Map-Reduce 프로그래밍 모델 기반의 나이브 베이스 학습 알고리즘

강대기

동서대학교

Naive Bayes Learning Algorithm based on Map-Reduce Programming Model

Dae-Ki Kang

Dongseo University

E-mail : dkkang@dongseo.ac.kr

요 약

본 논문에서는, 맵-리듀스 모델 기반에서 나이브 베이스 알고리즘으로 학습과 추론을 수행하는 방안에 대해 소개하고자 한다. 이를 위해 Apache Mahout를 이용하여 분산 나이브 베이스 (Distributed Naive Bayes) 학습 알고리즘을 University of California, Irvine (UCI)의 벤치마크 데이터 집합에 적용하였다. 실험 결과, Apache Mahout의 분산 나이브 베이스 학습 알고리즘은 일반적인 WEKA의 나이브 베이스 학습 알고리즘과 그 성능면에서 큰 차이가 없음을 알 수 있었다. 이러한 결과는, 향후 빅 데이터 환경에서 Apache Mahout와 같은 맵-리듀스 모델 기반 시스템이 기계 학습에 큰 기여를 할 수 있음을 나타내는 것이다.

ABSTRACT

In this paper, we introduce a Naive Bayes learning algorithm for learning and reasoning in Map-Reduce model based environment. For this purpose, we use Apache Mahout to execute Distributed Naive Bayes on University of California, Irvine (UCI) benchmark data sets. From the experimental results, we see that Apache Mahout's Distributed Naive Bayes algorithm is comparable to WEKA's Naive Bayes algorithm in terms of performance. These results indicates that in the future Big Data environment, Map-Reduce model based systems such as Apache Mahout can be promising for machine learning usage.

키워드

Map-Reduce model, Apache Mahout, Distributed Naive Bayes

Ⅰ. 서 론

우리가 분석을 수행해야 할 데이터는 그 양이 점점 커지고 있다. 예를 들어 트위터와 같은 소셜 네트워크 서비스에서 다수의 팔로잉과 팔로워를 가질 경우, 그 데이터의 복잡함은 매우 높다고 할 수 있다. 이러한 데이터 구성을 척도 없는 네트워크라고 부르면, 이에 대한 복잡계 분석 기반의 연구도 매우 활발하다[1]. 이러한 데이터들의 양은 매우 커져서, 이제 빅 데이터라는 이름으로 불리우고 있다.

데이터 마이닝과 지식 발견 기술은 대형 데이터베이스에서 자동으로 숨겨진 규칙을 찾아내는데, 매우 효율적인 것으로 알려져왔다[2]. 그러나, 데이터 마이닝이나 기계 학습 알고리즘들은 한 개의 작은 테이블로 구성된 데이터들에 대해서만 적용되어 왔다. 만일, 현실 세계의 이른바 빅 데이터에 대해 적용될 수 있는 기계 학습 알고리즘이 있다면, 이에 대한 연구는 기계 학습 및 데이터 마이닝 분야에 큰 발전을 가져올 수 있다.

본 논문에서는 이러한 빅 데이터 환경에 대한 컴퓨터의 데이터 모델 중 하나인 맵-리듀스 모델[3]에 기반한 Apache Mahout[4]를 소개하고, 이 시스템의 대표적인 기계 학습 알고리즘 중 하나인 분산 나이브 베이스 (Distributed Naive Bayes) 학습 알고리즘을, 역시 대표적인 기계 학습 벤치마크 데이터 중 하나인 University of California, Irvine (UCI)의 벤치마크 데이터 집합[5]에 적용하여 보았다.

II. 본 론

Apache Mahout은 Apache Software Foundation (ASF)에서 추진 중인 새로운 오픈 소스 프로젝트이다. 주요한 목적은 확장 가능한 기계 학습 알고리즘을 만드는 것이며 Apache 라이센스가 있으면 무료로 사용 가능하다.

Apache Mahout에는 클러스터링, 분류, 협업 필터링 (CF) 및 진화 프로그래밍을 위한 구현이 포함되어 있다. 게다가 Apache Hadoop 라이브러리를 사용하면 클라우드에서 Apache Mahout을 효과적으로 확장할 수도 있다.

Apache Mahout의 주요 기능은 다음과 같다.

- Taste Collaborative filtering (CF). - Taste는 SourceForge의 Sean Owen에 의해 시작된 CF를 위한 오픈 소스 프로젝트로 2008년에 Mahout으로 귀속됨
- k-Means, fuzzy k-Means, Canopy, Dirichlet 및 Mean-Shift를 포함한 여러 가지 Map-Reduce 사용 클러스터링 구현
- Distributed Naive Bayes 및 Complementary Naive Bayes 분류 구현
- 진화 프로그래밍을 위한 분산 적합성 함수 기능
- 행렬 및 벡터 라이브러리
- 위 모든 알고리즘의 예제

본 연구에서는 이러한 Apache Mahout를 대표적인 기계 학습 벤치마크 데이터 중 하나인 University of California, Irvine (UCI)의 벤치마크 데이터 집합 중 37 개에 적용하여 보았다. 또한 같은 데이터 집합들을 WEKA 기계 학습 알고리즘[6]에 적용하였다. 사용한 기계 학습 알고리즘은 Naive Bayes이며, 적용한 데이터의 집합은 다음과 같다.

Anneal, Audiology, Autos, Balance-scale, Breast-cancer, Breast-w, Car, Colic, Credit-a, Credit-g, Dermatology, Diabetes, Glass, Heart-c, Heart-h, Heart-statlog, Hepatitis, Hypothyroid, Ionosphere, Iris, Kr-vs-kp, Labor, Letter, Lymph, Mushroom, Nursery, Primary-tumor, Segment, Sick, Sonar, Soybean, Splice, Vehicle, Vote, Vowel, Waveform-5000, Zoo

III. 실험 결과 및 결론

성능 평가를 위해 각 데이터 집합에 대한 분류 정확도(accuracy)를 측정하였다. 실험 결과, Apache Mahout는 WEKA와 그 성능에 있어 별다른 차이가 없음을 알 수 있었다. 이러한 결과는, 향후 빅 데이터 환경에서 Apache Mahout와 같은 맵-리듀스 모델 기반 시스템이 기계 학습에 큰 기여를 할 수 있음을 나타내는 것이다.

감사의 글

본 연구는 지식경제부의 지원을 받는 동서대학교 유비쿼터스 지역혁신센터의 연구결과로 수행되었습니다.(No. B0008352)

참고문헌

[1] 강병남, *복잡계 네트워크 과학 : 21세기의 정보과학*, 집문당, 2010.

[2] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques,* 2nd ed., 2006.

[3] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *OSDI'04: Sixth Symposium on Operating System Design and Implementation,* San Francisco, CA, December, 2004.

[4] Cheng T. Chu, Sang K. Kim, Yi A. Lin, Yuanyuan Yu, Gary R. Bradski, Andrew Y. Ng, Kunle Olukotun, "Map-Reduce for Machine Learning on Multicore," In NIPS 2006, pp. 281-288, 2006.

[5] Frank, A. & Asuncion, A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations, Volume 11, Issue 1,* 2009.