 플랫폼들이 공개되고 있지만 일반 사용자들에게는 아직 진입 장벽이 높은 실정이다.

+ 교신저자 : yjs@chungbuk.ac.kr
본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업(IITP-2018-2013-1-00881), 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원[No.B0101-15-0266, 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발], 그리고 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업 (No. NRF-2017M3C4A7069432)의 지원을 받아 수행된 연구임

이러한 문제를 해결하기 위하여 연구가 되기 시작한 것이 자율 기계학습(Autonomic Machine Learning)이다. 자율 기계학습은 기계학습 과정에서 사용자가 수행하여야 하는 학습 데이터 시각화, 데이터 전처리, 하이퍼-파라미터 튜닝, 모델 선택 및 평가, 결과 보고서, 모델 사용(추론) 등의 작업을 자동으로 수행할 수 있는 연구이다. 최근 구글(Google), 파나소니(Panasonic) 등의 기업들을 중심으로 자율 기계학습 플랫폼들이 발표됐다.

본 논문에서는 고성능 자율 기계학습을 위한 인텔리전트 데이터베이스(이하 인텔리전트 DB) 플랫폼을 제안한다. 제안하는 기법에서는 인텔리전트 DB 플랫폼의 전체 시스템 구조를 설계하며, 제안하는 플랫폼을 이용하여 자율 기계학습을 수행하기 위한 기반 기술들을 도출한다.

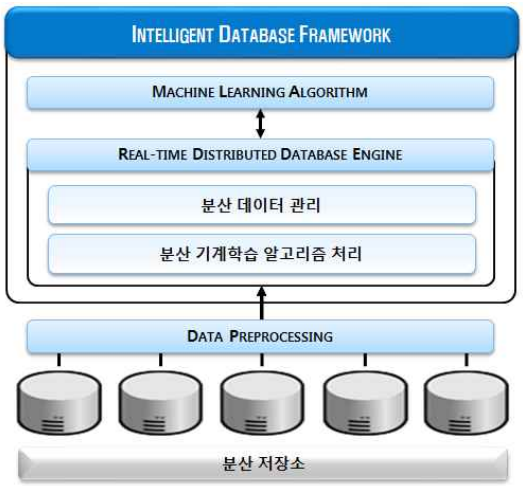
## Ⅱ. 제안하는 인텔리전트 DB 플랫폼

### 1. 전체 구조

인텔리전트 DB 플랫폼은 자율 기계학습을 위하여 데이터를 수집하고 전처리를 수행하며, 이를 학습하여 기계학습 알고리즘을 수행할 수 있도록 지원한다.

그림 1은 제안하는 인텔리전트 DB 플랫폼의 전체 구조를 보여준다. 제안하는 인텔리전트 DB 플랫폼은 실시간 분산 데이터베이스 엔진과 기계학습 알고리즘 실행 엔진으로 구성된다. 실시간 분산 데이터베이스 엔진은 기계학습을 알고리즘을 위한 데이터들을 수집/저장하고 이를 기계학습 알고리즘 엔진에 제공하는 역할을 수행한다.

분산 데이터베이스 엔진은 데이터 특성을 고려한 분산 저장 관리를 수행하는 다양한 분산 저장소와 연동된다. 기계학습 알고리즘 실행 엔진은 실시간 분산 데이터베이스 엔진으로부터 기계학습 데이터를 제공받아 기계학습을 수행한다. 기계학습 알고리즘 실행 엔진은 인텔리전트 DB 내부의 분산 저장소로 활용되는 노드들의 자원을 사용하여 알고리즘을 분산 처리한다. 기계학습 알고리즘 실행 엔진은 환경에서 따라서는 인텔리전트 DB 플랫폼의 분산 저장소가 아닌 원격의 위치에 알고리즘 수행만을 위한 별도의 시스템으로 존재하여 동작 할 수 있다.




▶▶ 그림 1. 인텔리전트 DB 전체 구조

## 2. 분산 데이터 관리 모듈

분산 데이터 관리 모듈은 실시간 분산 데이터베이스 엔진에서 데이터를 수집하고 저장한다. 분산 데이터 관리 모듈의 중요한 기능으로 동적 분산 데이터 관리, 동적 데이터 갱신, 패턴 분석 및 그룹화 등이 존재한다. 기계학습 데이터의 중요한 특징 중에 하나는 높은 정확도의 결과를 도출하기 위하여 잠재적으로 학습을 수행하는 데이터의 수가 방대하다는 것이다. 따라서 분산 저장소에 데이터들을 분산 저장하여야 하며, 인텔리전트 DB에 다양한 소스로부터 지속적으로 입력되는 데이터들에 대하여 동적으로 저장 관리를 수행할 수 있는 동적 분산 데이터 관리 기능이 필요하다. 또한 기계학습 데이터 중 온라인 문서 등을 다루는 응용에 활용되는 데이터들의 경우 기존의 데이터들을 대체하거나 수정 보완되는 이력 관리들이 필요한 데이터들이 존재한다. 하지만 이러한 이력 데이터들은 경우에 따라 원본 데이터의 몇 십 배에 해당하는 방대한 규모의 데이터로 나타날 수 있기 때문에 중복 데이터 등으로 인해 낭비되는 공간을 최소화하는 동적 데이터 갱신 기능이 필요하다. 마지막으로 기계학습의 중요한 주제인 추천과 관련된 연구에 사용하는 데이터의 경우 데이터들 간의 관계가 그래프 형태로 모델링 된다. 그래프 데이터의 경우 관련성 등을 고려하여 분산 저장을 수행하여야하는데 이를 위한 데이터들의 패턴 분석 및 그룹화를 수행하는 기능이 필요하다.

분산 데이터 관리 모듈에는 데이터를 저장 관리하는 위와 같은 주요 기능 외에도 분산 저장소와의 연동을 위한 기능들이 존재한다. 인텔리전트 DB에는 다양한 출처로부터 수집되는 다양한 형태의 기계학습 데이터들이 저장된다. 분산 데이터 관리 모듈은 분산 저장소에 저장되는 데이터들에 대하여 전처리를 수행하여 해당 데이터들이 기계학습 알고리즘에 입력 데이터로 활용 될 수 있도록 가공한다. 또한 분산 데이터 관리 모듈은 분산 저장소를 단일의 큰 저장소로 인식되도록 분산 저장소의 가상화를 수행하여 기계학습 알고리즘에 데이터가 입력 될 때 데이터의 저장소나 노드 위치에 관계없이 활용될 수 있도록 지원한다.

## 3. 분산 기계학습 알고리즘 처리 모듈

분산 기계학습 알고리즘 처리 모듈은 실시간 분산 데이터베이스 엔진을 통해 저장된 데이터들을 대상으로 기계학습 알고리즘을 처리한다. 분산 기계학습 알고리즘 처리 모듈의 중요한 기능으로 실시간 잡 스케쥴링, 인메모리 캐시 공유, 실체 뷰 생성 및 관리 등이 존재한다. 기계학습을 위한 최신 연구에서 높은 차원 및 복잡도를 가지는 기계학습 모델 및 알고리즘을 빠르게 처리하는 연구는 매우 중요한 연구이다. 특히 딥 러닝 등을 위해 사용하는 기계학습 알고리즘은 반복이 많은 계산 집중형 프로세스이다. 따라서 이러한 기계학습 알고리즘에 대하여 분산 노드에 작업을 적절히 분배하고 병합 및 조인하여 빠르게 결과를 도출하기 위해 노드의 부하 등을 고려한 실시간 잡 스케쥴링 기능이 필요하다. 또한 작업에서 많은 비용을 차지하는 디스크 I/O를 감소시키기 위하여 메모리를 사용하여 작업을 처리하는 인메모리 기반 처리 기법이나 자주 사용되는 데이터나 중간 결과를 재사용할 수 있도록 캐시나 실체 뷰를 활용하는 효율적인 질의 처리 기능이 필요하다.

## Ⅲ. 결론

본 논문에서는 고성능 자율 기계학습을 위한 인텔리전트 데이터베이스(이하 인텔리전트 DB) 플랫폼을 제안했다. 제안하는 기법에서는 인텔리전트 DB 플랫폼의 전체 시스템 구조를 설계하였으며, 제안하는 플랫폼을 이용하여 자율 기계학습을 수행하기 위한 기반 기술들을 도출했다. 향후 연구로는 인텔리전트 DB 플랫폼을 위한 기반 기술들에 대한 연구를 수행할 예정이다.

## ■ 참 고 문 헌 ■

[1] https://www.tensorflow.org/
[2] https://caffe2.ai/
[3] https://keras.io/