# 데이터 분석 도구 성능 비교 연구 -기계 학습을 적용하여-

권태희*
숭실대학교 소프트웨어특성화대학원 소프트웨어전공
e-mail:kth7121@gmail.com

# A Performance Comparison Study on Data Analysis Tool -Applying Machine Learning-

Tae-Hee Kwon*
*Dept. of Software Graduate School of Soong-Sil University

요 약

빅데이터 시대가 도래되면서 과거와 비교할 수 없을 만큼의 방대하고 다양한 데이터가 생산됨에 따라 기존의 데이터 분석 도구의 사용은 한계에 부딪히게 되었다. 따라서 기존의 분석 도구보다 효율적이고 정확성이 높은 데이터 분석 도구를 필요로 하게 되었고, 빅데이터를 처리할 수 있는 분석 도구들에 대한 많은 연구들이 진행되어 왔다. R과 Apache Spark는 대표적인 데이터 분석 도구로 기계 학습을 위한 기능을 제공하고 있다.

본 논문에서는 기계 학습을 활용하여 두 개의 널리 알려진 데이터 분석 도구인 R과 Apache Spark의 데이터 분석 성능을 비교함으로써 보다 효율적이고 정확성이 높은 도구를 모색하고자 한다.

## 1. 서론

2016년에 발표된 알파고는 구글에서 개발된 인공지능 바둑 프로그램으로서, 알파고의 출현은 전 세계적으로 기계 학습에 대한 관심을 급증시키는 계기가 되었다. 많은 데이터 분석 도구들이 기계 학습을 위한 기능을 제공하고 있으며, 본 논문에서 소개할 R과 Apache Spark가 대표적이다. R은 주로 통계 기반의 데이터 분석을 수행하는 데에 사용되는 분석 도구로서 기계 학습을 위한 기능을 제공하고 있고, Apache Spark는 데이터 처리 및 분석을 위한 다양한 기능을 포함하고 있으며 R과 마찬가지로 기계 학습을 위한 기능을 제공하고 있다. 기계 학습과 같이, 국가나 기업, 그리고 단체는 그들의 의사결정에 있어서 데이터를 적극적으로 활용하고 있으며, 이로 인한 데이터의 활용도가 증가함에 따라 정확하고 효율적인 의사결정을 위한 데이터 분석 분야의 중요성 또한 높아지고 있다.

본 논문에서는 Apache Spark와 R의 데이터 분석 성능 비교를 위해, 빅데이터 분석 기법 중 기계 학습 앙상블 학습 방법인 랜덤 포레스트(Random Forest)방법을 선정한다. 랜덤 포레스트를 활용한 결과로 무엇이 더욱 효율적이고 정확성이 높은 데이터 분석 도구인지 결론으로 맺으며 마친다.

## 2. 데이터 분석 도구 및 분석 기법

데이터 분석에 의한 의사결정의 중요성이 높아지고 오픈소스 소프트웨어에 대한 관심이 급증하면서 전 세계적으로 IT기업들이 자체 기술을 오픈소스로 발표하고 있으며, 오픈소스로 개발되는 데이터 분석 도구들도 각광을 받고 있다. 대표적으로 통계기반의 오픈소스 데이터 분석 도구인 R이 널리 알려져 있으며, Apache Spark 또한 다양한 데이터 분석 관련 라이브러리를 내장한 오픈소스로 큰 관심을 받고 있다.

이 장에서는 대표적인 분석 도구인 R과 Apache Spark와 이 두 개의 분석 도구를 비교 분석할 방안으로 사용될 랜덤 포레스트에 대하여 설명한다.

### 2.1. R

R은 통계 기반의 데이터 분석 도구로 널리 알려진 오픈소스 소프트웨어이자 하나의 언어로서, 벨 연구소에서 만들어진 통계 분석 언어인 S를 기반으로 만들어졌다. 최근 하둡과 Apache Spark 등과 같은 오픈소스 기반의 빅데이터 처리 및 분석 도구들뿐만 아니라 상용 소프트웨어도 R을 지원하고 있는 추세가 증가하고 있어 R에 대한 관심은 더욱 증가하고 있다. 대표적으로 RHive의 경우를 살펴보면, RHive는 넥스알이라는 기업이 Apache Software 재단에서 주관하는 빅데이터 분산처리기술인 하둡용 데이터 웨어하우스 Hive와 R을 결합하여 만든 빅데이터 분석 도구이다. RHive는 오픈소스로 공개되어 제공되고 있으며, 빅데이터 고급분석 통계 소프트웨어로 널리 알려져 많은 분야에서 활용되고 있다.

R은 통계, 기계 학습, 데이터 마이닝, 금융 분야 등에 이르는 5000개 이상의 다양한 통계 패키지를 내장하고 있

으며 누구에게나 이 패키지를 무료로 제공하고 있다. R에서 제공하고 있는 패키지는 데이터 분석뿐만 아니라 데이터 분석 결과의 시각화를 위한 기능들도 포함하고 있다. R의 오픈소스 소프트웨어라는 특징으로 인해 많은 패키지들이 커뮤니티 형태로 사용자들에 의해 개발되어 공개되고 있으며, 이로 인해 사용자 수가 증가하고 용도의 범위가 확장되고 있다.

R은 언어이지만 명령어 형태가 직관적이기 때문에 사용자가 배우기 쉽고, 윈도우, 리눅스, 맥 OS X에서 모두 지원하기 때문에 사용 시에 운영체제에 대한 제한이 크지 않다. 그리고 R은 정수, 부동소수, 문자열뿐만 아니라 데이터 처리에 용이한 벡터, 행렬, 리스트, 데이터 프레임과 같은 데이터 형태를 제공하므로 사용자가 데이터 분석 수행 시에 더욱 효율적으로 데이터 처리 및 분석이 가능하다.

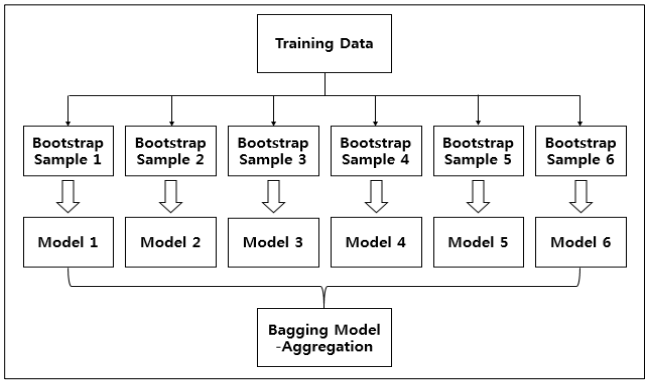
### 2.2. Apache Spark

Apache Spark는 Apache Software 재단에서 주관하고 있는 오픈소스 프로젝트 중 하나로서, 반복적인 대화형 연산 작업에서 비효율적인 맵리듀스의 단점을 보완하도록 설계된 클러스터용 연산 플랫폼이다. Apache Spark는 데이터 처리에 있어 인-메모리 기법을 적용하여 맵리듀스보다 빠른 속도로 연산 작업을 수행할 수 있고 맵리듀스 모델을 대화형 명령어 쿼리(Query)나 스트리밍(Streaming) 처리 등이 가능하도록 확장하였다. 그리고 기존에 각각 분리된 분산 시스템에서 작동되던 배치 애플리케이션(Batch Application), 반복 알고리즘(Iterative Algorithm), 대화형 쿼리, 스트리밍과 같은 다양한 작업들을 단일 시스템에서 수행할 수 있도록 지원하고 있다. 사용자들은 스파크에 내장되어 있는 파이썬(Python), 자바(Java), 스칼라(Scala), SQL API 그리고 다양한 라이브러리를 사용할 수 있으며 하둡과 카산드라 같은 다른 빅데이터 도구들과도 연동하여 활용할 수 있다. Apache Spark는 스파크SQL, 스파크 스트리밍, MLlib, 그래프X와 같은 기능들을 지원하고 있다. 이 중에서 MLlib는 기계 학습 기능들을 포함하고 있는 라이브러리로서 분류, 회귀, 클러스터링(Clustering), 협업 필터링(Collaborative Filtering) 등의 다양한 유형의 기계 학습 알고리즘뿐만 아니라 모델 평가 및 외부 데이터 불러오기와 같은 기능들도 제공한다.

### 2.3. 랜덤 포레스트

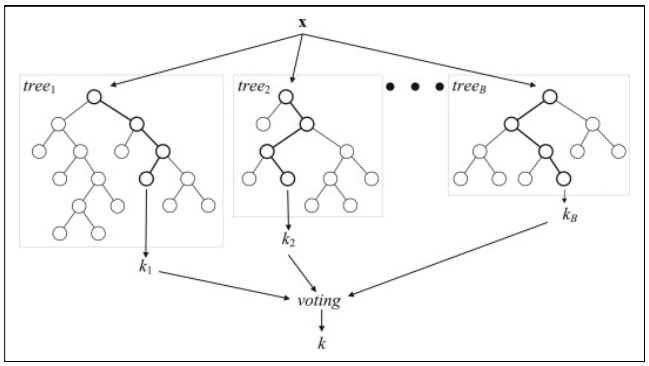
앙상블 학습이란 주어진 데이터로부터 여러 개의 알고리즘을 학습한 다음, 예측작업을 수행할 때 여러 알고리즘의 예측 결과들을 종합적으로 사용해서 작업의 정확도를 높이는 기법을 의미한다. 랜덤 포레스트 알고리즘은 기계 학습에서 지도 학습의 알고리즘 중 하나로서 앙상블 학습 기법을 사용한 알고리즘이며, 2001년 레오 브레이먼(Leo Breiman)에 의해 발표되었다.

랜덤 포레스트 알고리즘은 다수의 의사결정나무를 구성하는 학습 단계와 입력된 데이터를 분류하거나 예측하는 테스트 단계로 구성된다. 하나의 훈련 데이터로 학습 단계를 진행하는 의사결정나무 알고리즘과 달리, 랜덤 포레스트 알고리즘은 배깅이라는 기계 학습의 메타 알고리즘을 적용하여 다수의 훈련 데이터로 학습 단계를 진행한다. 배깅이란 부트스트랩(Bootstrap) 결합(Aggregating)의 약자로서 부트스트랩을 통해 생성된 다수의 훈련 데이터로 모델들을 학습시키고 학습된 모델들을 결합하는 앙상블 기법이다. 부트스트랩이란 기존의 훈련 데이터로부터 다수의 새로운 훈련 데이터를 복원 추출하는 과정이며 이를 통해 분산을 줄일 수 있다. 따라서 배깅은 모델의 분산을 줄여줌으로써 모델의 예측력을 향상시키는 기법이라 할 수 있기 때문에 일반적으로 과적합된 모형이나 분산이 큰 모델에 적합하다.




출처: DNI Institute에서의 인용

(그림 1) 배깅 알고리즘의 원리




(그림 2) 랜덤 포레스트 알고리즘 개념도

배깅 알고리즘의 원리를 나타내고 있는 (그림 1)을 살펴보면, 기존의 훈련 데이터인 Training Data로부터 부트스트랩을 이용하여 새로운 훈련 데이터들인 Bootstrap Sample들을 추출한다. 추출한 새로운 훈련 데이터들은 각각의 Model들을 학습시키고 학습된 Model들을 결합하여 최종 예측 모델인 Bagging Model-Aggregation을 수립한다. (그림 2)를 살펴보면, 배깅 과정을 통해 B개의 의사결정나무 알고리즘들인 tree들이 생성되고 새로운 데이터 x가 각 의사결정나무 알고리즘으로 입력된다. 그러면 각각

의 의사결정나무 알고리즘에서 B개의 결과 k들이 산출되며, 산출된 결과들로 과반수 투표 방식을 수행하여 입력 데이터 x에 대한 예측 결과인 k가 최종적으로 도출된다. 각각의 의사결정나무 알고리즘에서 산출된 결과들은 과반수 투표 방식뿐만 아니라 평균이나 곱하기와 같은 방식을 이용하여 최종 결과를 도출하기도 한다.

랜덤 포레스트 알고리즘이 배깅을 적용하여 알고리즘을 학습시키는 방법이므로 배깅의 특징이 랜덤 포레스트의 특징이라 말할 수 있다. 랜덤 포레스트의 또 다른 특징은 랜덤 포레스트 알고리즘에서 생성된 의사결정나무 모델 간의 상관성(Correlation)이 낮을수록 즉, 서로 독립일수록 예측오차가 작다는 점이다(Breiman, 2001). 의사결정나무는 작은 편향(Bias)과 큰 분산을 갖는 것이 특징인데 매우 깊은 의사결정나무는 훈련 데이터에 대해서 과적합 하게 될 가능성이 높아진다. 부트스트랩은 편향을 유지하면서 분산을 줄이고, 배깅은 서로 다른 훈련 데이터로 의사결정나무들을 학습시키면서 각각의 의사결정나무들 간의 상관성을 작게 만들어 불완전한 훈련 데이터에 대해서도 민감하지 않게 만든다. 그리고 이러한 특징들은 랜덤 포레스트 알고리즘의 성능을 향상시킨다.

## 3. 실험

### 3.1. 실험 방법

가상머신인 VMware Workstation을 사용하여 Ubuntu 운영체제에서 실험을 진행한다. 실험에서 진행할 데이터는 CSV 형식이며, 자동차와 관련된 7개의 속성에 대한 데이터로 이루어져 있다. 7개의 속성은 아래 <표 1>과 같으며 6개의 속성을 사용하여 자동차의 class를 분류하는 작업을 수행한다.

<표 1> 실험 데이터의 속성

| 속성 | 범위 |
|---|---|
| buying | low, med, high, vhigh |
| maint | low, med, high, vhigh |
| doors | 2, 3, 4, 5more |
| persons | 2, 4, more |
| lug_boot | small, med, big |
| safety | low, med, high |
| class | ACC, GOOD |

기존 데이터를 7:3의 비율로 나누어 실험을 위한 Training Data와 Test Data를 정하며, Training Data를 사용하여 랜덤 포레스트 모델을 학습시킨다. 그리고 각 학습을 통하여 생성된 모델의 성능 비교를 위해서, Test Data를 사용하여 분류에 대한 오차율을 계산하고 비교하여 학습된 모델들을 평가한다.

랜덤 포레스트 알고리즘을 활용한 데이터 분석을 위해서 R의 'randomForest' 패키지와 Apache Spark의 Spark MLlib 기능 안에 포함된 'Random Forest' 클래스를 활용한다. R의 실험은 Rstudio라는 R 언어 기반의 IDE에서 진행되고, Apache Spark의 실험은 IntelljJ라는 Java IDE에서 Scala 언어를 사용하여 수행된다.

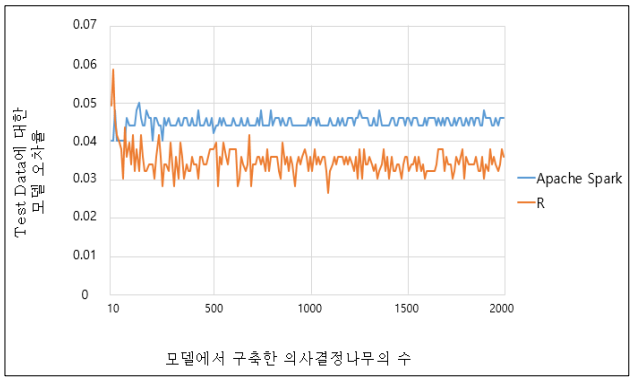
### 3.2. 실험 환경

<표 2>는 실험이 진행된 환경을 나타내고 있다.

<표 2> 실험 환경

| 구성요소 | 규격 |
|---|---|
| 가상머신 | VMware Workstation 11.1.0 |
| 운영체제 | Ubuntu 14.04.5 LTS (64-bit) |
| CPU | Intel Core i5-6500 3.20GHz |
| Java | 1.8.0_101 (64-bit) |
| Scala | 2.11.7 |
| R | 3.0.2 |
| Rstudio | 0.99.489 |
| Apache Spark | 1.6.2 |
| IntelliJ IDEA | 2016.2.4 |

### 3.3. 실험 결과




(그림 3) R과 Apache Spark 실험의 모델 오차율 그래프

<표 3> R과 Apache Spark 실험의 모델 오차율 수치

| 도구 / 의사결정나무 수 | R | Apache Spark |
|---|---|---|
| 500 | 0.03592 | 0.04409 |
| 1000 | 0.03592 | 0.04409 |
| 1500 | 0.03592 | 0.04409 |
| 2000 | 0.03592 | 0.04610 |

<표 4> R과 Apache Spark 실험의 모델 구축 시간

(단위: 초)

| 도구 / 의사결정나무 수 | R | Apache Spark |
|---|---|---|
| 500 | 1.140 | 4.612 |
| 1000 | 2.136 | 4.630 |
| 1500 | 3.122 | 5.607 |
| 2000 | 4.194 | 7.088 |

(그림 3)은 모델 학습 과정에서 구축한 의사결정나무의 개수에 따른 Test Data에 대한 모델의 분류 오차율을 나

타낸 그림이다. <표 3>은 (그림 3)에 표시된 오차율 중에서 대표적으로 구축된 의사결정나무가 500, 1000, 1500, 2000개 일 때의 오차율 수치를 소수점 5자리까지 반올림하여 나타내고 있는 자료이다. <표 4>는 <표 3>에서와 동일한 경우일 때의 랜덤 포레스트 모델 구축 시간에 대한 내용이다.

(그림 3)과 <표 3>을 통해서 R의 오차율이 Apache Spark보다 대략 0.01정도 낮은 것을 알 수 있으며, 랜덤 포레스트 알고리즘에 대해서 R이 Apache Spark보다 정확한 분류 성능을 가지고 있다고 판단할 수 있다. 그리고 <표 4>를 통해서 알 수 있듯이, 랜덤 포레스트 알고리즘에 있어서 Apache Spark보다 R의 처리 속도가 더 빠르다는 것을 볼 수 있다. 그러므로 본 논문에서는 Apache Spark보다 R이 더욱 정확하고 효율적인 데이터 분석 도구로 볼 수 있다.

## 4. 결론

단시간에 엄청난 양의 데이터가 폭발적으로 생산되는 환경에서 데이터의 힘은 점점 커지고 있다. 그리고 데이터에 대한 정확한 분석은 그 무엇보다 확실한 경쟁전략을 수립할 수 있다. 실제로 여러 기업들은 자신의 고객데이터를 활용하여 고객을 분석하고 고객 유지 전략을 수립하는데, 이러한 행동은 기업의 시장에서의 우위와 경제성에 직결된다. 데이터에 의한 의사결정의 영향력이 매우 넓고 중요하다는 것을 짐작할 수 있다.

데이터 분석 도구를 개발하는 많은 기업들은 R과 Apache Spark를 포함한 다양한 오픈소스 소프트웨어들을 활용하여 경쟁력 있는 자신들만의 솔루션을 개발하고 있으며, 개발된 솔루션들은 위에서 언급한 기업들을 포함하여 다양한 분야에서 데이터에 의한 의사결정을 돕고 있다. 그러므로 본 논문에서 수행한 데이터 분석 도구 성능 비교 실험은 보다 효율적이고 경쟁력 있는 데이터 분석 솔루션을 개발하고 정확한 의사결정을 내리는 데에 도움을 줄 수 있다고 판단된다.

현대에는 빅데이터 처리, 저장 및 분석을 위해서 대부분 분산 컴퓨팅 방식을 채택하고 있다. 그러나 본 논문의 실험은 단일 컴퓨터에서 수행되었고, 알고리즘 중에서 분류 분석 알고리즘인 랜덤 포레스트에 대해 진행되었다. 이 점은 본 논문의 한계점이라 할 수 있으므로 향후 연구에서는 분산 컴퓨팅 환경에서 다양한 종류의 알고리즘에 대한 실험을 진행하도록 한다.

## 참고문헌

[1] 서민구, R을 이용한 데이터 처리&분석 실무, 길벗(2014.10.31.), p488～p494.

[2] 홀든 카로, 앤디 콘빈스키, 패트릭 웬델, 마테이 자하리아, 러닝 스파크, 제이펍(2015.10.15.).

[3] 브레트 란츠, R을 활용한 기계 학습(데이터 분석을 위한 머신 러닝 이론과 적용), 에이콘(2014.9.23.), p30～p49.

[4] 유진은, “랜덤 포레스트 의사결정나무의 대안으로서의 데이터 마이닝 기법”, 2015.

[5] 황보선, 박철수, 장병탁, “상호 정보량을 이용한 동작 상상 뇌 신호 구분 방법론”, 『2014년 동계학술발표회 논문집』, 2014, p.614