The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.6.1107
29(6), 1107–1116

Analysis of massive data in astronomy

Min-Su Shin[a,1]

[a]Korea Astronomy and Space Science Institute

(Received September 19, 2016; Revised October 5, 2016; Accepted October 6, 2016)

Abstract

Recent astronomical survey observations have produced substantial amounts of data as well as completely changed conventional methods of analyzing astronomical data. Both classical statistical inference and modern machine learning methods have been used in every step of data analysis that range from data calibration to inferences of physical models. We are seeing the growing popularity of using machine learning methods in classical problems of astronomical data analysis due to low-cost data acquisition using cheap large-scale detectors and fast computer networks that enable us to share large volumes of data. It is common to consider the effects of inhomogeneous spatial and temporal coverage in the analysis of big astronomical data. The growing size of the data requires us to use parallel distributed computing environments as well as machine learning algorithms. Distributed data analysis systems have not been adopted widely for the general analysis of massive astronomical data. Gathering adequate training data is expensive in observation and learning data are generally collected from multiple data sources in astronomy; therefore, semi-supervised and ensemble machine learning methods will become important for the analysis of big astronomical data.

Keywords: astronomical data, statistical inference, machine learning, parallel computing, distributed computing

1. 서론

현대 관측 천문학의 최근 발전에 있어서 중요한 요소로, 검출 기기의 기술적 진보와 이를 통해서 대형화된 관측기기의 보편적 이용을 꼽을 수 있다. 관측 천문학이 이용하는 전자기파의 전파부터 감마선에 이르는 전 파장 영역에 있어서, 전자공학과 재료과학의 발전을 통해서 더 어두운 천체를 다양한 파장에서 광범위한 공간과 시간 영역에 있어서 관측할 수 있는 새로운 관측기기들이 출현하게 되었다. 특별히, 가시광선 파장 대역에서 Sloan Digital Sky Survey(SDSS) 탐사관측 (Gunn 등, 2006)의 경우 지금까지 단일 가시광선 프로젝트로서는 가장 방대한 양의 관측 자료를 생산했는데, 분석 전 원래 자료의 크기만 고려했을 때 그 자료의 양은 약 10^6개 규모의 스펙트럼과 10^4deg^2 규모의 하늘을 관측한 사진에 이른다 (Abazajian 등, 2009). 이러한 방대한 양의 자료획득이 가능했던 것은 6개의 필터를 이용하여 광대한 하늘을 동시에 관측할 수 있는 대형 모자익 카메라를 이용할 수 있었기 때문이다.

대용량 관측 자료의 생산에 더불어서, 또 하나의 중요한 관측 천문학 연구의 변화 요소는 손쉬운 대용량 관측 자료의 공유를 가능케 한 자료 공개, 배포, 접근 수단의 변화이다. 1990년대에 들어서 급격하게

[1]Korea Astronomy and Space Science Institute, 776, Daedeokdae-ro, Yuseong-gu, Daejeon 34055, Korea. E-mail: msshin@kasi.re.kr

천문학 연구의 중요한 도구로 쓰이게 된 인터넷 환경은, 2000년대에 들어서 빠른 속도로 전 세계적으로 얻어지는 값진 관측 자료들의 공유를 촉진시키는 기술의 발전으로 연결되었다. 앞서 언급한 SDSS의 경우, 관측 천문학 연구에서 처음으로 전 세계 천문학자들에게 전체 자료의 접근을 허용하고, 손쉽게 기본적인 자료 분석과 획득을 가능케 하는 환경을 제시하였다 (Szalay 등, 2000). 이러한 환경의 제공은, 원래 프로젝트에 참여하지 않은 연구자들도 대용량 관측 자료의 분석을 통한 연구를 수행하는 것을 가능케 하는 획기적인 방향의 전환이었다. 이러한 변화의 결과로, 2000년대에 가상 천문대라는 개념의 국제적인 관측 자료 공유/배포 환경의 구축이 이루어진다 (Golombek, 2004).

이러한 대용량 관측 자료의 획득과 공유를 통한 자료 분석이라는 연구 환경의 변화는, 천문학자들이 자연스럽게 대용량 자료 분석에 있어서 몇 가지 현실적인 문제들을 경험하게 하였다. 첫째, 대용량 자료를 분석할 수 있는 컴퓨터 알고리즘과 분석 환경의 중요성이 새롭게 인식되었는데, 이는 컴퓨터 과학자들과의 협업을 통해서 병렬 분산 컴퓨팅 기술이 대용량 천문학 자료 분석에 필수적으로 활용되는 변화로 이어졌다. 둘째, 기존의 다양한 분석 방법을 큰 규모의 자료 분석에 적용하기 위해서, 이들 분석 방법에 대한 새로운 구현의 필요성이 대두되었다. 이는 병렬 분산 처리라는 분석 환경에 맞추어 자료 분석을 구현하는 것이다. 이 부분에 있어서 기존에 'Astroinformatics'라는 분야로 이루어지던 통계학자 및 전산학자들과의 공동 연구가 가지는 중요성이 새롭게 인식되고 있다 (Zhang과 Zhao, 2015). 셋째, 자동화된 대용량 자료 분석과 그 과정에서 잘못된 검출과 판단의 규모를 정량적으로 추정하는 것의 중요성이 부각되었다. 이 목적을 위하여 기계학습 방법이 다양한 목적을 가지고 활용되기 시작하였는데, 아직까지 천문학에서의 활용은 본격적이지는 않은 상황으로 기계학습 연관 분야와의 공동 연구가 요구되고 있다.

이 논문에서는, 위에 기술한 최근의 대용량 천문 자료 분석에서 관심이 되고 있는 주제를 중심으로, 최근의 자료 분석의 핵심 문제들과 분석 사례를 소개하고자 한다. 이를 통해서 이 논문이 국내 통계학자들이 대용량 천문학 자료를 분석하는 문제에 참여하는 기회를 모색하는데 도움이 되었으면 한다. 이를 위해서 먼저 천문학 관측 자료의 특징에 대해서 다음 2절에서 간략히 기술하고, 3절에서 핵심 문제들에 대해서 최근 논의의 내용을 제시한다. 마지막으로 4절에서 요약과 결론을 제시하며, 앞으로 응용통계학 연구에서 천문학 대용량 자료 분석을 통한 협력 연구의 가능성을 제시하고자 한다.

이 논문에 보완하여, 대용량 천문학 자료 분석에 있어서의 다양한 기계학습 활용에 대한 포괄적인 소개는 'Advances in Machine Learning and Data Mining for Astronomy' (Way 등, 2012)를 참고하고, 여러 통계 분석 방법의 적용 사례에 대해서는 이번 해에 여섯 번째를 맞이했던 'Statistical Challenges in Modern Astronomy VI' 학회 발표 자료들과 그 이전 학회들의 출판물 (Feigelson과 Babu, 2012)들을 참고하기 바란다. 나아가서 이러한 대용량 자료의 분석과 관련하여 실질적인 분석 환경 구현과 관련된 사례들은, 1991년부터 연례적으로 열리는 'Astronomical Data Analysis Software & Systems' 학회 발표 내용들을 참고하면 도움이 될 것이다.

2. 천문학 대용량 관측 자료의 특징

천문학은 실험이 아닌 관측을 중심으로 자료를 획득하는 분야로서, 대부분의 자료가 우주를 관측하는 망원경을 통하여 획득된다. 망원경을 이용해 획득되는 관측 자료는, 관측 환경에 영향을 주는 자연적 요소와, 관측 기기의 상태라는 두 가지 요소에 의해서 영향을 받는다. 전자의 경우 관측자가 그 상태를 조정 및 관리할 수 없기에, 환경 요소 그 자체를 정확히 측정하는 과정을 통하여 파악이 된다. 반면에 후자는, 실제 관측자가 관리 및 측정이 용이한 기기 특성으로서, 일반적으로 관측을 통하여 자료 획득이 얻어지기 전이나 관측 과정에서 파악이 용이하다. 이들 두 요소들은 결국 측정이 되어야 하는 것들로, 확률 분포를 가지는 변수로 고려되어야 한다.

광학 망원경을 이용해서 얻어지는 관측 자료를 생각해 볼 때, 관측 환경의 자연적 요소로 대표적인 예는 대기 영향으로 나타나는 관측 이미지에서 천체들의 형상 변화가 있다. 관측 기기의 상태로 나타나는 요소의 예로는 기기가 만들어내는 이미지의 위치 분포가 있겠다. 이러한 정보들은 과학 자료를 획득하는 기기에서 얻어지는 자료로부터 추정을 하거나, 독립적인 관측 기기를 사용하여 과학 관측이 이루어지는 환경 정보를 동시에 측정하는 방법들이 이용되고 있다.

관측 기기의 상태 정보는, 기기의 측정치를 물리적인 양으로 전환하는데 영향을 줄 수 있는 가능한 모든 요소들을 파악하는 방법으로 획득된다. 예를 들어, 밤하늘을 관측한 이미지 자료로부터 측정되는 밝기 값을 물리적인 의미가 있는 값으로 전환을 하는 필수 과정이 있다. 이 과정에서 기기가 실제 밝기에 반응하는 반응성이 파악되어야 하고, 또한 기기가 빛에 노출되지 않은 경우에도 가지게 되는 신호의 세기 등이 확인이 되어야 한다.

이러한 자연 요소와 관측 기기의 요소들이 물리량의 측정에 어떻게 영향을 주는지 정량적으로 결정하는 과정에, 다수의 통계적인 추론 과정이 결부 된다. 첫째, 측정치에 대한 오차를 추정하기 위해 관측되는 신호가 가지는 통계 분포가 일반적으로 가정되어야 하는데, 흔히 Poisson 분포나 Gaussian 분포가 가정된다. 둘째, 측정치와 물리량 사이의 전환은 측정치가 가지는 오차를 고려하여 회귀분석의 방법으로 흔히 얻어지게 된다. 예를 들어, 이미지에서 얻어지는 천체들의 위치를 실제 물리적인 좌표계로 전환하는데 있어서, 2차원의 두 좌표들 사이의 회귀분석을 통해서 가장 적절한 전환 식을 추정하게 된다 (Pier 등, 2003). 천체가 내는 빛의 에너지 분포인 스펙트럼을 얻는 경우에도, 관측으로 얻어지는 이미지 자료에서의 위치 정보를 스펙트럼의 빛의 파장과 연결 짓는 과정에서 회귀분석이 필수적으로 이용된다.

따라서 이러한 관측 측정값의 물리량으로의 전환에 관여하는 여러 과정들에서 얻어지는 통계적 추론들의 정확성과 신뢰성에 대한 자동화된 접근이, 대용량 천문 자료 분석에 필수적으로 요구된다. 기존의 소규모 관측 자료 분석을 통한 연구에서는 이러한 자료의 검증 과정이 개별 연구자의 판단에 의해서 수동적으로 이루어졌다. 그러나 근래의 대용량 관측 자료 증가에 맞추어 이러한 기존의 방식은 한계를 가지게 되었고, 통계적 추론의 정확성에 대한 정보를 이용하여 대용량의 자료들을 자동으로 평가하고 활용하는 것이 필수인 상황이다. 또한 관측의 자연 요소나 기기들의 상태가 안정적이지 않고 다양하게 변동될 수 있기에, 이러한 변화하는 요소들을 자동으로 인지하고, 자료 검증에 기계학습을 적용하고자 하는 노력이 논의되고 있다.

최근의 대용량 천문학 자료 증가는, 크게 공간, 시간, 파장의 세 가지 방향에서 이루어지고 있다. 먼저 더 많은 하늘의 영역을 관측하거나, 같은 영역의 하늘이라도 더 자세하게 관측하는 경우이다. 또는 더 어두운 천체를 관측할 수 있어서 더 먼 천체들을 관측하면서, 더 큰 공간을 관측할 수 있게 되는 것도 관측 자료 증가로 이어지고 있다. 이와 더불어서, 천체들의 시간에 따른 변화를 추적하기 위해서, 동일한 영역을 수차례에 걸쳐서 관측하는 것 역시 중요한 탐사 관측의 일종으로서, 최근의 대용량 자료 생산에 주된 과정이 되고 있다. 나아가서 빛의 다양한 파장 대역에서 동일한 천체를 관측하는 것 역시 최근의 대용량 자료 생산의 한 방법이다. 따라서 천문학 대용량 자료의 분석에 있어서, 광범위한 하늘의 영역에서 발견되는 다양한 천체에 대해서 수차례에 걸쳐서 다양한 파장 대역으로 관측되는 것을 어떻게 통합적으로 분석할 수 있는지가 중요한 문제이다.

다른 종류의 관측 자료와 달리, 천문학의 관측 자료는 기상이나 천체의 위치와 같은 관측 제한 조건들로 인한 자료의 불 균질성의 문제를 가지고 있다. 시간 자료의 경우 규칙적으로 장기간 얻어지기 어렵다는 특징을 가지고 있다. 따라서 일반적으로 규칙적으로 측정되는 자료의 분석에 이용되는 방법들은 그 적용이 불가능하며, 경우에 따라서는 적용하더라도 그 결과의 해석이 명확하지 않은 문제를 가진다 (Vio 등, 2013). 또한 시간에 따른 자료의 질 역시 변화를 보인다는 문제도 존재한다. 이 경우 단순한 정규분포를 가정하고 정규분포 검증을 수행하는 것과 같은 기존의 분석들은 한계를 가지게 된다. 공간의 경우

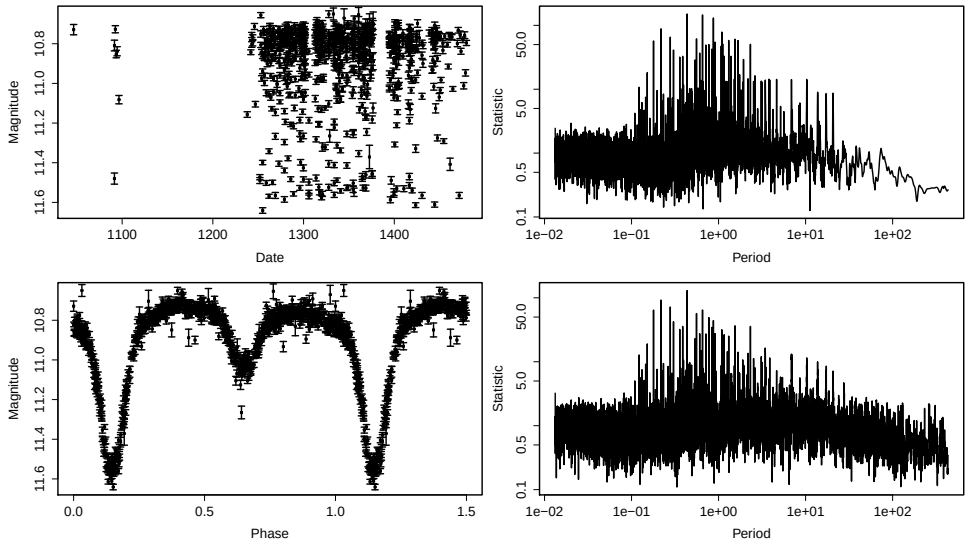

Figure 3.1. Example of time-series data (top left) and periodograms estimated by two different methods for the example time-series data (right; Shin and Byun, 2004). The phase-folded time-series data at a period of 0.435801 days (bottom left) imply that the variability in the time-series data is real rather than the result of noise.

에서도 비슷한 문제가 발생하는데, 전체 하늘에 대해서 관측 자료의 질이 가지는 불 균질성과 같은 문제를 고려하여야 한다 (Szapudi 등, 2005). 다 파장 관측 자료의 경우에도, 특정 파장의 경우에 해당 천체가 검출되지 않거나 신호의 강도가 다르게 얻어지는 등의 불 균질성 문제가 발생할 수 있게 된다.

3. 대용량 천문학 자료 분석의 문제들

3.1. 병렬 분산 환경에서의 자료 분석

최근의 대용량 자료 분석을 중심으로 한 연구에서, 분석 방법을 병렬 분산 환경에서 구현하는 것이 중요한 문제로 등장하고 있다. 대용량 자료를 분석하는 경우에, 대부분의 기존 분석 방법은 전체 자료를 분석하는데 너무 많은 시간이 소요되는 문제를 가지게 된다. 이를 위해서 병렬 분석이나 분산 분석이 필수가 되는데, 이를 위해서는 기존의 알고리즘을 이에 맞추어서 새롭게 구현하거나, 혹은 새로운 알고리즘을 적용할 필요성이 대두된다.

아직 천문학의 대용량 자료 분석에 있어서, 병렬 혹은 분산 분석 알고리즘의 적용은 굉장히 제한적으로 특정 분석 방법에만 적용되어 왔다. 대표적인 예가, 공간적으로 두 지점 사이의 밀도 분포가 가지는 상관성을 측정하는데 천문학에서 흔히 이용되는 아래 수식 (3.1)의 spatial two-point correlation function 분석이 이에 해당된다. 이에 대해서도 Graphics Processing Unit(GPU) 등과 같은 새로운 병렬 분산 처리 방식이 필수가 되고 있다 (Alonso, 2012). 다수의 천체들이 공간에 분포할 때, 두 천체사이의 변위가 $\vec{r}$이고 각각이 차지하는 부피가 dV_1과 dV_2라면, 평균 밀도 $\bar{n}$에 비해서 두 천체가 발견될 초과 확률은

$$\langle dP \rangle = \bar{n}[1 + \xi(\vec{r})]\, dV_1\, dV_2 \tag{3.1}$$

으로 규정되고, 여기서 ξ가 two-point correlation으로 측정된다. 또한 천체의 시간에 따른 변화를 측정

한 시계열 자료에서 주기적인 변화의 강도를 추정하는 periodogram 분석 (Figure 3.1) 역시, 관측되는 천체들의 수가 빠르게 증가하는 대용량 관측 시대에는 GPU를 이용한 병렬 분석 (Townsend, 2010) 혹은 Open Multi-Processing(OpenMP)을 활용한 병렬 분석 (Shin과 Byun, 2004)과 같은 노력이 요구되고 있다.

현재 대용량 자료의 분석에 흔하게 이용되는 일반적인 도구들의 천문학 대용량 자료 분석에서의 활용은 더딘 상황이다. 대표적인 공개 소프트웨어 도구로서 대용량 자료에 대한 일반적인 분석을 위해서 최근에 주목을 받아 개발되고 있는 R (Ihaka와 Gentleman, 1996)의 경우에도, 통계학 사용자들에게는 흔하게 이용되지만 천문학자들에게는 활발하게 이용되고 있지 않은 상황이다. 산업계에서도 활발히 활용되고 있는 병렬 분산 자료 분석 환경인 Apache Mahout 등과 같은 도구의 적용 및 활용도 굉장히 제한적이다 (Hahm 등, 2012). 최근에 천문학자들이 광범위하게 이용하고자 하는 도구는 Python 프로그래밍 언어에 기반을 둔 병렬 분산 분석 도구들로서 (Ivezić 등, 2014), 범용성을 중심으로 두기보다는 아직까지는 특정 대용량 자료 분석 목적을 기반으로 개발 및 활용되고 있다 (Singh 등, 2013).

근래에 활발하게 범용 병렬 분산 자료 분석을 위한 도구로 주목 받고 있는 부분은 Markov Chain Monte Carlo(MCMC) sampler이다. MCMC 방법은 과거 10여 년간 베이지안 통계 추론이나 기계학습 방법이 천문학 자료 분석에서 광범위하게 적용되면서 중요한 도구로 발전하였다 (Allison과 Dunkley, 2014; Zuntz 등, 2015). 대용량의 자료를 분석하는데 있어서, 이러한 MCMC sampler의 속도를 높이기 위하여 병렬 분산 계산 환경에서 효율적으로 이용할 수 있게 하는 것은 중요한 과제이다. 이를 위해서 알고리즘의 개선이나 새로운 계산 환경에서의 MCMC sampler가 구현된 것들이, 대용량 천문학 자료 분석에 꾸준히 이용되고 있다. 광범위하게 천문학자들에 의해서 이용되는 MCMC sampler의 구현된 예로는 MULTINEST (Feroz 등, 2009), PyMC (Patil 등, 2010), emcee (Foreman-Mackey 등, 2013)와 같은 소프트웨어들이 있다.

3.2. 기계학습 방법의 활용

대용량 천문학 자료 분석에서도 다른 분야의 대용량 자료 분석의 경우와 마찬가지로, 최근에 기계학습을 활용한 대용량 자료 분석을 빠르게 적용하고 있는 상황이다 (Ball과 Brunner, 2010; Borne, 2013). 특별히 다 변수 다차원 자료 분석은 기존의 단순 통계 분석 방법에서 나아가서, 분류(classification)나 군집분석(clustering)과 같은 기계학습 방법의 활용을 촉진하고 있다 (Bhat, 2011; Al-Jarrah 등, 2015).

천문학 대용량 자료의 분석에서 기계학습을 활용하는 첫 단계는, 자료를 학습 가능하게 기술하는 자료 특성 추출(feature extraction)이다 (Zheng과 Zhang, 2008). 즉, 원 자료에서 학습이 필요한 의미 있는 정보들을 추출하는 과정이 필요하다. 이는 다차원의 자료에 담긴 정보를 낮은 차원으로 축약하는 것에 해당한다. 많은 경우 이들 방법은 기존의 다양한 통계치를 측정하는 것에 근거하기도 한다. 예를 들어, Figure 3.1에 제시된 시간에 따른 천체의 밝기 변화와 같은 관측 자료의 경우, 통계치를 이용한 통계적 추론과 상관없이, 아래의

$$\eta = \frac{\sum_{n=1}^{N-1}(x_{n+1} - x_n)^2/(N-1)}{s^2} \tag{3.2}$$

와 같은 이미 알려진 통계치(von Neumann, 1941)를 기계학습에 활용할 수 있는 자료 특성으로 이용하기도 한다. 식 (3.2)에서 x_n은 전체 N개의 관측 중에서 n번째 관측된 밝기를 나타내고, s는 관측된 밝기 값들의 표준편차 값을 나타낸다. 이 경우에, 이 통계치를 이용한 가설 검증과 같은 통계적 추론을 이용하지는 않지만, 실제 밝기 변화가 있다면 이 통계치가 어떻게 결정되어야 할지에 대한 이해가 있어서, 기계학습의 특성 추출의 한 가지로 이용되는 것이다. 반면에 이미 알려진 통계치를 이용하기 보다는, 다

음과 같이 임의적으로 규정한 자료 특성을 이용하기도 한다 (Stetson, 1996).

$$J = \sum_{n=1}^{N-1} \mathrm{sign}(\delta_n \delta_{n+1}) \sqrt{|\delta_n \delta_{n+1}|}, \tag{3.3}$$

여기서 δ_n은 전체 N개의 관측 자료 중에서 n번째 관측치 x_n에 대해서, 푸아송 분포와 같이 가정하는 분포에 따라서 추정되는 그 측정 오차 e_n와 전체 평균 μ를 이용해서, $\sqrt{N/(N-1)}(x_n - \mu)/e_n$로 정의된다. 이들 추출된 자료 특성들은 다양한 기계학습 방법을 활용하는데 입력 자료로 이용된다.

기계학습에 이용되는 자료 특성 추출은, 기계학습 방법 활용에 있어서 자료와 기계학습으로 풀고자 하는 문제를 가장 잘 아는 천문학자들이 가장 중요한 역할을 하는 부분에 해당한다. 즉, 해당 연구 분야에 따라서 특성(feature)을 선정하는 것 역시 그 가능성이 다양하다 (Saeys 등, 2007; Hira와 Gillies, 2015). 현재 이용되는 여러 통계치들의 활용이나 자유롭게 정의되는 특성 측정치들이 가지는 한계와 유용성에 대한 연구가 계속되고 있는 상황이다. 특별히 중요한 문제로는 이들 특성 추출 과정에서 어떻게 관측 자료의 측정 오차를 특성에 반영할 수 있는 지이다. 광범위하게 이용되는 기계학습법들은 일반적으로 입력 자료 개개의 특성이 가지는 오차를 고려하지 않는다. 이러한 자료의 오차가 가지는 효과는 전체 입력 자료의 분포로만 대부분의 기계학습 자료 활용에서 고려되고 있다. 또 다른 접근 방법으로, 식 (3.3)과 같이 입력 자료의 특성을 계산할 때 관측 자료의 측정 오차를 반영하는 것을 고려할 수 있다. 그러나 이러한 접근법을 유용하게 만들기 위해서는, 자료 특성으로 이용되는 통계치나 다른 측정치를 정의할 때, 어떻게 엄밀하게 자료 측정의 오차를 반영할 수 있는가에 대한 연구가 필요하다.

입력 자료의 질을 직접 고려하는 기계학습 방법의 활용 역시, 다양한 질의 대용량 자료를 분석하는 최근 상황에서 주목할 필요가 있다. 예를 들어, 유사한 자료의 묶음을 찾는 과정(clustering)에서 입력 자료에 각 자료의 정확도에 따라서 다른 가중치를 할당하는 것을 생각해 볼 수 있다 (Gebru 등, 2015). 가중치의 사용은 단지 측정치의 정확도를 나타내는 것 외에도, 입력 자료의 불 균질성 등을 보정하는 용도 등으로도 이용될 수 있다. 따라서 관측 자료의 부정확성을, 실제 기계학습 입력 자료의 가중치로 연결하는 방법에 대한 깊이 있는 논의가 필요하다. 대용량 자료 분류(classification)의 경우에도 가중치를 이용하는 방법을 생각해 볼 수 있으나, 이를 입력 자료의 부정확성과 연결 지어 제시된 방법은 드물다.

천문학 자료의 기계학습을 이용한 분석에서 흔히 겪는 문제 중의 하나는 학습용 자료(training sample)의 규모가 적용에 쓰이는 자료의 규모보다 일반적으로 훨씬 작다는 사실이다. 관측을 통해서 자료를 획득하는 천문학에서, 다량의 다양한 경우를 고려해서 충분한 기계학습을 위한 자료를 획득하는 것은 쉽지 않다. 특히 다량의 학습용 자료를 준비하기 위해서, 출처가 다양한 자료들을 모아서 불 균질한 학습 자료를 이용하는 경우도 있게 된다. 학습 입력 자료 구성에 따라서 기계학습법 적용의 성패가 다르기 때문에, 불 균질적인 학습 자료나 소수의 학습 자료에서 학습된 결과를 압도적으로 더 큰 적용 자료에 활용하는 방법에 대한 추가적인 연구가 요구되는 상황이다. 이러한 점에서 다양한 출처의 학습 자료로 부터 얻어지는 개개의 학습 결과를 결합해서 적용하는 앙상블 학습법(ensemble learning; Zhou, 2015)의 활용 가능성이 크다. 또한, 준 지도학습(semi-supervised learning; Chapelle 등, 2010)과 같이 그 정답을 아는 학습 자료의 양이 부족한 경우를 대상으로 하는 방법들의 활용 가능성이 크다.

기계학습의 활용과 관련되어 천문학 대용량 자료 분석에서 또 하나 중요한 문제는, 예측이나 분류 문제에 대해서 엄밀하고 신뢰할 수 있는 예측 및 분류 정확도의 제시이다. 대용량 자료를 기계학습을 이용해서 자동으로 빠르게 분석하는 것이 필수적인 상황에서, 그 분석의 정확도를 측정하는 것 (Borra와 Di Ciaccio, 2010)은 천문학 연구에서 여러 문제에 관련된다. 대부분의 경우 기계학습을 통해서 얻게 되는 추가적인 측정치는 다른 분석에 이어져서 이용되는데, 그 정확도에 대한 엄밀한 측정이 없이는 다음 분석 과정이 정확히 진행되기 어렵게 된다. 대표적인 예로, 은하들의 색상에 근거해서 그 은하들의 종류

와 우리로부터의 거리(정확히는 적색편이)를 추정하는 문제의 경우, 기계학습 방법 중에서도 특별히 지도학습(supervised learning)의 문제인 회귀분석의 일종으로 꾸준히 활용되어 왔다 (Cavuoti 등, 2015). 이 문제는 제한된 수의 은하들에 대해서 그 종류와 적색편이가 알려져 있고, 이들 학습 자료를 이용해서 다수의 종류와 거리가 알려지지 않은 은하들에 대해서 그 종류와 적색편이를 추정하는 것을 목적으로 한다. 이렇게 추정된 적색 편이 정보는 우주의 은하 분포가 가지는 거대 구조를 연구하거나, 거리에 따른 은하 진화를 연구하는 후속 연구에 중요한 기본 자료로 이용된다. 따라서 추정된 적색편이의 정확성에 대한 정보가 후속 분석 과정에 필요하게 된다.

예측의 부정확성을 엄밀하고 신뢰할 수 있는 정보로 제시하는 방법들에 대한 요구는 계속되어 강조될 것으로 예상된다. 현재까지 천문학 자료의 기계학습 분석에 있어서 가장 보편적으로 이용되는 방법으로는 교차타당성(cross-validation)이 있다. 앞서 언급한 은하의 적색편이 추정에 있어서도, 교차타당성 등과 같은 방법으로 적색편이의 범위를 추정하기도 한다. 주의할 점은 현재 모든 기계학습 방법들이 적색편이의 분포인 $P(z)$를 제시하는 것은 아니라는 사실이다.

4. 요약 및 결론

천문학 대용량 관측 자료의 증가와 더불어서, 다양한 관측 기기 및 환경적인 요소에 의해서 발생하는 문제들과, 불완전한 시간 및 공간 분포를 가지는 자료 획득 문제를 천문학 자료의 특징으로 제시하였다. 이러한 문제들을 고려한 자료 분석 방법들에 대한 필요성은, Large Synoptic Survey Telescope(LSST) (Ivezic 등, 2008)와 같은 미래 대용량 탐사 관측 자료 분석을 위해서도 당연히 강조되고 있다. LSST 경우 10^{13}회 규모 정도의 위치, 시간에 따른 천체의 특성을 측정할 예정으로, 천문학에서 가장 큰 규모의 관측 자료를 생산할 것으로 예상된다. 이 경우에도 다양한 관측에서의 시간 및 공간 제약으로 야기되는 문제들을 고려한 분석이 요구된다 (Liao 등, 2015).

최근의 대용량 천문 자료 분석은 필연적으로 병렬 분산 환경을 이용할 필요가 있고, 더불어서 자료 분석의 자동화를 가능케 하는 기계학습의 활용이 필수적인 과정이 되고 있다. 앞서 제시된 것처럼, 광범위한 경우에 활용할 수 있는 병렬 자료 분산 환경의 개발 및 사용은 더딘 상황이다. 그러나 베이지안 추론이 적용되는 사례의 확장 및 대용량 자료에 대한 모델 적용에서 요구되는 MCMC sampler의 광범위한 활용이 최근 연구에서 활발하다. 더불어서 다양한 기계학습 방법의 활용도 주목을 받고 있다. 기계학습 방법의 효과적인 활용을 위해서는, 무엇보다도 대용량 자료의 특성을 충분히 기술할 수 있는, 특성 추출 단계에서의 세심한 고려가 필수이다. 이 부분에 있어서 다양한 통계치에 대한 이해가 중요한 역할을 할 것으로 보인다.

다른 분야에서의 기계학습과는 달리, 관측을 통한 대용량의 학습 자료 획득이 쉽지 않고, 그래서 다양한 출처의 자료를 통합해서 기계학습을 적용해야하는 문제에 대한 인식이, 천문학 대용량 자료 분석에서 필요하다. 나아가서 관측으로 측정되는 물리량의 측정 오차를 반영한 기계학습법 활용이라는 본질적인 문제에 대해서 방법론적인 연구가 필요할 것이다.

앞서 제시한 문제들과 관련하여, 우리나라의 통계학 연구자들과 천문학자들이 관심을 가지고 협력 연구를 진행할 수 있는 근 미래의 대용량 자료로는, 무엇보다도 LSST를 처음으로 고려할 수 있다. LSST를 통해서 국내 연구자들이 얻게 될 자료는, 기본적으로 다양한 빛의 파장 대에서 획득되는 공간-시간 자료(spatio-temporal data)이다. 이를 통해서, 시계열 자료(time-series data) 분석이 더욱더 중요해 질 것으로 예상된다. 구체적으로 시계열 자료들에서 특정 유형을 검출하고, 이미 알려져 있는 시계열 변화 형태에 근거해서, 얻어지는 자료를 분류하는 것과 그 분류 정확도를 측정하는 문제가 대두될 것이다. 아주 먼 우주의 천체에서 얻어지는 밝기가 변하는 시계열 자료의 경우, 우주 팽창의 속도를 측정하는데 쓰

일 수 있는데, 이 경우에는 불규칙하게 얻어진 자료를 이용해서 모델을 검증하는 것을 필요로 한다. 또한 움직이는 천체로 검출이 될 태양계 소행성 천체를 발견하고, 그 궤도를 모델과 비교하여 정확히 추정하고 그 정확도 역시 추정할 필요가 있게 된다. 나아가서 시간에 따라 밝기가 변하는 시계열 자료로부터 이러한 소행성의 모양을 추정하는 문제를 풀어야 할 것이다.

LSST 자료 분석에 대한 준비는 자료 분석 방법에만 제한되는 것이 아니라, 대다수의 국내 천문학자들이 이러한 대용량 자료 분석을 수행할 수 있는 환경의 구성도 포함한다. 병렬 분산 환경에서 자료 분석이 가능한 데이터베이스와 도구가 마련될 필요가 있다. 특별히 R과 같은 도구를 천문학자들이 거대 자료 분석에 이용하는 것에 있어서, 국내 통계학자들이 협력할 수 있는 부분이 있을 것으로 기대된다.

감사의 글

이 논문의 기반이 되는 여러 주제에 대해서, 대용량 천문학 자료 분석과 관련하여 다양한 의견을 주신, 김대원, 변용익, 이한, 장서원 등의 분들께 감사드립니다.

References

Abazajian, K. N. Abazajian, K. N., Adelman-McCarthy, J. K., Ağueros, M. A., Allam, S. S., Prieto, C. A., An, D., *et al.* (2009). The seventh data release of the Sloan Digital Sky survey, *The Astrophysical Journal Supplement*, **182**, 543–558.

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., and Taha, K. (2015). Efficient machine learning for big data: a review, *Big Data Research*, **2**, 87–93.

Allison, R. and Dunkley, J. (2014). Comparison of sampling techniques for Bayesian parameter estimation, *Monthly Notices of the Royal Astronomical Society*, **437**, 3918–3928.

Alonso, D. (2012). CUTE solutions for two-point correlation functions from large cosmological datasets, *ArXiv e-prints*, 1210.1833. Available from: https://arxiv.org/abs/1210.1833

Ball, N. M. and Brunner, R. J. (2010). Data mining and machine learning in astronomy, *International Journal of Modern Physics D*, **19**, 1049–1106.

Bhat, P. C. (2011). Multivariate analysis methods in particle physics, *Annual Review of Nuclear and Particle Science*, **61**, 281–309.

Borne, K. (2013). Virtual observatories, data mining, and astroinformatics. In *Planets, Stars and Stellar Systems* (pp. 403–443), Springer Netherlands

Borra, S. and Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods, *Computational Statistics & Data Analysis*, **54**, 2976–2989.

Cavuoti, S., Brescia, M., De Stefano, V., and Longo, G. (2015). Photometric redshift estimation based on data mining with PhotoRApToR, *Experimental Astronomy*, **39**, 45–71.

Chapelle, O., Schlkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*, The MIT Press.

Feigelson, E. D. and Babu, J. (2012). *Statistical Challenges in Modern Astronomy V*, (Volume 902 of Lecture Notes in Statistics), Springer, New York.

Feroz, F., Hobson, M. P., and Bridges, M. (2009). MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics, *Monthly Notices of the Royal Astronomical Society*, **398**, 1601–1614.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). emcee: The MCMC Hammer, *Publications of the Astronomical Society of Pacific*, **125**, 306–312.

Gebru, I. D., Alameda-Pineda, X., Forbes, F., and Horaud, R. (2015). EM algorithms for weighted-data clustering with application to audio-visual scene analysis, *CoRR*, Available from: https://arxiv.org/abs/1509.01509

Golombek, D. (2004). Archives, databases and the emerging virtual observatories, *Astrophysics and Space Science*, **290**, 449–456.

Gunn, J. E., Siegmund, W. A., Mannery, E. J., Owen, R. E., Hull, C. L., Leger, R. F., *et al.* (2006). The 2.5 m telescope of the sloan digital sky survey, *The Astronomical Journal*, **131**, 2332–2359.

Hahm, J., Kwon, O.-K., Kim, S., Jung, Y.-H., Yoon, J.-W., Kim, J., Kim, M.-K., Byun, Y.-I., Shin, M.-S., and Park, C. (2012). Astronomical time series data analysis leveraging science cloud, In *Lecture Notes in Electrical Engineering*, **181**, 493–500.

Hira, Z. M. and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data, *Advances in Bioinformatics*, **2015**, Article ID 198363.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Ivezic, Z., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., AlSayyad, Y., *et al.* (2008). LSST: from science drivers to reference design and anticipated data products, *ArXiv e-prints*, 0805.2366, Available from: https://arxiv.org/abs/0805.2366

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., and Gray, A. (2014). *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*, Princeton University Press.

Liao, K., Treu, T., Marshall, P., Fassnacht, C. D., Rumbaugh, N., Dobler, G., *et al.* (2015). Strong lens time delay challenge. II. Results of TDC1, *The Astrophysical Journal*, **800**, 11.

Patil, A., Huard, D., and Fonnesbeck, C. (2010). PyMC: Bayesian stochastic modelling in python, *Journal of Statistical Software*, **35**, 4.

Pier, J. R., Munn, J. A., Hindsley, R. B., Hennessy, G. S., Kent, S. M., Lupton, R. H., *et al.* (2003). Astrometric calibration of the sloan digital sky survey, *The Astronomical Journal*, **125**, 1559–1579.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23**, 2507–2517.

Shin, M.-S. and Byun, Y.-I. (2004). Efficient period search for time series photometry, *Journal of Korean Astronomical Society*, **37**, 79–85.

Singh, N., Browne, L.-M,. and Butler, R. (2013). Parallel astronomical data processing with Python: Recipes for multicore machines, *Astronomy and Computing*, **2**, 1–10.

Stetson, P. B. (1996). On the automatic determination of light-curve parameters for Cepheid variables, *Publications of the Astronomical Society of the Pacific*, **108**, 851–876.

Szalay, A. S., Kunszt, P. Z., Thakar, A. R., Gray, J., and Slutz, D. (2000). The sloan digital sky survey and its archive, *Astronomical Data Analysis Software and Systems IX, ASP Conference Proceedings*, **216**, 405–414.

Szapudi, I., Pan, J., Prunet, S., and Budavári, T. (2005). Fast edge-corrected measurement of the two-point correlation function and the power spectrum, *The Astrophysical Journal*, **631**, L1–L4.

Townsend, R. H. D. (2010). Fast calculation of the Lomb-Scargle periodogram using graphics processing units, *The Astrophysical Journal Supplement*, **191**, 247–253.

Vio, R., Diaz-Trigo, M., and Andreani, P. (2013). Irregular time series in astronomy and the use of the Lomb-Scargle periodogram, *Astronomy and Computing*, **1**, 5–16.

Way, M. J., Scargle, J. D., Ali, K. M., and Srivastava, A. N. (2012). *Advances in Machine Learning and Data Mining for Astronomy* (1st ed.), Chapman & Hall/CRC.

Zhang, Y. and Zhao, Y. (2015). Astronomy in the big data era, *Data Science Journal*, **14**, 1–9.

Zheng, H. and Zhang, Y. (2008). Feature selection for high-dimensional data in astronomy, *Advances in Space Research*, **41**, 1960–1964.

Zhou, Z.-H. (2015). Ensemble learning, *Encyclopedia of Biometrics*, Springer US, Boston.

Zuntz, J., Paterno, M., Jennings, E., Rudd, D., Manzotti, A., Dodelson, S., Bridle, S., Sehrish, S., and Kowalkowski, J. (2015). CosmoSIS: Modular cosmological parameter estimation, *Astronomy and Computing*, **12**, 45–59.

Von Neumann, J. (1941). Distribution of the ratio of mean square successive difference to the variance, *The Annals of Mathematical Statistics*, **12**, 367–395.

천문학에서의 대용량 자료 분석

신민수[a,1]

[a]한국천문연구원

(2016년 9월 19일 접수, 2016년 10월 5일 수정, 2016년 10월 6일 채택)

요 약

최근의 탐사 천문학 관측으로부터 대용량 관측 자료가 획득되면서, 기존의 일상적인 자료 분석 방법에 큰 변화가 있었다. 고전적인 통계적인 추론과 더불어 기계학습 방법들이, 자료의 표준화로부터 물리적인 모델을 추론하는 단계까지 자료 분석의 전 과정에서 활용되어 왔다. 적은 비용으로 대형 검출 기기들을 이용할 수 있게 되고, 더불어서 고속의 컴퓨터 네트워크를 통해서 대용량의 자료들을 쉽게 공유할 수 있게 되면서, 기존의 다양한 천문학 자료 분석의 문제들에 대해서 기계학습을 활용하는 것이 보편화되고 있다. 일반적으로 대용량 천문학 자료의 분석은, 자료의 시간과 공간 분포가 가지는 비 균질성 때문에 야기되는 효과를 고려해야 하는 문제를 가진다. 오늘날 증가하는 자료의 규모는 자연스럽게 기계학습의 활용과 더불어 병렬 분산 컴퓨팅을 필요로 하고 있다. 그러나 이러한 병렬 분산 분석 환경의 일반적인 자료 분석에서의 활용은 아직 활발하지 않은 상황이다. 천문학에서 기계학습을 사용하는데 있어서, 충분한 학습 자료를 관측을 통해 획득하는 것이 어렵고, 그래서 다양한 출처의 자료를 모아서 학습 자료를 수집해야 하는 것이 일반적이다. 따라서 앞으로 준 지도학습이나 앙상블 학습과 같은 방법의 역할이 중요해 질 것으로 예상된다.

주요용어: 천문학 자료, 통계 추론, 기계학습, 병렬 컴퓨팅, 분산 컴퓨팅

[1](34055) 대전광역시 유성구 대덕대로 776, 한국천문연구원. E-mail: msshin@kasi.re.kr