빅데이터 처리를 위한 Spark MLlib 성능 Benchmark

김용표, 오상윤*
아주대학교

primayy@ajou.ac.kr, *syoh@ajou.ac.kr

Benchmark on the Spark MLlib Benchmark for Big Data

Yong Pyo Kim, Sangyoon Oh*
Ajou Univ.

요 약

기계학습에서 학습시간은 일반적으로 데이터의 크기와 학습 모델의 복잡성에 영향을 받는다. 따라서 빅 데이터를 빠르게 처리할 수 있는 Spark 를 사용해 모델을 학습시킨다면, 프레임워크를 사용하지 않았을 때와 비교해 더 빠른 시간 내에 학습이 가능할 것이다. 본 연구에서는 Spark MLlib 에서 지원하는 기계학습 알고리즘을 이용하여 데이터 크기가 다른 두 종류의 데이터셋을 학습시킨다. 그리고 학습된 모델의 결과로부터 Spark MLlib 를 벤치마킹하고 데이터의 크기가 기계학습에 영향을 주는 정도를 확인한다.

Ⅰ. 서 론

빅데이터 처리에 필요한 대규모 연산을 위해 클러스터를 기반으로 분산 데이터 처리 환경에서 디스크 기반의 처리를 하는 Hadoop 과 인 메모리 기반으로 처리를 하는 Spark 프레임워크가 제안되었으며, 현재 빅데이터 처리를 위해 많은 응용 분야에서 이 두 프레임워크가 사용되고 있다. 빅데이터 처리를 위해 기계학습 알고리즘을 사용하는 경우, 학습시간은 일반적으로 데이터의 크기, 학습 모델의 복잡성에 영향을 받는다. 본 연구에서는 Spark 를 기계학습 기반의 빅데이터 처리에 사용하는 경우, 모델의 학습 시간과 데이터 및 모델의 복잡도와의 관계를 설명하기 위한 Benchmark 를 수행하고 결과를 분석 및 고찰한다.

Ⅱ. 관련 연구

Mehdi Assefi et al [1]은 본 연구의 Benchmark 목표인 Spark MLlib 에 대해, SVM, Random Forrest 및 Naïve Bayes 의 성능을 측정하였으며 이를 위해 UCI Machine Learning Repository [2]으로부터 얻은 5 개 데이터셋과 미 정부의 교통연구소(RITA) [3]로부터 얻은 데이터셋을 사용하였다. 다양한 실험 결과 분석을 통해 보편적으로 사용되는 Weka 와 비교하여 높은 성능으로 인한 상대적 장점을 가지는 것을 보였다.

또한 Fu et al [4]는 그들의 논문에서 Spark 프레임워크가 기계학습에 대해 가지는 장점을 보이기 위해, 스파크의 핵심기능과 지원하는 라이브러리에 대한 간단한 설명과 함께 Spark MLlib 에서 지원하는 linear regression model 의 성능분석을 수행하였다.

Ⅲ. Benchmark 실험 환경

1. 데이터 및 실험 환경

학습하려는 데이터의 크기에 따라 Spark 를 사용했을 때의 이점이 다르게 작용할 수 있기 때문에, 크기가 다른 두 개의 데이터셋을 사용해 실험을 수행했다.

표 1 실험 데이터

Amazon Fine Food Review	
Size	355.53MB
Number of Review	568,454
Feature (Column)	10
Amazon Movie Review	
Size	4.69GB
Number of Review	7,911,696
Feature (Column)	8

클러스터는 1 Master, 6 Slave node 로 구성되어 있으며, 각 노드 별 하드웨어 Specification 은 다음과 같다.

표 2 Specification of Cluster

CPU	Intel® Core™ i7-8700@ 3.20GHz, 12 cores
RAM	32GB
OS	Ubuntu 18.04.4 LTS
Hadoop	3.2.1
Spark	3.0.0 preview2

2. 데이터 전처리

데이터 전처리를 Spark 를 사용해 수행했다. 먼저, Review 의 Score Column 의 값이 3 이상이면 긍정으로 1, 3 보다 작으면 부정으로 0 Label 을 추가했다. 다음으로 주어진 문장에서 유의미한 단어만을 추출하기 위해 Text Column 의 데이터에서 html tag 와 불용어 그리고 특수문자를 제거했다. 다음으로 Clean text 로 변환한 문장을 단어로 분리하기 위한 Tokenize 과정을 거쳤다. 각 단어 별 등장하는 빈도수를 파악하기 위해 주어진 단어 집합의 단어를 카운트했다. 마지막으로 클러스터링 알고리즘(LDA, K-means, Bisecting K-means)을 이용할 때, 각 단어들마다 중요한 정도를 나타내는 TF-IDF 매트릭스를 위해 IDF 를 생성했다.

3. 모델 학습

데이터를 학습시키기 위해 Spark MLlib 에서 사용한 기계학습 알고리즘은 총 6 개이다. 클러스터 환경에서 병렬화의 효과가 큰 알고리즘이 무엇인지 알기 위하여 Classification 에 속하는 Multinomial Naïve Bayes

(MNB), Linear SVC 알고리즘과, Clustering 에 속하는 LDA, K-means, Bisecting K-means 알고리즘 그리고 Regression 에 속하는 linear regression 알고리즘을 사용해 결과를 측정했다. 각 알고리즘을 이용한 모델을 구현하기 위해 파이썬 환경에서 spark 를 사용할 수 있게 해주는 pyspark 를 사용했다. 각 알고리즘 수행의 핵심 파라메터 설정은 다음과 같으며, Linear Regresion 의 경우는 모두 default 설정값을 사용하였다.

표 3 모델 파라메터 설정

MNB (Smoothing)	1 (default)
Linear SVC (Regulation)	0.01
LDA (k)	1
K-means (k)	10
Bisecting K-means (k)	10

학습을 위해 데이터셋은 8:2 비율로 training set, test set 으로 나누었고, 학습에 사용하는 worker 수에 따른 학습시간의 차이를 확인하기 위해 클러스터의 worker 수를 1~6 개로 조절해가며 학습을 수행했다. 마지막으로 학습시간을 측정하기 위해 각 모델의 학습은 10 번씩 수행하여 최소 학습시간, 최대 학습시간, 평균 학습시간을 측정했다

Ⅳ. 성능 분석 결과

1. Amazon Fine Food Reviews Dataset

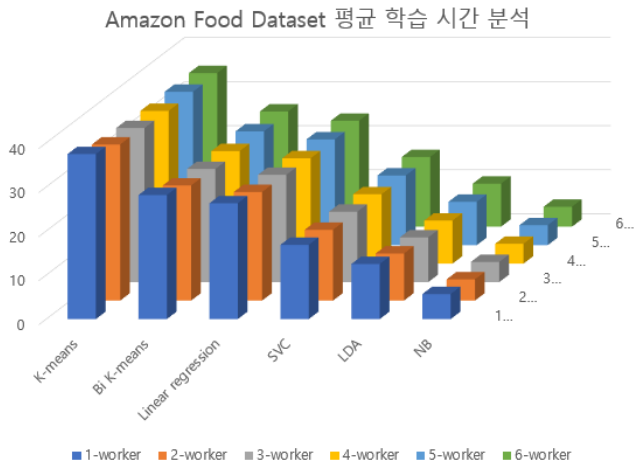

그림 1 Amazon Food Dataset 학습 시간 분석

Amazon Fine Food Reviews 데이터셋을 6 가지 기계학습 알고리즘을 이용해 학습시킨 결과, 모델 중 하나의 worker 를 사용했을 때를 기준으로 학습 시간이 가장 크게 줄어든 모델은 LDA 모델로 6 개의 worker 를 사용했을 때의 학습 시간이 하나의 worker 를 사용했을 때 보다 평균 2.78 초 단축(21.85%)되었다. 반대로 가장 적은 학습 시간 단축을 보여준 모델은 Linear SVC 모델로 4 개의 worker 를 사용했을 때 평균 1.09 초가 단축되었고, 이는 하나의 worker 를 사용했을 때에 학습 시간의 약 6.47%에 해당하는 시간이었다. Worker 의 수를 6 개까지 점차 늘려가며 평균 학습시간을 측정했을 때, 대부분의 기계학습 알고리즘 모델에서 worker 의 수가 4 개일 때 학습에 가장 적은 시간이 소요되었고, 이 이상으로 worker 의 수를 늘릴 경우 오히려 학습에 더 오랜 시간이 소요되는 모습을 확인할 수 있었다. 따라서 본 실험에서는 데이터의 크기가 작아 Spark 에서의 기계학습 진행에 대한 이점을 충분히 확인할 수 없었다고 판단된다.

2. Amazon Movie Reviews Dataset

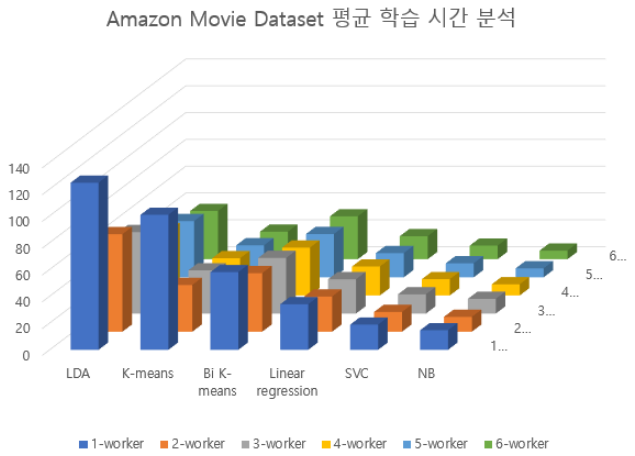

그림 2 Amazon Movie Dataset 학습 시간 분석

Amazon Movie Reviews 데이터셋을 6 가지 기계학습 알고리즘을 이용해 학습시킨 결과, 모델 중 하나의 worker 를 사용했을 때를 기준으로 학습 시간이 가장 크게 줄어든 모델은 K-means 모델로 6 개의 worker 를 사용했을 때의 학습시간이 하나의 worker 를 사용했을 때보다 평균 80.24 초가 단축되었다. 이는 하나의 worker 를 사용했을 때에 학습 시간의 약 79.63%에 해당하는 시간이었다. 반대로 가장 적은 학습 시간 단축을 보여준 모델은 Bisecting K-means 모델로 6 개의 worker 를 사용했을 때 25.99 초가 단축(44.83%)되었다. Worker 의 수를 6 개까지 점차 늘려가며 평균 학습시간을 측정했을 때, 사용한 모든 기계학습 알고리즘 모델에서 worker 의 수가 6 개일 때 학습에 가장 적은 시간이 소요되는 모습을 확인할 수 있었다. 본 실험에서는 Spark 를 통한 병렬화의 효과를 확인할 수 있었다.

Ⅴ. 분석 및 고찰

본 연구에서는 Spark 를 기계학습 작업에 사용하는 경우의 학습 시간을 측정함으로써 Spark MLlib 를 벤치마킹하는 것이었다. 학습에 사용한 데이터의 크기가 충분히 크지 않아 병렬도를 높였을 때의 효과를 완전히 확인하기 어려웠다. 이는 데이터의 크기가 더 Amazon Movie Reviews 데이터셋의 결과에서 병렬처리의 효과가 더 큰 것으로 확인하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의

SW 중심대학사업의 수행결과로 추진되었음 (2015-0-00908)

참 고 문 헌

[1] Mehdi Assefi et al, “Big Data Machine Learning using Apache Spark MLlib,” 2017 IEEE International Conference on Big Data, Dec. 2017.

[2] “UCI machine learning repository,” [Online] Available: http://archive.ics.uci.edu/ml/index.html.

[3] “US government’s bureau of transportation research and innovative technology administration (RITA),” [Online] Available: http://transtats.bts.gov

[4] Fu Jian et al, “SPARK – A Big Data Processing Platform for Machine Learning” International Conference on Industrial Informatics – ICIICI), Dec. 2016