The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.6.999
29(6), 999–1005

Current trends in high dimensional massive data analysis

Woncheol Jang[a,1] · Gwangsu Kim[b] · Joungyoun Kim[c]

[a]Department of Statistics, Seoul National University
[b]Data Science for Knowledge Creation Research Center, Seoul National University
[c]Department of Information Statistics, Chungbuk National University

(Received October 19, 2016; Revised October 19, 2016; Accepted October 19, 2016)

Abstract

The advent of big data brings the opportunity to answer many open scientific questions but also presents some interesting challenges. Main features of contemporary datasets are the high dimensionality and massive sample size. In this paper, we give an overview of major challenges caused by these two features: (1) noise accumulation and spurious correlations in high dimensional data; (ii) computational scalability for massive data. We also provide applications of big data in various fields including forecast of disasters, digital humanities and sabermetrics.

Keywords: big data, computation scalability, digital humanities, noise accumulation, spurious correlations

1. 서론

21세기는 데이터의 시대라고 해도 과언이 아닐 정도로 우리는 지금 데이터 홍수시대에 살고 있다. 전직 구글 최고경영자인 Eric Shmidt는 2003년까지 인류가 만들어 낸 자료의 크기는 5 exabytes정도로 추정되는데 요즘 이정도 데이터는 이틀에 한번 꼴로 생성된다고 지적했다. 이런 빅데이터는 산업전반과 우리생활에 많은 혁신과 변화를 가져왔다. 빅데이터의 시대에 데이터 못지않게 중요한 것은 데이터의 분석능력이다. Gartner Research의 부사장인 Peter Sondergaard은 "정보가 21세기의 기름이라면 분석은 연소엔진"이라는 말로 분석의 중요성을 강조했다.

빅데이터의 등장은 데이터 과학으로 일컬어지는 데이터 관련 학문분야인 통계학, 응용수학, 전산학 교육과 연구에 큰 변화를 가져왔다. 미국의 경우 2012년에 대통령 주도로 빅데이터의 연구와 개발 계획이 발표되었으며 수많은 교육기관에서 "데이터 사이언스" 전문석사 과정이 개설되고 있다.

사실 빅데이터의 경우 통일된 정의는 존재하지 않고 IBM에서 제안한 4V로 일컬어지는 크기(volume), 속도(velocity), 다양성(variety), 정확성(veracity)을 빅데이터의 주요특징으로 들고 있다. 통계학적 관

Woncheol Jang's research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A4A1007895).
[1]Corresponding author: Department of Statistics, Seoul National University, Gwanak-ro 1, Gwanka-Gu, Seoul 08826, Korea. E-mail: wcjang@snu.ac.kr

점에서 이러한 특징을 가지는 빅데이터를 크게 "길쭉한(massive)" 과 "뚱뚱한(high-dimensional)" 자료로 나눌 수 있다. 예를 들면 신용카드 거래자료의 경우 수백만 명의 카드 소지자의 다양한 소비성향을 나타내며 이러한 자료를 사용자를 열로, 카드 사용처를 행으로 정리한다면 길쭉한 자료가 된다. 반면에 특정질병과 유전자의 관계를 알고자 유전자발현자료를 살펴볼 경우 백명 남짓의 사람들에게서 대략 2만개 정도의 유전자의 발현 정도를 알아보는 것이 일반적으로 이 경우는 행의 길이가 훨씬 긴 뚱뚱한 자료가 될 것이다. 대용량과 고차원으로 대표되는 빅데이터의 두가지 특징은 빅데이터 분석에 있어서 기존 분석방법의 한계를 노출시켰으며 최근 10여년간의 통계학 분야의 연구는 이러한 한계를 극복하는데 초점을 맞추고 있다.

본 논문에서는 빅데이터 연구를 크게 두 주제로 나누어서 다룬다. 첫 번째는 빅데이터의 분석 방법에 관한 내용, 두 번째는 빅데이터의 응용사례에 관해 초점을 맞추었다. 첫 번째 경우의 분야로는 기계학습, 최적화, 계산확장성, 자료시각화 등을 들 수 있으며 두 번째 응용사례는 천문학, 뇌인지과학, 생물정보학, 입자물리, 개별화 의료, 소셜 네트워크, 재난 예측, 스포츠 통계, 디지털 인문학 등을 들 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 고차원 자료분석의 주요 문제점인 소음축적과 위 상관관계에 대해 살펴보며 3절에서는 대용량 자료처리의 핵심인 계산확장성에 대해 알아본다. 4절에서는 빅데이터의 다양한 응용분야를 소개하며 5절에서 맺음말을 제시한다.

2. 고차원 자료의 특징

고차원 자료의 가장 큰 특징으로는 자료의 크기 n보다 차원의 크기 p가 크다는 데 있다. 이러한 고차원 자료에서 생기는 주요특징으로 소음 축적(noise accumulation)과 위 상관관계(spurious correlation)가 있으며 이러한 고차원 자료분석에 가장 많이 사용되는 방법은 차원축소와 벌점화를 통한 성김(sparsity)을 들 수 있다.

2.1. 소음 축적과 위 상관관계

일반적으로 고차원자료분석에서 성김현상을 가정한다. 즉 대부분의 변수들은 소음(noise)이며 예측이나 분류를 위한 의미있는 변수의 갯수는 상대적으로 적다고 가정하는 것이다. 예를 들면 마이크로 어레이에서 특정 질병과 관련있는 유전자의 갯수는 극소수라고 가정하자. 이 경우 모든 유전자를 이용하여 로지스틱 회귀분석을 하는 것은 비효율적이다. 즉 특정질병과 관련이 있지 않은 유전자들의 경우 소음으로 작용하여 많은 관련없는 유전자가 모형에 들어가 있다면 소음축적으로 모형의 분류오차가 커지게 됨을 알 수 있다. 이러한 소음축적의 해결책으로는 변수 선택을 들 수 있다. 하지만 고차원에서의 변수선택은 위 상관관계로 인해 어려운 문제로 간주될 수 있다.

위 상관관계는 서로 관련이 없는 확률변수들이 고차원상에서 실제로 높은 표본상관계수 값을 가지는 현상을 말한다. Fan과 Lv (2008)는 실제로 과학적으로 서로 연관이 없는 변수와 중요한 변수들이 높은 위 상관관계를 가질 수 있음을 보였다. 가장 간단한 예로 p차원 벡터 $\mathbf{Y} = (Y_1, \ldots, Y_p) \sim N_p(\mathbf{0}, \mathbf{I}_p)$가 p차원 정규분포를 따른다고 가정하고 우리가 $\mathbf{y}_1, \ldots, \mathbf{y}_n$를 관측한다고 하면 첫 번째 변수와 나머지 변수간의 표본상관관계의 최대 절대값을 다음과 같이 정의할 수 있다.

$$r = \max_{j \geq 2} \left| \widehat{\mathrm{Cor}}(Y_1, Y_j) \right|.$$

컴퓨터 모의실험을 할 경우 p가 커질수록 r값도 점점 높아진다는 것을 쉽게 볼 수 있다. 따라서 이러한 위 상관관계는 변수선택 시 심각한 영향을 끼쳐 잘못된 모형선택으로 이어지게 할 수 있다. 이러한 위 상관관계로 인한 영향을 방지하기 위해서는 스크린을 통한 차원 축소가 요구된다 (Fan과 Lv, 2008).

2.2. 차원축소와 벌점화

차원축소는 고차원자료분석에서 가장 많이 사용되는 분석방법으로 자료시각화를 위한 전초작업으로 많이 사용된다. 주성분분석은 차원축소에 사용되는 대표적인 도구이다. 고차원 자료에서 주성분분석을 실행하는데 가장 큰 어려운 부분은 singular value decomposition의 구현이다. Witten과 Candès (2013)은 random matrix projection을 이용하여 singular value decomposition을 근사하였다.

벌점화를 통한 성긴 모형(sparse model)을 구현하는 것도 저차원 모형 구조를 찾아내는 대표적 방법이다. Lasso (Tibshirani, 1996)는 벌점화의 대표적인 방법으로 이외의 대표적인 벌점화 방법으로 SCAD (Fan과 Li, 2001)와 MCP (Zhang, 2010)을 들 수 있다. Lasso 추정치에 관한 추론을 위해 van de Geer 등 (2014)은 편의제거를 위한 추가과정을 제안하였다.

3. 계산확장성

빅데이터의 두 번째 유형인 대용량 자료의 경우 그 크기로 인해 전통적인 분석방법과 컴퓨팅 기반시설에 일대 혁신을 가져왔다. 이 절에서는 이러한 대용량 자료의 처리를 위한 컴퓨팅 환경과 빅데이터 분석에 있어서 중추적 역할을 하는 최적화이론의 최근 동향에 대해서 알아본다.

3.1. 컴퓨팅 환경

대용량 자료 처리의 가장 기본적인 접근방법은 "divide and conquer"이다. 즉 보다 작은 단위로 잘게 나눈 후 병렬 컴퓨팅을 이용하여 분석하는 방식이다. 하지만 데이터의 크기가 굉장히 큰 경우에는 이러한 방법을 적용하기 힘들 수 있다. 예를 들면 수백만개의 core process가 동원된 경우 이 중 일부가 컴퓨팅 중에 작동을 멈출 수도 있다. 뿐만 아니라 이런 방대한 계산량을 공평하게 나누어 분배하는 것도 쉬운일은 아니다. 이러한 대용량 계산을 위한 기반시설로 하둡(hadoop)과 맵리듀스(Mapreduce)를 들 수 있다. 하둡은 자바기반의 분산자료 파일시스템이며 맵리듀스는 구글에서 개발한 분산 프로그래밍 모형이다. Park 등 (2013)은 하둡과 맵리듀스에 대한 자세한 리뷰와 실제 계산알고리즘을 맵리듀스로 구현한 예를 제시하고 있다. 하지만 맵리듀스 사용시 매 작업마다 디스크에서 자료를 입출력해야 한다는 문제점 때문에 클러스트 플랫폼인 아파치 스파크가 최근들어 각광받고 있다. 관심있는 독자는 Ko과 Won (2016)을 참조하기 바란다.

클라우드 컴퓨팅은 빅데이터의 저장과 처리를 값싼 가격에 제공해주는 환경으로 대표적인 클라우드 컴퓨팅 서비스로는 Amazon Web Services(AWS)를 들 수 있으며 아파치 스파크도 지원하고 있다.

3.2. 최적화

최적화는 벌점화 방법을 이용한 모형선택, 딥 러닝, 신경망 분석 등 빅데이터 분석방법에서 핵심적인 요소이다. 많은 경우 빅데이터 분석에서 최적화문제는 비 볼록(non-convex) 최적화문제로 상당히 어려운 최적화 문제로 간주된다. 하지만 경우에 따라 비 볼록 최적화 문제를 볼록 최적화 문제로 이완하는 것이 가능하다. 예를 들면 변수선택에 있어서 변수의 갯수에 대한 벌점을 주는 경우는 L_0 벌점화 함수이지만 대신 회귀계수의 절대값의 합에 벌점을 주는 L_1 벌점화 함수로 이완할 수 있으며 이렇게 만들어진 L_1 벌점화 함수를 이용한 모형선택방법이 Lasso이다. Lasso의 경우 최적화를 통한 조절모수의 선택이 훨씬 쉽지만 이 과정에서 모수 추청치에 편의(bias)가 들어가게 된다.

가장 많이 사용되는 최적화 방법은 gradient decent 알고리즘이지만 종종 gradient의 계산이 너무 오래 걸리는 단점이 있다. 이를 보완하기 위해 pathwise coordinate descent 알고리즘이 제안되었는데 기본

적인 아이디어는 gradient를 모든 방향에서 계산하는 대신 순차적으로 한 축에서만 gradient를 계산하여 최적화에 사용하는 것이다.

최근 비 볼록 벌점화 함수를 바탕으로 하는 변수선택 방법들이 많이 제안되었으며 이러한 비 볼록 최적화문제를 풀 수 있는 알고리즘으로는 local quadratic 알고리즘을 들 수있다 (Fan과 Li, 2001).

4. 응용사례

고차원 대용량 자료의 적용분야는 매우 다양하며 이번 응용통계특집호에서는 특별히 천문학, 뇌인지과학, 생물정보학, 디지털인문학 등의 분야에 대한 리뷰논문과 사례연구를 소개하고 있다. 이 절에서는 국내 통계학분야에서 상대적으로 덜 알려진 재난예측, sabermetrics, 디지털 인문학에 대해서 알아본다.

4.1. 재난예측

우리나라도 얼마 전 규모 5.8의 경주 지진을 계기로 대형 재해 재난에 대한 경각심이 높아졌다. 세계적으로는 9/11 테러를 시발점으로 대규모 테러 또는 자연재해의 예측에 대한 관심이 높아졌다. 지진과 같은 대형재해의 경우 건물들의 내진 설계의 기준을 위해 규모별 발생확률을 예측하는 것이 중요하다. 이러한 extreme event에 대한 확률모형은 일반적으로 멱함수(power law) 모형을 이용한다. Clauset 등 (2009)은 멱함수의 모수추정을 위해 물리학분야에서 많이 사용하고 있는 최소제곱법의 문제점을 지적하고 우도비(likelihood ratio)와 Kolmogorov-Smirnov 적합도 검정을 결합하여 보다 일반적인 멱함수에서 모수추정 방법을 제시하였다. Clauset과 Woodard (2013)은 이 방법을 이용하여 9/11과 같은 대규모 테러가 일어날 확률을 계산했는데 최소 10명이상의 인명살상자료를 바탕으로 일반화 멱함수를 적합시킨 결과 9/11과 같은 대규모 테러가 1968년에서 2007년 사이에 최소한 한번을 일어날 확률은 11–35%라고 보고하였다.

메르스와 같은 전염병의 확산 또는 탄저균 유출같은 경우를 대비한 감시시스템의 구축도 재난대비의 필수불가결한 요소 중의 하나이다. 이런 경우 hot spot을 탐지하는 것이 감시 시스템의 주요 임무 중의 하나이며 spatial scan statistic이 유용한 것으로 알려져 있다 (Kulldorff, 1997).

4.2. Sabermetrics

흔히 기록경기라고 일컬어지는 야구의 경우 기록지 한장으로 한경기의 내용을 요약할 수 있었다. Sabermetrics는 이러한 야구자료분석을 전문으로 하는 분야이다. 머니볼이라는 영화를 통해서 야구에서의 분석의 중요성은 널리 알려져 있으며 많은 메이저리그 구단에서 이러한 분석전문가들이 있다. 강정호 선수가 속해있는 Pittsburgh Pirates의 경우 기록분석을 이용한 수비 시프트가 20년만의 포스트시즌 진출에 큰 역할을 했다고 알려져 있다 (Sawchik, 2015).

펜실베니아 대학교의 통계학과 교수들이 개발한 수비능력평가 시스템인 Spatial Aggregate Fielding Evaluation(SAFE)의 경우 상대적으로 수비를 잘하는 것으로 알려진 뉴욕양키즈의 스타 유격수 데릭 지터가 이 시스템의 분석결과 실제로는 수비능력이 떨어진다고 밝혀 논란이 일기도 했다.

빅데이터의 시대는 sabermetrics에서도 새로운 변화를 불러왔다. 2014년 뉴스워크지는 빅데이터가 머니볼에 스테로이드를 주입한 것 같은 효과를 거둘 것이라며 예측하였다. 실제 메이저리그의 경기는 PITCH f/x, HIT f/x, FIELD f/x와 같은 시스템을 통해 선수들의 일거수일투족을 기록할 수 있으며 경기당 쏟아지는 데이터의 양은 GB단위의 양이다. R package pitchRx를 이용할 경우 PITCHf/x자료의 다양한 분석과 시각화가 가능하다 (Sievert, 2015).

4.3. 디지털 인문학

최근 구글은 "구글 도서관 프로젝트(google library project)"를 통하여 1450년 이후에 출판된 모든 책의 12%에 달하는 약 1500만권의 책을 디지털형태로 변환하였다. 구글은 N-gram viewer라는 온라인 서치엔진을 통해 "구글 도서관 프로젝트"에 있는 영어, 독일어, 스페인어 등 다양한 언어로 쓰여진 책들에서 N개의 연달아 등장하는 단어들의 빈도를 찾아볼 수 있도록 하였다.

구글은 또한 2011년부터 전세계의 박물관들의 전시품에 대해서도 "구글 아트 프로젝트(google art project)"라는 사업을 시도하여 작품들의 고화질 이미지 라이브러리를 구축하였다. 이렇게 디지털화된 인류의 문화유산에 관한 정량적 분석을 통하여 이때까지 알려지지 않는 여러 가지 흥미있는 특성을 찾아낼 수 있다. Kim 등 (2014)은 이러한 온라인 갤러리 중의 하나인 Web Gallery of Art (http://www.wga.hu)에 있는 서양화 8,798점의 디지털 이미지를 이용하여 서양미술의 정량적 분석을 통한 변천사를 살펴보았다.

한국의 대표적인 디지털 인문학 자료로는 조선왕조실록과 같은 역사적 문헌을 들 수 있다. 조선왕조실록은 유네스코 세계기록유산으로 1392년부터 1863년까지 25대 임금의 재위기간동안 일어났던 일에 대한 기록으로 총 6400만자에 달하는 분량이 저술되어 있다. 조선왕조실록의 경우 서양의 문헌에 비해 오랜기간 동안 동일한 방식으로 기록되었기 때문에 시대별 변화를 보다 효과적으로 측정할 수 있다는 장점이 있다. 또한 국사편찬위원회에서는 조선왕조실록의 디지털 자료를 무료로 온라인으로 제공하고 있어서 누구나 쉽게 접근하여 분석에 사용할 수 있는 장점이 있다. Bak과 Oh (2015)에서는 텍스트 마이닝을 이용하여 각 왕들의 통치스타일에 대한 분석를 시도했으며 Lee 등 (2016)에서는 조선왕조실록과 더불어 다음 4가지 역사기록 (승정원일기, 비변사등록, 고려사, 고려사절요)에서 한자들의 사용빈도의 분포를 분석하여 문체의 변화를 정량적으로 분석하였다. 이렇게 예술적 표현양식을 정량적으로 연구하는 분야를 양식측정학(stylometry)이라고 부르며 19세기 후반에 태동하여 문헌분석에 주로 초점을 두고 있다. 구글의 N-gram viewer과 같이 많은 경우 말뭉치(corpus)에서 주요단어들의 빈도수를 이용한 탐색적 분석에 초점을 맞추고 있으며 최근들어 텍스트 마이닝의 고급기법 중 하나인 토픽모형을 이용한 분석이 늘어나고 있다. 서양문헌에 비해서 조선왕조실록과 같은 우리나라의 역사적 문헌에 관한 연구는 손꼽을 정도로 아직은 더디게 진행되고 있으며 상대적으로 복잡계, 전산전공자들의 주도로 이루어지고 있다.

5. 결론

이 눈문에서 통계학적 관점에서 빅데이터 분석에 대하여 알아보았다. 빅데이터의 주요특징인 고차원과 대용량으로 야기되는 여러가지 문제점에 대해서 알아보고 이를 해결하기 위한 통계학 분야의 최근 동향과 다양한 응용분야에 대해서 살펴보았다. 이 분야에 대한 보다 포괄적인 리뷰는 Fan 등 (2014)과 Franke 등 (2016)을 참조하기 바란다.

빅데이터의 등장은 통계적 사고방식과 계산방법에 있어서 패러다임의 전환을 가져왔으며 기계학습과 인공지능의 발전에 있어서 통계학의 역할이 더욱 강조되고 있다. 데이터의 시대를 맞이하여 "데이터로 부터 배우는 학문"인 통계학이 보다 중추적인 역할을 할 수 있기를 기대한다.

References

Bak, J. Y. and Oh, A. (2015). Five centuries of monarchy in Korea: mining the text of the annals of the Joseon dynasty. In *Proceedings of the 9th SIGHUM workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities at the 53rd Annual Meeting of the Association for*

Computational Linguistics, 10–14.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data, *SIAM Review*, **51**, 661–703.

Clauset, A. and Woodard, R. (2013). Estimating the historical and future probabilities of large terrorist events, *Annals of Applied Statistics*, **7**, 1838–1865.

Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis, *National Science Review*, **1**, 293–314.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion), *Journal of the Royal Statistical Society Series B*, **70**, 849–911.

Franke, B., Plante, J.-F., Roscher, R., Lee, E.-S. A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., and Reid, N. (2016). Statistical inference, learning and models in big data. *International Statistical Review*, To appear.

Kim, D., Son, S.-W., and Jeong, H. (2014). Large-scale quantitative analysis of painting arts, *Scientific Reports*, **4**, 7370.

Ko, S. and Won, J.-H. (2016). Processing large-scale data with Apache Spark, *The Korean Journal of Applied Statistics*, To appear.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

Lee, B., Kim, D., Kim, D., and Jeong, H. (2016). N-gram web service and stylometric analysis of Korean historical documents, *New Physics: Sae Mulli*, **66**, 502–510.

Park, J.-H., Lee, S.-Y., Kang, D. H. and Won, J.-H. (2013). Hadoop and Mapreduce, *Journal of the Korean Data & Information Science Society*, **24**, 1013–1027.

Sawchik, T. (2015). *Big Data Baseball: Math, Miracles, and the End of a 20 Year Losing Streak*, Flatiron Books.

Sievert, C. (2015). Tools for harnessing 'MLBAM', 'Gameday' data and visualizing 'pictchfx', http://cpsievert.github.com/pitchRx

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

van de Geer, S. Bühlmann, P. Ritov, Y. A., and Dezeure, R. (2014). On asymptotically optimal confidence regions and test for high-dimensional models, *Annals of Statistics*, **42**, 1166–1202.

Witten, R. and Candès, E. (2013). Randomized algorithms for low-rank matrix factorizations: sharp performance bounds, *Algorithmica*, **63**, 355–363.

Zhang, C.-H. (2010.) Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, **38**, 894–942.

고차원 대용량 자료분석의 현재 동향

장원철[a,1] · 김광수[b] · 김정연[c]

[a]서울대학교 통계학과, [b]서울대학교 데이터과학및 지식창출 연구센터, [c]충북대학교 정보통계학과

(2016년 10월 19일 접수, 2016년 10월 19일 수정, 2016년 10월 19일 채택)

요 약

빅 데이터의 출현은 여러가지 과학적 난제에 대답 할 수 있는 기회를 제공하지만 흥미로운 도전을 또한 제공한다. 이러한 빅데이터의 주요 특징으로 "고차원"과 "대용량"을 들 수가 있다. 본 논문은 이러한 두 가지 특징에 동반되는 다음과 같은 도전문제에 대한 개요를 제시한다 : (1) 고차원 자료에서의 소음 축적과 위 상관 관계; (ii) 대용량 자료분석을 위한 계산 확장성. 또한 본 논문에서는 재난예측, 디지털 인문학과 세이버메트릭스 등 다양한 분야에서 빅 데이터의 다양한 응용사례를 제공한다.

주요용어: 빅데이터, 계산 확장성, 디지털 인문학, 소음 축적, 위 상관관계

본 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단 지원을 받아 수행된 기초연구사업임 (No. 2014R1A4A1007895).

[1]교신저자: (08826) 서울시 관악구 관악로 1, 서울대학교 통계학과. E-mail: wcjang@snu.ac.kr