유기소재 속성 예측을 위한 가우시안 프로세스 기반의 능동 학습

곽하녹 [1], 윤상웅 [2], 한철호 [1], 심문보 [3], 장병탁 [1,2]

[1] 서울대학교 컴퓨터공학부, [2] 서울대학교 협동과정 뇌과학전공, [3] 삼성전자 종합기술원

{ hnkwak, swyoon, chhan}@bi.snu.ac.kr, munbo.shim@samsung.com, btzhang@bi.snu.ac.kr

Active Learning with Gaussian Process for Predicting Organic Chemical Materials

Hanock Kwak[1], Sangwoong Yoon[2], Cheolho Han[1], Munbo Shim[3], Byoung-Tak Zhang[1,2]

[1]School of Computer Science & Engineering, Seoul National University

[2]Interdisciplinary Program in Neuroscience, Seoul National University

[3]Samsung Advanced Institute of Technology, Samsung Electronics Co., Ltd.

요 약

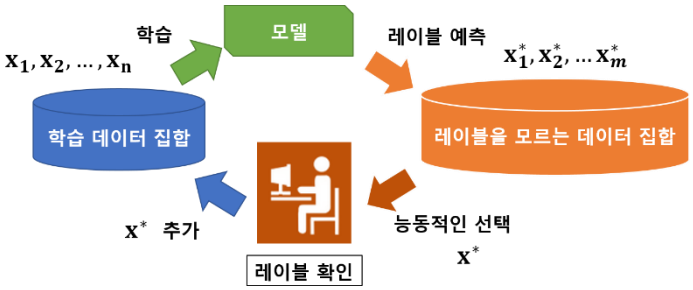
전자 구조 이론에 따른 고수준의 양자 화학적 계산법으로 분자 구조와 속성 사이의 의미 있는 상관 관계를 알아내는데 전례 없는 양의 데이터와 시간을 요구한다. 또한 기하급수적으로 늘어나고 있는 가상의 분자 구조의 속성을 매번 정확히 계산하는데 한계가 있다. 그래서 이러한 문제점을 보완하기 위해 다양한 기계학습적인 방법이 새롭게 시도되고 있고 일부 데이터에서는 성공적인 결과가 나왔다. 그러나 기계학습적인 방법 또한 많은 학습 데이터를 필요로 한다. 이러한 학습 데이터를 구축하는데 비용이 많이 들며 이러한 문제를 보완하기 위해서 능동 학습 관점의 접근 방식이 필요하다. 본 연구는 가우시안 프로세스 기반의 능동 학습 알고리즘인 ALM과 ALC를 이용하여 최소한의 학습 데이터로 높은 정확도의 예측을 가능하게 하였다.

1. 서 론

유용하고 효율적인 유기소재를 빠르게 찾을 수 있는 데이터 처리 방법 또는 과학적 계산법에 대한 산업계의 수요가 크다. 그러나 현재의 고수준의 양자 화학적 계산법으로는 분자 구조와 속성 사이에 내재된 상관관계를 명확히 밝혀내지 못하고, 분자의 속성을 정확히 알아내는데 상당한 시간이 필요하다. 게다가 구조적 규칙을 준수하는 가상의 분자 구조가 지속적으로 설계되고 있는 와중에, 매번 속성을 정확히 계산하는데 한계가 있다. 이러한 문제점을 보완하기 위해 다양한 통계학적, 기계학습적인 방법이 연구되었다[1, 3].

본 연구는 C, N, O, S 원자를 포함한 23 가지 원자로 구성할 수 있는 안정적인 분자 구조 정보를 갖는 GDB-13 데이터의 일부분을 사용하였다. 이 데이터의 분자 구조 정보로부터 높은 정확도로 속성(원자화 에너지, HUMO와 LUMO의 고유치, 들뜸 에너지, 평균 분자 분극 등)을 예측하는 인공 신경망 기반의 모델이 연구되었다[1].

본 논문은 기존의 접근 방식과는 달리 능동 학습 관점에서 이 문제를 다룬다. 능동 학습은 준감독 기계학습의 일종으로 최소한의 학습 데이터로 높은 성능을 내기 위해 능동적으로 새로운 학습 데이터를 선택하는 학습 방법이다[4]. 능동 학습은 데이터의 복잡성이 높고 학습 데이터를 확보하기 어려운 경우 응용된다[2]. [그림 1]은 능동 학습의 과정을 묘사하고 있다. 여기서 레이블은 예측하고자 하는 대상을 의미하며, 분자 구조 데이터인 경우 분자의 속성을 의미한다. 학습 효과가 큰 데이터를 학습 데이터 집합에 지속적으로 추가하여 비용이 많이 드는 레이블 확인 작업을 최소화할 수 있다.



[그림 1] 능동 학습의 과정

가우시안 프로세스 회귀모델은 데이터의 레이블을 평균과 분산의 확률 분포로 예측하는 기계학습 모델이다. 측정된 확률 분포는 능동 학습에 직접 활용될 수 있고, [1]에서 연구된 인공 신경망과 비교하여 비슷한 수준으로 원자화 에너지를 예측하였다. 또한 인공 신경망은 온라인 학습이 어렵다는 단점[7]이 있는 반면 가우시안 프로세스 회귀모델은 역행렬 정리를

이용하여 온라인 학습이 빠르고 정확하다.

본 연구에서 가우시안 프로세스 기반의 능동 학습 알고리즘으로 ALC와 ALM[5]을 사용하였고 성능을 비교하였다. 두 알고리즘 모두 능동 학습의 다양한 선택 전략 중에서 모델에서 예측하는 레이블에 대한 분산을 최대한 감소시키는 전략을 따른다.

2. 가우시안 프로세스 기반의 능동 학습

2.1 가우시안 프로세스

가우시안 프로세스 회귀모델은 확률적으로 정의된 함수 분포에서 데이터에 대한 함수를 예측하는 모델이다[6]. 임의의 n 개의 학습 데이터 $D = \{(\mathbf{x}_i, f_i) \mid i = 1, 2, \dots, n\}$ 를 토대로 임의의 m 개의 테스트 데이터 $\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_m^*$에 대한 함수 $f_1^*, f_2^*, \dots, f_m^*$을 추정한다. 가우시안 프로세스는 전체 입력 공간의 함수에 대해서 다변수 가우시안 분포를 정의한다. 이에 따라 함수 공간 내의 임의의 유한 집합 $\{f_1, f_2, \dots, f_k\}$은 다음과 같은 다변수 가우시안 분포를 갖는다.

$$[f_1, f_2, \dots, f_k]^T \sim \mathcal{N}[m, K] \qquad (1)$$

여기서 K 는 공분산 행렬, m 은 평균 벡터이다. K 의 (i,j)번째 원소는 임의의 두 데이터의 함수의 공분산을 구하는 커널 $k(\mathbf{x}_i, \mathbf{x}_j)$ 이다. 커널 k 는 사용자가 직접 설계하며 커널에 따라 추정되는 함수의 형태가 크게 달라지는데 일반적으로 유클리드 거리가 가까운 데이터의 함수는 상관성이 높다고 가정하는 RBF 커널을 사용한다.

정의에 따라 학습 데이터 D 로 임의의 데이터 $\mathbf{x}^*$ 의 함수 f^*를 다음과 같이 추정한다.

$$f^*|D, \mathrm{x}^* \sim \mathcal{N}\left(k^{*T}K^{-1}\mathrm{f},\ \mathrm{k}(\mathrm{x}^*, \mathrm{x}^*) - k^{*T}K^{-1}k^*\right) \qquad (2)$$

여기서 k^*는 $[k(\mathbf{x}_1, \mathbf{x}^*), k(\mathbf{x}_2, \mathbf{x}^*), \dots, k(\mathbf{x}_n, \mathbf{x}^*)]^T$ 이다. 함수 f^*는 가정된 다변수 가우시안 분포에서 학습 데이터를 기준으로 주변화되어 일차원의 가우시안 분포로 예측된다.

2.2 ALM, ALC 알고리즘

ALM 알고리즘[5]은 레이블을 모르는 데이터 중에서 분산이 가장 큰 데이터를 선택한다. 이러한 데이터는 현재의 학습 데이터로부터 예측이 어렵고 학습했을 때의 효과가 크다. 그러나 선택된 데이터가 전체 데이터 분포에서 대표성이 많이 떨어지는 경우, 학습 효과가 낮을 수 있다.

ALC 알고리즘[5]은 ALM하고는 다르게 전체 데이터 공간에 대한 분산을 최대한 감소시키는 데이터를 찾는다. 예측에 대한 평균 제곱근 에러는 편의-분산 분해(bias-variance decomposition)로 나누어지는데 모델이 올바르다고 가정한 경우 분산에 비해 편의의 영향은 작다. 이런 관점에서 ALC 알고리즘의 선택 기준이 합리성을 갖는다. 전체 데이터 공간에 대한 분산을 알기 어려우므로 데이터의 분포를 잘 나타내는 r 개의 참조 데이터 $\{\xi_1, \xi_2, \dots, \xi_r\}$ 를 기준으로 삼는다. 그러면 특정 후보 데이터 $\mathbf{x}$가 학습 데이터로 추가되었을 때, 참조 데이터의 총 분산의 변화 $\Delta\sigma^2$는 다음과 같다.

$$\mathbf{k}_\xi = [k(\mathbf{x}_1, \xi), k(\mathbf{x}_2, \xi), \dots, k(\mathbf{x}_n, \xi)]^T$$

$$\Delta\sigma^2(\mathbf{x}) = \sum_{i=1}^{r} \frac{\left(\mathbf{k}_{\xi_i}{}^T K^{-1}\mathbf{k}_\mathbf{x} - k(\mathbf{x}, \xi_i)\right)^2}{k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_\mathbf{x}{}^T K^{-1}\mathbf{k}_\mathbf{x}} \qquad (3)$$

ALC는 $\Delta\sigma^2$가 가장 큰 후보 데이터를 선택한다.

2.3 가우시안 프로세스 회귀모델의 온라인 학습

(2)식에서 역행렬 K^{-1}을 구하는 부분의 시간 복잡도가 $O(|D|^3)$ 으로 여기서 가장 많은 계산이 이루어진다. 새로운 학습 데이터 $\mathbf{x}^*$가 추가되었을 때, 다음과 같이 기존의 K^{-1} 을 활용하여 새로운 역행렬 K^{*-1} 을 $O(|D|^2)$의 시간복잡도로 구할 수 있다.

$$K^{*-1} = \begin{bmatrix} K & \mathbf{k}_{\mathbf{x}^*} \\ \mathbf{k}_{\mathbf{x}^*}{}^T & \mathbf{x}^* \end{bmatrix}^{-1} = \begin{bmatrix} [K^{-1} + \mu^{-1}\mathbf{g}\mathbf{g}^T] & \mathbf{g} \\ \mathbf{g}^T & \mu \end{bmatrix} \qquad (4)$$

$$\mathbf{g} = -\mu K^{-1}\mathbf{k}_{\mathbf{x}^*}, \qquad \mu = \left(k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_{\mathbf{x}^*}{}^T K^{-1}\mathbf{k}_{\mathbf{x}^*}\right)^{-1}$$

3. 실험 내용 및 결과

3.1 가우시안 프로세스와 인공 신경망 비교

GDB-13 데이터의 일부분에 대해서 인공 신경망(MLP)을 통해 실험한 결과[1] 중에 원자화 에너지를 가지고 가우시안 프로세스(GP)와 비교하였다. 분자 구조 데이터의 전처리 방식과 사용된 학습 데이터와 테스트 데이터 모두 동일하다. 총 5000개의 분자로 학습하였고 2211개의 분자의 원자화 에너지 값을 예측하였다. 실험에 사용된 분자의 원자화 에너지의 평균은 -1538.04, 표준편차는 223.92이고 단위는 kcal/mol이다.

	RMSE(kcal/mol)
GP	8.6787
MLP	8.3018

[표 1] 가우시안 프로세스와 인공 신경망 성능 비교

예측된 값의 RMSE(root mean square error)를 두 모델에서 비교한 결과가 [표1]에 나와있다. 가우시안 프로세스가 인공 신경망에 비해 정확도가 조금 낮지만 원자화 에너지의 표준편차를 고려한다면 미미한 수준이다.

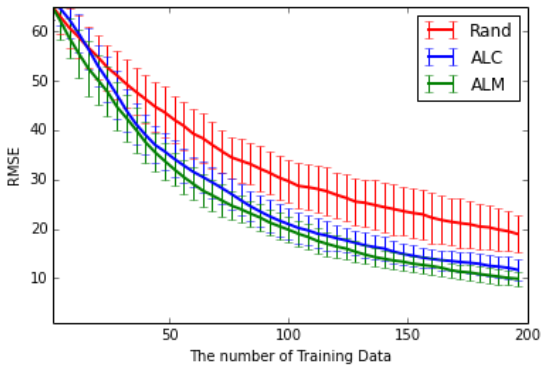
3.2 장난감 문제에서의 능동 학습

능동 학습의 효과를 알아보기 위해 가우시안 프로세스 기반의 능동 학습 모델로 다음과 같은 egg holder 함수를 예측하는 실험을 수행했다.

$$-(\mathrm{y} + 47)\sin\left(\sqrt{\left|x + \frac{y}{2} + 47\right|}\right) - x\,sin\left(\sqrt{|x - (y + 47)|}\right)$$

전체 데이터는 정의역 $[-100, 100]^2$에서 임의로 선택된 1000개의 점으로 하였고, 이 안에서 임의의 200개의

점을 테스트 데이터로 지정했다. 처음에는 하나의 학습 데이터로 시작하여 지속적으로 새로운 학습 데이터를 전체 데이터 중에 테스트 데이터를 제외한 부분에서 선택하여 추가하였다. 새로운 학습 데이터가 추가될 때마다 테스트 데이터의 함수를 예측하고 RMSE를 측정하였다. 임의로 선택하는 방법과 ALC, ALM을 사용하여 각각 100회 반복 실험해서 RMSE의 평균과 표준편차를 구했다. ALC의 참조 데이터는 테스트 데이터를 제외한 800개의 데이터로 하였다.

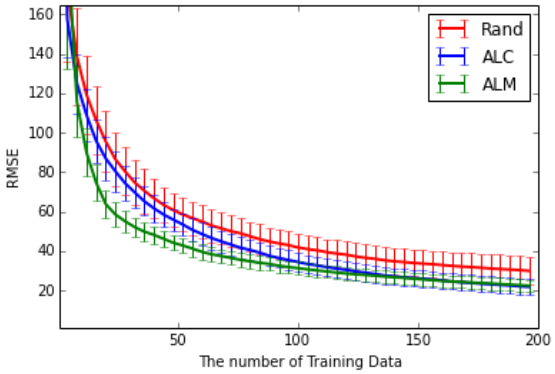


[그림 3] 임의선택, ALM, ALC의 egg holder 함수 예측에 대한 능동 학습 비교. x축은 학습 데이터의 수를 의미하고 y축은 RMSE를 나타낸다. Rand는 임의선택을 의미한다.

[그림 3]의 실험 결과를 보면 ALM과 ALC가 임의선택에 비해 같은 학습 데이터 수에서 더 정확한 예측을 하고 있다. 이 결과는 학습 데이터가 부족한 상황에서 능동 학습이 갖는 강점을 잘 설명한다.

3.3 유기소재 데이터에서의 능동 학습

본 연구의 주된 실험으로서 유기소재 데이터에 대한 가우시안 프로세스 기반의 능동 학습의 효과를 실험하였다.



[그림 4] 임의선택, ALM, ALC의 원자화 에너지 예측에 대한 능동 학습 비교. 형식은 [그림 3]과 동일하다.

GDB-13 데이터에서 1000개의 분자를 선정하였고 이중에 임의의 200개의 분자를 테스트 데이터로 나누었다. 능동 학습을 진행하면서 분자의 원자화 에너지를 예측하여 학습 효과를 알아보았다. 실험 방법 및 결과 도식은 장난감 문제와 동일하다. [그림 4]의 실험 결과를 보면 ALM과 ALC가 임의선택에 비해 같은 학습 데이터 수에서 더 정확한 예측을 하고 있다.

본 실험 결과를 보면 ALM이 ALC보다 더 좋은 성능을 내고 있다. 하지만 데이터에 따라 ALC가 ALM보다 더 좋은 성능을 내는 경우도 있다[5]. ALM을 적용할 때, 분산이 큰 데이터들이 전체 데이터 분포에서 동떨어져 있는 경우 학습 효과가 떨어진다. ALC는 참조 데이터가 얼마나 전체 데이터 분포를 잘 나타내는지에 따라 성능 차이가 많이 날 수 있다.

4. 결론

가우시안 프로세스 회귀모델이 유기소재 속성 예측에 활용될 수 있다는 가능성을 보였고, 이 모델을 기반한 ALM, ALC와 같은 능동 학습을 이용하여 효율적으로 학습 데이터를 모을 수 있다. 또한 이 모델은 역행렬 정리에 따라 새로운 학습 데이터를 쉽게 수용하여 온라인 학습이 용이하다. 기하급수적으로 설계되고 있는 미지의 분자 구조들로부터 학습 효과가 크다고 판단되는 일부로 학습 데이터를 만들고 그 이외의 분자들의 속성은 기계학습적인 방법으로 추측하여 비용을 절감할 수 있다.

감사의 글

이 논문은 삼성전자 종합기술원의 지원을 받아 수행된 연구이며, 정부(미래창조과학부 및 정보통신기술진흥센터)의 정보통신·방송 연구개발사업 지원(R0126-15-1072-SW스타랩, 10035348-mLife, 10044009-HRI.MESSI)과 정부(미래창조과학부 및 한국연구재단)의 지원(NRF-2010-0017734-Videome)을 일부 받았음.

참고문헌

[1] G. Montavon, et al. "Machine learning of molecular electronic properties in chemical compound space." *New Journal of Physics* 15.9 (2013): 095003.

[2] S. Oh, M.S. Lee, and B.T. Zhang. "Ensemble learning with active example selection for imbalanced biomedical data classification." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*8.2 (2011): 316-325.

[3] R.M. Balabin and E.I. Lomakina. "Support vector machine regression (LS-SVM)-an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data." *Physical Chemistry Chemical Physics*13.24 (2011): 11710-11718.

[4] B. Settles. "Active learning literature survey." *University of Wisconsin, Madison* 52.55-66 (2010): 11.

[5] S. Seo, et al. "Gaussian process regression: Active data selection and test point rejection." *Mustererkennung 2000*. Springer Berlin Heidelberg, 2000. 27-34.

[6] C.E. Rasmussen. "Gaussian processes for machine learning." (2006).

[7] B.T. Zhang. "An incremental learning algorithm that optimizes network size and sample size in one trial." *Neural Networks, 1994. IEEE World Congress on Computational Intelligence, 1994 IEEE International Conference on*. Vol. 1. IEEE, 1994.