기계학습에 기반을 둔 소재물성 예측

DOI: 10.3938/PhiT.27.001

현 윤 경

Predicting Materials' Properties by Using Machine Learning

YunKyong HYON

Big Data has been already attracting attention in almost all research fields. Data analytics plays an important role in finding new values in almost all fields, such as society, culture, science, information communication, and the analysis and interpretation of many phenomena. Base on big data, machine learning has also been attracting attention as a method to solve problems in various fields such as social science and the physical and biological sciences. In particular, the development of deep learning methods and its numerous applications have been the focus of much attention. The research on and the development of machine learning algorithms, as well as the findings in fields to which they have been applied, such as the physical and the biological science, are expanding. Machine learning algorithms can also be used as modeling methods for highly complex problems and, in combination with object recognition, speech recognition, sensor data analysis, etc., in autonomous automobiles. Such algorithms are also being introduced into medical care in efforts to extending life and to provide for human welfare. Recently, machine learning has been suggested as a way to address issues with both industrial productivity and efficiency. Furthermore, many more applica-

tions and uses of machine learning can be found in many other areas. In this article, we present a mechanical learning methodology that can predict the properties of the new material that could not be obtained through conventional computer simulation. We discuss this machine learning method and show that its use should shorten for new materials development times by predicting their physical properties before having to perform material synthesis experiment.

들어가기 (빅데이터)

인터넷의 발달과 함께 1990년대 말 등장한 빅데이터는 2000년 들어 다른 어떤 것들보다 많은 주목을 받았다. 기술과 컴퓨터의 발달에 힘입어 대용량의 데이터를 처리할 수 있는 도구가 갖추어졌으며, 기존에 사용되지 않았던 데이터로부터 새로운 가치창출이라는 사회적 수요와 맞아 떨어졌다. 일부에서는 거품이라는 반론도 있었지만 기존에 버려졌던 데이터의 발굴과 활용은 충분한 가치를 가지고 있다. 빅데이터를 통하여 기존과학이 가지고 있던 한계를 극복하기 위한 새로운 과학의 패러다임도 만들어졌다. 최초 과학의 패러다임은 자연현상을 관찰하여 기록/분류/목록화하고 탐구하는 것이었다면, 다음 단계로는 관찰된 결과를 바탕으로 이론을 도출하고 연구하는 이론과학이 등장하였다. 또 한 단계 넘어 컴퓨팅 기술의 발전에 따라 현상의 해석을 위해 도출된 이론과 모델을 컴퓨터 시뮬레이션을 통해서 검증하는 계산과학의 시대가 열렸다. 오늘날은 컴퓨터의 발전과 계산방법에 발전이 지속적으로 이루어지고 이전 시대의 과학을 통해 쌓여진 데이터의 활용이 맞물려 지금 우리에게 많은 관심과 새로운 과학연구의 방향을 제시하고 있는 데이터과학(Data Science, Data-Intensive Science)이 새로운 과학의 시대를 이끌어 가고 있다. 이 새로운 과학의 흐름은 새로운 패러다임을 제시했고, (그림 1) 이는 기존의 실험과 가설 기반의 전통적 방법론을 지양하며 고기능·고정밀 시험 기기와 인터넷으로 연결된 세계 곳곳에 위치한 측정·관찰 기기로부터 생산된 대용량 데이터의 수집과 처리 그리고 분석을 통

저자약력

현윤경 박사는 KAIST 수리과학과에서 일반사각격자에서의 혼합유한요소 개발과 해석을 주제로 박사 학위를 받았으며, 이후 미국 미네소타대학(University of Minnesota) 내의 Institute for Mathematics and its Applications (IMA)에서 복잡유체에 대한 수리모델링과 계산을 연구했다. 2012년도 4월부터 국가수리과학연구소 선임 연구원으로 재직 중이다. 2013년부터 빅데이터 분석모델 개발과 수리과학 기반의 기계학습 기반의 데이터분석 기술 연구를 진행 중이다. (hyon@nims.re.kr)

Fig. 1. The Evolution of Science Paradigms.^[1]

Fig. 2. Weekly percentage of sentinel physician visits related to influenza-like illness (ILI) reported by the Centers for Disease Control and Prevention (CDC) and estimated using Google Flu Trends (GFT), United States, October 2010-March 2014.^[2]

해 과학적 발견을 시도하고 있다.^[1] 허블망원경으로부터 얻은 대용량이미지 데이터의 분석을 통한 새로운 초신성의 발견과 유럽의 강입자(hadron) 충돌실험을 통하여 힉스입자의 존재를 증명한 것이 데이터과학의 대표적인 예이다. 이러한 새로운 발견과 증명을 위한 과학방법에서 중요한 것은 실험을 위한 거대한 실험 장치를 만들고 운영하는 것뿐만 아니라 이로부터 수집되는 대용량의 데이터 분석에 많은 연구자들의 공동노력이 이러한 결과를 도출했다는 것이다. 데이터과학의 중요한 역할은 확보된 대용량의 데이터를 기반으로 이전에는 시도하지 못했던 복잡한 현상을 규명하고 거대 도전 과제의 해결을 모색할 수 있는 여건을 갖추었다는 것이다. 그러나 이러한 새로운 발견은 데이터 공유·융합뿐만 아니라 학제 간 연구 분야 간 협력이라는 기반이 반드시 필요하다. 빅데이터의 활용분야는 기존에 도전하지 못했던 거대한 과학적 발견에만 국한되지는 않는다. 인간의 삶과 질을 개선할 수 있는 사회과학적 분야

Fig. 3. Neural network training pipeline and architecture.^[3]

에서 활용이 보다 적극적이다. 데이터분석을 통한 추이(trend)와 패턴(pattern)의 예측에서 대표적인 사례 중에 하나가 구글 트렌드(Google Trends) 서비스를 활용한 독감의 유행 수준을 파악한 것이다.(그림 2) 구글은 검색기술의 장점을 활용하여 독감자체의 발생보다는 독감에 걸리면 나타나는 증상들에 관한 검색어가 얼마나 자주 검색됐는지를 파악해 독감 확산을 예측했다.^[2] 당시 이것은 미국 질병 통제 예방센터(Centers for Disease Control and Prevention)의 공표보다 일주일에서 열흘 앞서 독감의 창궐을 탐지해 냈다. 이외에도 다양한 사례들이 존재한다. 특히 구글번역기는 우리 일상생활 속으로 들어온 빅데이터 활용의 전형적인 예이다. 이러한 빅데이터를 기반으로 하여 대용량 데이터 처리와 분석에 적합한 알고리즘 혹은 프레임워크의 개발이 더욱더 중요해졌으며 이를 자동으로 처리하고 분석하는 기계학습이 다시금 주목을 받게 되었다.

REFERENCES

- [1] The Fourth Paradigm: Data Intensive Scientific Discovery, 2009 Microsoft research.
- [2] Martin LJ, Xu B, Yasui Y, PLoS ONE 9(12), e109209 (2014). doi:10.1371/journal.pone.0109209.
- [3] David Silver *et al.*, Nature 529, 484 (2016).

Fig. 4. Empirical evaluation of AlphaGo Zero.^[4]

기계학습

여기서 우리가 인공지능을 다루는 것은 아니지만 기계학습을 언급하면서 인공지능을 떼어놓고 말하기 어렵다. 최근 기계학습을 기반으로 하는 인공지능의 연구와 개발이 더욱더 활발해지고 있으며, 벌써 2년 전에 바둑기사들의 수많은 기보 데이터를 학습(그림 3)한 구글딥마인드(Google DeepMind)의 알파고(AlphaGo)가 큰 반향을 일으켰다.^[3] 발전을 거듭하여 얼마 전에는 규칙기반(rule based)의 강화학습(reinforcement learning)으로 무장한 알파고제로(AlphaGo Zero)가 바둑에 있어서이지만 인간의 한계를 뛰어 넘었다.(그림 4)^[4] 인공지능의 개발은 과학의 새로운 발견과 해석에 대한 접근보다는 인간의 삶의 질을 향상하는 기술개발 특히 의료데이터 분석을 통한 질병예측에 많은 노력을 기울이고 있다.^[3,5] 기계학습의 태생 자체가 인공지능을 개발하기 위한 방법으로 제시되었으며 빅데이터 이전에 기계학습과 관련된 많은 연구들이 이루어져오고 있었다. 그럼에도 불구하고 빅데이터의 등장인 기계학습의 중요성을 더욱더 강조하며 기계학습과 인공지능의 시대를 이끄는 포석이 된 것은 사실이다. 여기에서는 인공지능을 다루지 않고 기계학습의 장점을 활용하여 새로운 소재를 예측하는 연구와 결과에 대하여 다음 섹션에서 집중적으로 다룰 것이다. 우선 기계학습은 크게 지도학습(supervised learning)과 비지도학습(unsupervised learning)으로 구분하며^[6] 지도학습은 예측하고자 하는 것에 대한 답을 알고 있는 트레이닝 데이터(training data)라 불리는 데이터집합(data set)이 주어진 상태에서 기계학습 알고리즘을 통하여 예측모델을 만들어 답을 모르는 테스트 데이터(test data) 집합에 대하여 특성을 분석하거나 예측하는 일련의 학습을 의미한다. 대표적으로 서포트 벡터 머신(support vector machine), 은닉 마르코프 모델(Hidden Markov model), 회귀 분석(Regression), 신경망(Neural network), 나이브 베이즈 분류(Naive Bayes Classification) 등이 지도학습 알고리즘에 해당한다.^[7] 후자인 비지도학습은 트레이닝

Fig. 5. Typical CNN architecture.^[9]

데이터가 없는 상태에서 해당 속성을 예측하는 기계학습 방법을 말하며, k-means, mixture models, hierarchical clustering과 같은 군집화(Clustering) 방법과 독립 성분 분석(Independent Component Analysis)이 대표적인 예이다. 또한 최근 많은 관심을 받고 있는 인공신경망(Artificial Neural Network)을 이용한 Autoencoders, Deep Belief Nets(DBN), Hebbian Learning, Generative Adversarial Networks(GAN) 등이 있다.^[8] 인공신경망은 현재의 기계학습 연구와 활용을 이끄는 중요한 축 중 하나이다. 과거 인공신경망의 단점이었던 수학연산 XOR 계산능력을 숨겨진 층(hidden layer)을 도입하여 해결한 것이 심층학습(Deep Learning)의 핵심이다.^[10] 합성곱 신경망(Convolutional Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN), 제한된 볼츠만 머신(Restricted Boltzmann Machine, RBM), 심층 신뢰 신경망(Deep Belief Network, DBN), 심층 Q-네트워크(Deep Q-Networks) 등 다양한 심층학습들이 연구 개발되고 있으며,(그

REFERENCES

- [4] David Silver *et al.*, Nature **550**, 354 (2017).
- [5] <https://www.ibm.com/watson/>.
- [6] Christophe Bishop, *Pattern Recognition and Machine Learning* (Springer, 2011).
- [7] https://en.wikipedia.org/wiki/Supervised_learning.
- [8] https://en.wikipedia.org/wiki/Unsupervised_learning.
- [9] https://en.wikipedia.org/wiki/Convolutional_neural_network.
- [10] https://en.wikipedia.org/wiki/Deep_learning.

Fig. 6. The machine (or statistical) learning methodology.^[11]

림 5) 활용분야는 영상인식(Image recognition), 자연어처리(Natural language processing), 신약개발과 독성학(Drug discovery and toxicology), 추천시스템(Recommendation systems), 생물정보학(Bioinformatics), 자동음성인식(Automatic speech recognition) 등과 같이 광범위하게 확장되고 있다. 심층학습의 장점 중 하나는 보통의 데이터 분석에 있어서 데이터 도메인에 대한 통찰력을 반영한 속성(feature)을 선택하고 속성 간의 상관관계를 분석하는 속성에 대한 전처리가 필요하지 않다는 것이다. 오히려 분석에 필요한 속성을 자동으로 선택한다. 이러한 장점은 데이터의 속성에 대한 기존의 통찰력이 없거나 매우 복잡해서 분석해 내기가 어려운 경우에 매우 유용하다. 반대로 중요한 속성을 알고 있는 경우 이를 활용하기에는 구조적으로 복잡하다는 단점이 있다.

기계학습 기반 소재물성 예측

양자역학에서 도출된 밀도범함수 이론(Density Functional Theory, DFT)에 기반을 둔 제일원리계산(first-principles calculations)은 소재분야에서도 물성을 계산하는데 주요한 방법으로 등장했다. 빅데이터에서와 마찬가지로 컴퓨팅 알고리즘과 계산자원의 확대는 제일원리계산의 확대와 이를 통하여 소재와 관련한 수많은 계산이 가능하게 되었다. 이 결과로써 다양한 물성에 대한 데이터가 충분히 축적이 되었다. 이는 데이터과학을 적용할 수 있는 만큼 유효한 데이터 양이 되었으며, 향후 더 많은 데이터의 전략적인 축적은 더 효율적이고 정확한 예측모델을 구축할 수 있을 것이다. 하지만 어떤 경우는 계산 알고리즘에 종속성이 높고, 계산자의 기호에 따른 알고리즘 활용과 변수들의 적합도 문제는 여전히 기계학습의 적용에 있어서 해결해야 할 많은 어려운 점이 있다.

Fig. 7. Learning performance of electron charge density-based fingerprint vectors.^[11]

Fig. 8. Prediction performance of the developed linear and non-linear learning models.^[13]

소재정보학에 있어서 데이터과학의 적용은 기존 계산으로 도출된 입력과 출력의 결과데이터를 바탕으로 기존의 제일원리계산을 하지 않고 소재의 물성을 예측함으로써 계산에 필요한 시간을 단축할 뿐만 아니라 관찰과 계산에서 찾기 어려웠던 물성을 예측하고자 하는 목적을 가지고 있다.(그림 6) 전자의 경우 많은 연구를 통하여 예측가능성이 입증되고 있다.^[11-13] 또한 소재의 비선형적인 물성을 보다 정확하고 효과적으로 예측할 수 있는 속성(feature)에 대한 연구와 그에 적합한 기계학습 방법론의 연구가 진행되고 있다.^[12] 6개의 기본함수인 x , $x^{1/2}$, x^2 , x^3 ,

REFERENCES

- [11] G. Pilania *et al.*, Scientific Reports 3, 2810, DOI: 10.1038/srep02810.
- [12] G. Pilania *et al.*, Scientific Reports 6, 19375 | DOI: 10.1038/srep19375.
- [13] C. Kim *et al.*, Chem. Mater. 28, 1304 (2016).

Fig. 9. Prediction of bandgap in Gradient boosting tree regression with Computational Materials Repository (CMR) double perovskite data.

Fig. 10. Feature importance in gradient boosting tree regression with Computational Materials Repository (CMR) double perovskite data.

$\ln(1+x)$, e^x 로부터 쉽게 1,245,944개의 합성속성을 구성할 수 있으며 이 합성속성들에 대한 기계학습의 적용을 통하여 유효 속성을 탐색할 수 있으며 비선형물성에 대한 예측모델을 구성하여 물성을 예측할 수 있다.(그림 7, 8)^[12] 그리고 소재에 있어서도 기존의 이론과 계산방법으로 설명하기에 매우 복잡한 물리현상의 경우 데이터 기반의 모델을 세우고 기계학습을 적용할 경우 효과적으로 해석이 가능해졌다.^[13] 이러한 비선형 특성을 가지고 있는 데이터의 분석에 적용이 가능한 방법으로 gradient boosting 타입의 결정트리(decision tree) 방법이 있다.^[14] 이를 이용하여 이중 페로프스카이트(double perovskite) 데이터를 기반으로 예측모델을 만들고 밴드갭(bandgap)을 예측할 수 있으며 그에 대한 결과는 그림 9와 같다.

이 기계학습 방법은 비선형 특성을 가지고 있는 데이터를 분석하고 예측하는 데에 장점을 가지고 있으며, 또한 속성이 예측에 있어서 어느 정도로 역할을 하는지 그 중요도(feature importance)로 정의하여 제공하고 있다.(그림 10) 이는 해당 데이터 도메인에 대한 기존 통찰력을 통한 데이터 분석의 효과를 기계학습 방법론과 비교할 수 있는 중요한 요소를 제공하고 있어서 기계학습을 통한 예측결과의 해석과 설명에 필요한 근거를 제시

하고 있다.

데이터속성의 중요도는 기계학습에서의 차원축소법(dimensional reduction) 방법과 연계될 수 있으며 데이터분석에 있어서 효율적인 속성의 활용에 중요한 정보를 제공할 수 있다.

이와 같이 소재정보학 또는 신소재의 예측이나 후보물질의 도출에 있어서도 충분한 소재 데이터의 확보 위에 기계학습 방법을 적용하면 기존에 과학적 접근방법보다 나은 효율을 얻을 수가 있다. 더욱더 다양한 소재데이터와 기계학습의 접목이 기존에 해결하지 못한 문제들을 해결할 수 있을 것으로 기대되며, 향후 다양한 소재데이터에 대한 기계학습 방법론의 적용에 대한 연구결과들이 쏟아져 나올 것으로 기대된다.

앞으로

앞서 기술된 기계학습 방법론 이외에도 전통적인 이론에 기반을 둔 새로운 방법론들이 등장하고 있어 이를 간단히 소개하고자 한다. 소재에 있어서 분자의 구조정보는 합성 후의 물성을 결정하는데 중요한 역할을 한다는 것은 물성을 예측하는 기계학습의 적용에 있어서 매우 많은 정보를 제공한다. 특히 기계학습의 예측정확도와 효율을 결정하는 속성(feature)의 결정에 있어서 매우 중요하다.

1. 그래프 합성신경망(Graph Convolutional Neural Networks, GCNN)

분자구조 등 구조의 정보를 나타내는 방법으로는 그래프이론을 적용해 볼 수 있다. 그래프이론은 구조의 위치, 결합과 관련된 속성을 나타낼 수 있는 도구를 제공하며 이를 해석해 낼 수 있는 엄밀한 이론을 제공함으로써 보다 정확하고 효과적인 결과를 도출할 수 있다. 이를 적용한 것 중에 하나가 그래프 합성신경망이다.^[15] 그림 11은 이를 crystalline 소재 디자인에 적용한 예이며, GCNN의 구조를 표현한 것이며, 이 방법을 활용한 예측결과와 그림 12와 같다.

2. 위상학적 데이터 분석(Topological Data Analytics, TDA)

구조의 본질적인 속성을 해석해내는 이론이 위상학이다. 이를

REFERENCES

- [14] Chen, Tianqi, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System" (2016). CoRR. abs/1603.02754. arXiv:1603.02754.
- [15] T. Xie *et al.*, Crystal Graph Convolutional Neural Networks for Accurate and Interpretable Prediction of Material Properties, arXiv:1710.10324.

Fig. 11. Illustration of the crystal graph convolutional neural network (CGCNN).^[15]

Fig. 12. Extraction of site energy of perovskites from total formation energy.^[15]

데이터분석에 적용한 것이 위상학적 데이터 분석(TDA)이다. 위상학적 데이터 분석은 기존에 데이터분석 방법론이 분석해 내지 못한 새로운 것을 찾아내기도 한다.^[16-18] 그림 13은 위상학적 데이터 분석을 통하여 암에 대한 분석결과를 데이터시각화를 통하여 나타낸 것이다. 소재분야에서 있어서도 위상학적 물성에 대한 통찰력을 바탕으로 위상학적 데이터분석을 적용한 연구결과가 나오고 있다.^[19] 향후 데이터분석과 예측에 있어서 다양한 활용이 이루어질 것으로 기대된다.

나오는 글

빅데이터 시대를 지나 기계학습과 인공지능 기반의 4차 산업 혁명이 이끄는 시대에 들어가고 있다. 기계학습과 인공지능의 역할은 사회·과학적 요구와 맞물려 시간이 갈수록 더욱더 증가하고 있다. 이러한 흐름에서 과학적 패러다임 역시 새로운 국면으로

Fig. 13. Progression Analysis of Disease analysis of the Netherlands Kanker Instituut data.^[18]

접어들고 있다. 이러한 변화 속에서 데이터의 가치를 인식하고 데이터의 적극적 활용이 필요한 시기이다. 다른 분야와 달리 소재정보학 분야에서의 기계학습도입은 조금은 늦었지만 많은 연구자들이 관심을 가지고 적극적으로 연구를 진행하고 있으며, 앞서 살펴본 것과 같이 여러 연구결과들이 신소재 개발 및 소재정보학에서의 기계학습의 가능성을 입증하고 있다. 우선 해결해야 할 것은 유효한 데이터의 확보를 통한 빅데이터 구축이며, 연구에 있어서는 여러 분야 간의 융합이다. 또 하나 중요하게 생각해야 할 것은 기계학습과 인공지능이 전통적인 과학적 방법을 완벽히 대체하는 것이 아니라는 것이다. 데이터의 관점에서만 보더라도 기계학습과 인공지능은 데이터분석과 활용 면에서의 역할이라는 위치를 가짐으로써 기존의 연구방법과 상호보완적인 위치에 있다. 이러한 새로운 과학방법에 대한 이해를 통해 유효성을 확인하고 관심분야에서의 적극적인 활용을 함으로써 새로운 가치의 발굴과 새로운 현상의 이해가 이루어질 것으로 생각된다.

REFERENCES

- [16] Gunnar Carlsson, Bulletin of The American Mathematical Society **46**, 255 (2009).
- [17] Gurjeet Singh *et al.*, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, Eurographics Symposium on Point-Based Graphics **22**, (2007).
- [18] Monica Nicolau *et al.*, PNAS **108**, 7265 (2011).
- [19] Yongjin Lee *et al.*, Nature Communications **8**, 15396 (2017).