기계학습을 이용한 소재 연구

DOI: 10.3938/PhiT.26.035

임진오

Machine Learning in Materials Science

Jino IM

The great success of the human genome project not only changes research paradigm in various fields of biology, but also impacts other field outside biology. In material science, the concept of materials gene has been developed as a counter part of biological gene, and it suggests new research paradigm and enables inverse design of materials. This new paradigm is named as “material informatics.” Machine learning is central in materials informatics, which provides an efficient way to understand materials database and to perform data-driven research. This article aims to introduce the concept of materials genome and material informatics, and the application of machine learning to real research of materials science.

Materials Genome

“인간게놈 프로젝트(Human Genome Project)”는 21세기 최대의 과학적 성과 중의 하나로 손꼽힌다.^[1] 이 프로젝트에서는 인간의 모든 유전자 염기 서열에 대한 정보, human genome의 구조(structure and organization)와 인간의 수명 등에 기여하는 특별한 기능(function)을 가진 유전자와 단백질 역시도 밝히며, 인간을 구성하는 요소를 이해하는 바탕을 제시하였다. 이는 의학, 생명공학의 발전을 가져왔으며, 광범위한 영역에서 생물학 연구의 새로운 패러다임을 제시하였다. 더 나아가 이러한 인간 게놈 프로젝트의 성공은 다른 과학 분야에까지 많은

저자약력

임진오 박사는 서울대학교 물리학과에서 이학박사(2009) 학위를 받은 후 미국 University of California, Davis에서 Postdoctoral Scholar(2009-2010), Northwestern University에서 Postdoctoral Fellow(2010-2014)를 거쳐 현재 한국화학연구원에서 선임연구원으로 재직 중이다. (imjino@krci.re.kr)

Fig. 1. Upper panel shows a schematic picture of biological gene and corresponding protein. Lower panel shows schematic picture of materials gene and corresponding property.

영향을 끼쳤고, 재료 과학 분야에서는 새로운 패러다임으로서 “materials gene”의 개념의 도입과 “소재 정보학(material informatics)”이 제시되었다.

Material gene의 개념은 1990년대 말부터 생물학적인 연구 방법을 재료과학의 연구에도 도입해보려는 시도와 함께 발전되었고,^[2-4] 특히 미국의 Materials Genome Initiative^[5]가 시작되면서 많은 소재 연구에서 등장하는 개념이 되었다. 하지만 아직까지 material gene의 개념은 하나의 고유한 물리량으로써 정의되지는 못하였고, 물질을 구성하는 구성 성분으로서 결정 구조의 특성이나 구성원소의 특성으로 널리 이해되고 있다. 그림 1은 materials gene의 하나의 예시로서 물질의 밀도, 화학식, 구성원소의 질량 등의 정보를 보여준다. 이중 특정한 material gene의 집합이 어떠한 물질의 특성과 깊은 상관관계가 있는 경우, 생물학적 유전자(biological gene)의 배열(DNA sequence)로부터 특별한 기능을 하는 단백질이 발현되는 과정과 유사성을

REFERENCES

- [1] <https://ghr.nlm.nih.gov/primer/hgp/description>.
- [2] Alan L. Mackay, J. Mol. Struct. **336**, 293 (1995).
- [3] J. H. E Cartwright and A. L. Mackay, Phil. Trans. R. Soc. A **370**, 2807 (2012).
- [4] Krishna Rajan, Annu. Rev. Mater. Res. **45**, 153 (2015).
- [5] <https://www.mgi.gov>.

찾을 수 있다. 이러한 유사성은 인간게놈 프로젝트에서 추구되었던 것과 같이 materials gene의 배열과 물성들의 상관관계를 모두 이해함으로써 “inverse design”^[6]과 같은 소재 연구의 새로운 패러다임이 가능함을 보여준다.

하지만 소재의 결정구조나 분자구조의 복잡성 때문에 materials gene과 물성들 간의 상관관계를 찾는 것은 어려운 문제이다. 이러한 문제를 풀기 위해서 최근에 주목받고 있는 방법은 통계학적인 접근 방법이다. 이 새로운 방법에서는 신소재를 찾기 위해 기존의 Edisonian, 즉 하나의 물질을 집중해서 이해하는 방법에서 벗어나, 많은 소재의 구조와 물성 정보를 담고 있는 데이터베이스를 구축하여, 신소재에서 요구되는 물성과 가장 상관관계가 큰 materials gene을 통계적으로 분석하는 것이다. 소재정보학은 이러한 통계학적 연구방법을 실제 소재 연구에 이용하는 학문을 지칭한다.^[7]

지난 10년간 새로운 연구 패러다임으로서 소재정보학이 관심 받게 될 수 있었던 이유 중 하나는 소재 데이터베이스의 폭발적인 확장을 들 수 있다. 조합 화학으로 대표되는 대량 합성방법의 발전은 high-throughput 합성 및 평가를 가능하게 함으로서 대량의 실험 결과를 데이터베이스화할 수 있게 하였다.^[8] 더 나아가 지난 10년간의 멀티 코어 CPU의 발전은 컴퓨팅 파워를 기하급수적으로 증진시켰고, 이는 수십만 개의 물질에 대한 양자역학 기반의 물성계산을 가능하게 하였다. 이렇게 대량 생산된 물성 정보는 데이터베이스로서 저장되고 축적되어 “소재 빅데이터(materials big data)”로 진화하게 되었다.

이러한 소재 빅데이터를 생성하기 위해서는 많은 계산 자원과 인적 자원이 필요하기 때문에 다수의 프로젝트가 국가적 지원을 통해서 진행되었다. 미국의 경우에는 Materials Genome Initiative(MGI)가 2011년부터 NSF, DOE, NIST 등의 다수의 국가 연구 기관과 다수의 대학연구진들이 참여하여 진행하고 있고,^[9] 유럽의 경우 독일을 중심으로 Novel Materials Discovery (NOMAD) 프로젝트가 진행되고 있다.^[9] 가까운 일본의 경우도 재료연구소(NIMS)를 중심으로 MAT Navi라는 물성 데이터베이스를 생성하고 있다.^[10]

소재 빅데이터가 실제 연구에 이용되기 위해서는 데이터로부터 의미 있는 정보를 추출하는 과정이 중요하다. 이를 위해서 데이터 마이닝, 진화 알고리즘, 통계적 분석과 같은 정보학의 방법들이 이용될 수 있다. 이 중 최근 가장 각광을 받고 있는 것은 machine learning이다. Machine learning은 데이터베이스로부터 효과적인 컴퓨터 학습을 통해 데이터들 사이에서 나타나는 특별한 패턴을 분석하거나 예측모델을 만드는 방법들을 지칭한다.^[11] Machine learning은 소재 정보학에서 본격적으로 응용되기 이전부터 컴퓨터 체스나 바둑 같은 cognitive game theory나 얼굴인식이나 지문분석과 같은 pattern recognition,

그리고 bioinformatics 분야에서 눈부신 성과를 보여주었다. 소재 정보학에서는 지난 5년간 machine learning을 이용하는 기초적인 연구가 다수의 국외의 연구 그룹에서 활발히 진행되었고, 최근 들어 국내의 연구진에 의해서도 시도되고 있다.

Machine learning과 소재정보학

Machine learning을 실제 소재 연구에 적용하기 위해서는 1) 데이터베이스, 2) machine learning 알고리즘, 그리고 3) Features의 세 가지 구성요소가 필요하다.^[11] 소재 정보학에서는 많은 물질들의 구조 정보 및 물성 정보가 첫 번째 요소인 데이터베이스에 포함된다. 정확하고 효과적인 machine learning 분석을 위해서는 데이터베이스에 포함되는 물질의 개수가 많아야 하고, 정형적이고 완전한 형태의 데이터베이스가 유리하다. 이러한 관점에서 실험에서부터 얻어지는 데이터는 제한적이기 때문에, 소재정보학에서는 계산으로 얻어지는 물성 데이터베이스의 활용도가 높다. 하지만 이러한 경우 실험에서 측정된 물성 값과 계산된 물성 값의 오차가 고려되어야 한다.

두 번째 구성요소인 machine learning algorithm은 데이터베이스의 형태와 분석하고자 하는 목적에 따라 supervised learning과 unsupervised learning으로 구분된다. Supervised learning의 경우에는 input과 output data 모두의 정보를 가지고 data fitting(regression for continuous variables and classification for discrete variables)을 하여, 주어진 새로운 input에 대하여 output을 예측하는 모델을 만든다. 예를 들면 소재정보학에서는 결정 구조의 특성을 input variable로 사용하여 밴드갭과 같은 전자 구조적인 물성을 예측하고자 할 때 supervised learning 방법이 사용될 수 있다.^[12] 또 다른 하나인 unsupervised learning은 output이 없이 input data의 특정한 statistical pattern이나 clustering을 찾는다. Unsupervised learning 방법을 사용하는 연구의 예로서 결정 구조의 유사함을 판단하여 phase diagram으로 나타내는 일을 예로 들 수 있다.^[13] 표 1에 이러한 machine learning 방법의 분류와 분류에 따른 응용분

REFERENCES

- [6] <http://www.centerforinversedesign.org/approach.html>.
- [7] J. R. Rogers and D. Cebon, MRS Bulletin **31**, 975 (2006).
- [8] X. D. Xiang, *et al.*, Science **268**, 1738 (1995).
- [9] <https://www.nomad-coe.eu>.
- [10] http://mits.nims.go.jp/index_en.html.
- [11] C. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).
- [12] G. Pilania *et al.*, Scientific Report **6**, 19375 (2016).
- [13] C. J. Long *et al.*, Review of Scientific Instruments **78**, 072217 (2007).

Table 1. Machine learning algorithm and its applications.^[14]

Machine learning	Algorithm	Applications
Supervised learning	Regularized least squares Support vector machines Kernel ridge regression Neural networks Decision trees Genetic programming	Predict processing-structure property relationships; Develop model Hamiltonians; Predict crystal structures; Classify crystal structures; Identify descriptors
Unsupervised learning	K-means clustering Mean shift theory Markov random fields Hierarchical cluster analysis Principal component analysis Cross-correlation	Analyze composition spreads from combinatorial experiments; Analyze micrographs; Identify descriptors; Noise reduction in data sets

아가 정리되어 있다.

마지막으로 machine learning을 통해 소재 정보학의 여러 문제를 잘 해결하기 위해서는 적절한 feature를 선택하는 것이 중요하다. Feature는 데이터베이스의 개별적인 데이터를 정의 하는데 사용되는 값으로써 machine learning을 적용하는데 있어서 input variable이 된다.^[11] 소재 정보학에서는 데이터베이스의 기본적인 구성요소가 개별적인 물질이 되기 때문에 여기서의 feature는 개별적인 물질들의 특성을 설명할 수 있는 정보들, 예를 들면 결정 구조의 symmetry, chemical composition, density 등이 포함될 수 있다. Features는 데이터를 정의하고 설명할 수 있다면 어떠한 값도 가능한데, 데이터베이스에 포함되는 정보의 성격이 응용분야에 따라 판이하게 다르기 때문에 일률적인 방법으로 선택하여 사용할 수 없다. 따라서 각각의 응용분야(domain)에 따라 그 분야의 전문적 지식, 즉 다시 말해 domain knowledge를 이용하여 feature를 선택하는 것이 중요하다. 이렇게 domain knowledge를 이용하여 feature를 선택하는 것을 feature engineering이라고 부르는데 machine learning을 이용하는데 있어 가장 중요한 단계로 여겨진다. Machine learning의 권위자 중에 한 명인 Andrew Ng 교수는 다음과 같은 말로 feature engineering에 대한 중요성을 설명하였다.

“Coming up with features is difficult, time-consuming, requires expert knowledge. “Applied machine learning” is basically feature engineering.”^[15]

소재 정보학에서는 feature engineering을 통해서 선택된 정보들이 의미 있는 materials gene으로 이용될 수 있기 때문에, 많은 소재 정보학 연구들이 어떤 feature를 선택하여 이용할 것인지에 대한 주제를 다루고 있다. 그 한 가지 예로서 Luca M. Ghiringhelli와 그의 동료들은 최근 논문을 통해서 소재 정보학

에서 feature가 가져야 할 필수적인 성질에 대해서 다음과 같이 설명하고 있다.^[16]

1. Feature는 물질의 기본적인 성질로서 물질을 고유하게 정의하여야 한다.
2. 서로 크게 다른 물질들은 feature에서도 큰 값의 차이가 있어야 한다.
3. Feature는 물질로부터 쉽게 얻어질 수 있는 값으로 정의되어야 한다.
4. Feature space의 dimension은 작을수록 좋다.

다음 섹션에서는 Machine learning을 실제 소재 연구에 적용한 몇 가지의 예를 소개해 보고자 한다.

Machine learning을 이용한 소재 연구 예시

Prasanna V. Balachandran과 그의 동료들은 2011년에 발표된 논문을 통해서 선형 regression 방법을 사용하여 high-temperature piezoelectric perovskite 물질의 Curie temperature를 예측하는 연구를 소개하였다.^[17] 이 연구에서는 실험을 통해서 측정된 20개의 piezoelectric perovskite 물질의 Curie temperature(T_c)를 포함하는 데이터베이스와 principal component analysis(PCA)와 partial least square 방법을 사용하여

REFERENCES

- [14] T. Mueller, A. G. Kusne and R. Ramprasad, *Reviews in Computational Chemistry* (Wiley, 2016), Vol. 29, Chap. 4.
- [15] Andrew Ng, *Machine Learning and AI via Brain simulations* (Stanford University).
- [16] Luca M. Ghiringhelli *et al.*, Phys. Rev. Lett. **114**, 105503 (2015).
- [17] P. V. Balachandran, S. R. Broderick and K. Rajan, Proc. R. Soc. A **467**, 2271 (2011).

Fig. 2. Multivariate predicted model in comparison with measured T_c . Filled circles, training set; filled triangle, test set; plus symbols, new predictions. (from Ref. 17)

Fig. 3. Comparison between predicted and computed DOS_F . (from Ref. 18)

regression을 수행하였다. PCA는 feature engineering을 수행하는 가장 기본적인 방법 중에 하나로서 여러 개의 feature를 포함하는 feature space에서 서로 연관성이 떨어지는 feature를 추출할 수 있고, 이것을 통해서 축소된 feature space에서 데이터를 분석할 수 있게 된다. 이 연구에서는 30개의 feature들을 PCA를 통해서 6개의 주요 feature로 축소하였고, 축소된 feature space에서 linear regression을 통해 T_c 에 대한 예측모델을 제시하였다.(그림 2) 이를 통해서 높은 T_c 를 얻기 위한 design rule을 제시하여 두 개의 새로운 high T_c piezoelectric

a) 300 K

b) 1000 K

Fig. 4. Classification trees for HH compounds at 300 K (top) and 1000 K. Printed on each node are the number of compounds it contains, and the fraction of them above the third quartile of the ZT distribution. Left-leaning branches correspond to increased percentages of compounds in the best- ZT category. (from Ref. 19)

perovskite 물질을 디자인하는 데 사용하였다.

앞에서 소개한 선형 방법은 적절한 feature를 선택하였을 경우 매우 유용한 정보를 제공하지만, 보다 복잡한 관계식을 예측하기 위해서는 비선형 machine learning 방법을 적용할 필요가 있다. 지금까지 많은 비선형 machine learning 알고리즘이 개발되어 사용되고 있는데, 대표적으로 support vector machine, Kernel ridge regression, tree-regression, artificial neural network와 같은 방법들이 널리 사용된다. K. T. Schütt와 그의 동료들은 2014년에 발표된 논문에서 Gaussian 및 Laplacian Kernel을 사용한 Kernel ridge regression을 이용하여 7000개의 물질의 Fermi level density of state(DOS_F)를 예측하는 연구를 수행하였다.^[18] 이 연구에서 사용된 feature는 partial radial distribution function인데 이것은 삼차원상에 분포되어 있는 원자들의 상대적인 분포도를 1차원에 투영한 물리량인데, 이를 통해서 두 물질의 구조적 차이를 수치화하여 machine learning의 feature로써 사용하였다. 그 결과로서 0.68×10^{-2} states/eV/Å³의 에러 값을 가지는 DOS_F 예측모델을 도출하였다

REFERENCES

- [18] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).
- [19] Jesús Carrete, Natalio Mingo, Shidong Wang and Stefano Curtarolo, Adv. Func. Mater. **24**, 7427 (2014).

Fig. 5. Schematic to illustrate differences between standard, single-group research (left) and new opportunities afforded by large materials databases (right). The standard research cycle is considerably longer because testing and refining hypotheses typically involves several time-consuming data collection steps. In contrast, the availability of large data sets in the data-driven approach enables multiple hypotheses to be tested simultaneously through novel informatics-based approaches, reducing the burden of data collection. (from Ref. 20)

다.(그림 3) J. Carrete와 그의 동료들은 또 다른 비선형 방법인 tree-based regression 방법을 사용하여 75개의 half-heusler(HH)의 ZT 값을 분석하였다. 그 결과로서 구성원소의 electronegativity 차이, 원소 번호 차이, 주기율표에서의 group의 정보를 input variable로 이용하여 75개의 HH를 높은 ZT값을 가지는 그룹과 낮은 ZT값을 가지는 그룹으로 나누는 decision-tree를 구축하였다.(그림 4) 이를 통해 High ZT 물질을 만들기 위한 design rule을 제시하였다.

Perspective & Summary

고전적인 소재 연구 방법은 주어진 소수의 물질에 집중하여 실험과 이론적 분석을 수행하여 물질의 물성을 이해하는 방향으로 진행되었다. 하지만, 이러한 연구 방법은 복잡하고 다양한 물성 조건을 만족시키는 새로운 물질을 찾는 응용분야에서는 trial-and-error를 통해 새로운 물질을 제시할 수밖에 없기 때문에 소재 부분의 연구가 전체적인 연구 개발 속도를 늦추는 걸

림돌이 되었다. 소재 정보학은 이러한 에디슨이언(Edisonian) 방법에서 탈피하여 물질을 보다 포괄적으로 이해하는 새로운 관점을 제시한다. 소재 정보학에서는 materials gene과 물성 사이의 상관관계로부터 새로운 물질에 대한 design rule을 이해함으로써 물질 중심이 아닌 물성 중심의 inverse design을 목표로 한다. Machine learning은 소재 빅데이터를 효율적으로 분석하여 물성에 대한 정확하고 빠른 예측모델을 제시하고 구조와 물성 간에 보다 정량적인 상관관계를 도출함으로써 소재 정보학을 가능하게 한다.(그림 5) Machine learning을 소재 연구에 응용하는 것은 아직 초기 단계로서 feature engineering에 대한 연구에 집중되어 있으나, 추후에 재료과학과 정보학의 교류가 더욱 활발해짐에 따라 소재 정보학이 보다 빠르게 발전하기를 기대한다.

REFERENCES

- [20] Anubhav Jain, Kristin A. Persson and Gerbrand Ceder, *APL Mater* **4**, 053102 (2016).