시뮬레이션 비용을 줄이기 위한 머신러닝을 활용한 결과 및 시간 예측

강민규*, 김지수**, 권 훈*, 이정철***
*제주대학교 컴퓨터공학전공
**제주대학교 전산통계학과
***한국과학기술정보연구원 계산과학플랫폼센터
e-mail:jclee@kisti.re.kr

Predict results and time using Machine Learning to reduce simulation costs

Mingyu Kang*, Ji-su Kim**, Hoon Kwon*, Jeongcheol Lee***
*Dept of Computer Engineering, Jeju National University
**Dept of Computer Science & Statistics, Jeju National University
***Korea Institute of Science and Technology Information

요 약

최근 계산과학 분야에서 시뮬레이션을 활발하게 이용하고 있다. 그리고 IT 기술의 발전에 힘입어 이제는 시뮬레이션을 별도의 복잡한 절차 없이 온라인으로 할 수 있게 되었다. LCAODFTLab은 EDISON 플랫폼에서 제공하는 나노물리 분야의 시뮬레이션 프로그램이다. LCAODFTLab은 온라인 환경에서 사용자에 의해 주어진 원자구조에 대한 전자구조 시뮬레이션을 제공한다. 이처럼 실제로 실행하기 어려운 실험을 간단히 행하는 시뮬레이션은 얼마나 정밀하고 복잡한가에 따라 비용이 발생한다. 이러한 비용은 대체로 사용자에게 부담이 되고, 따라서 시뮬레이션의 수행비용을 줄이는 것은 아주 중요한 요소이다. 따라서 본 논문에서는 EDISON 플랫폼 사용자의 데이터와 머신러닝을 활용하여 시뮬레이션 비용을 줄이는 방법을 소개한다. 이를 통해 사용자는 큰 비용을 지불하지 않고 결과를 예측할 수 있고, 대략적인 수행시간을 알 수 있다.

1. 서론

최근 계산과학 분야는 복잡한 문제를 해결하기 위해 고성능 컴퓨팅 자원이 필요하다. IT 기술의 발전과 함께 계산과학 분야는 이제 더 복잡하고 어려운 문제를 해결할 수 있게 되었다. 그러나 시뮬레이션의 정밀도와 복잡도가 증가함에 따라 시뮬레이션을 수행하는 비용도 많이 증가하고 있다. 시뮬레이션을 가능한 적은 비용으로 처리하는 것은 매우 중요하다. 우리는 EDISON 플랫폼에서 제공하는 시뮬레이션 프로그램인 LCAODFTLab을 가지고 시뮬레이션 비용을 줄이는 방법을 제안한다.

LCAODFTLab은 EDISON 플랫폼에서 제공하는 Linear Combination of Atomic Orbitals 기반 Density Functional Theory 전자구조 계산 소프트웨어이다. SIESTA에 구현된 원자 궤도 함수(Atomic orbital basis sets) 기반 밀도범함수 이론(DFT) 계산을 바탕으로 전자구조계산, 구조 최적화 그리고 분자 동역학 등의 시뮬레이션을 수행할 수 있는 도구이다.

※ “본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음”(2018-0-01863)

SIEASTA는 효율적인 전자구조 계산 및 분자와 고체의 제 1원리 분자 동역학 시뮬레이션을 수행하는 방법 및 프로그램이다. SIESTA의 효율성은 엄격하게 지역화된 기저 함수(basis set)의 사용과 적합한 시스템에 적용할 수 있는 선형 스케일링 알고리즘의 구현에서 비롯된다. 이 프로그램의 매우 중요한 특징은 빠른 탐색 계산에서 평면파 및 모든 전자 방법과 같은 다른 접근법들과 같은 품질의 매우 정확한 시뮬레이션까지, 정확도와 비용을 광범위하게 조정할 수 있다는 것이다. 주요 특징으로는 국소 밀도 근사(LDA)와 일반 구배 근사(GGA)를 포함한다는 것과 의사 포텐셜(pseudo potential)을 사용한다는 것이다. 그리고 공간적으로 유한한 기저 함수를 사용하고 사용하는 기저 함수의 개수와 모양을 조절할 수 있다. 가능한 계산들에는 힘(force), 응력 텐서(stress tensor), 전자밀도(electron density), 구조 최적화(geometry optimization) 등이 있다.[1]

LCAODFTLab은 시뮬레이션을 하기 위하여 반복적이고 복잡한 계산을 수행한다. 따라서 주어진 원자구조에 따라 시뮬레이션 비용이 많이 발생할 수 있다. 여기서 비용이라고 하면 주로 계산자원이나 수행시간을 뜻한다. 이러

한 비용을 줄이기 위한 여러 방법이 있다. 그 중 첫 번째는 시뮬레이션을 수행하는 컴퓨터 하드웨어의 성능을 높이는 것이다. 대표적으로 멀티코어 GPU를 활용하여 시뮬레이션을 병렬로 수행하거나, 다수의 컴퓨터를 네트워크로 연결한 컴퓨터 클러스터를 구축하여 컴퓨터의 성능을 높여 시뮬레이션 수행시간을 줄이는 방법이다. 두 번째는 시뮬레이션을 수행하는 알고리즘을 개선하는 것이다. 알고리즘의 불필요한 계산을 제거하거나 병렬성을 증가시켜 시뮬레이션에 들어가는 비용을 줄일 수 있다. 본 문서에서는 머신러닝을 활용하여 시뮬레이션에 들어가는 비용을 줄이는 방법을 제시한다. 사용자가 시뮬레이션을 수행하기 전에 결과와 수행시간을 예측하는 방법을 사용한다. 기존의 사용자들이 시뮬레이션에 사용한 데이터를 수집하여 모델을 만들고, 머신러닝 알고리즘을 활용하여 예측했다.

2. 시뮬레이션 비용을 줄이기 위한 예측

대부분의 시뮬레이션은 복잡하고 반복적인 계산을 하고, 그에 따라 비용이 발생한다. 그래서 우리는 머신러닝을 통해 사용자가 입력하는 데이터를 가지고 복잡한 계산 과정 없이 결과를 예측해보고자 한다. 머신러닝을 통하여 시뮬레이션의 결과를 예측하려면 충분히 많은 데이터가 있어야 한다. 그러나 무작정 데이터만 있어서는 안 되고 그 데이터를 통해 유의미한 결과를 얻을 수 있어야 한다. 그래서 우리는 EDISON 플랫폼에서 LCAODFTLab을 사용한 사용자들의 데이터를 가지고 결과와 수행시간을 예측했다.

우선, EDISON 플랫폼을 통해 LCAODFTLab을 사용한 사용자들의 데이터를 수집했다. 수집한 데이터는 사용자가 입력한 데이터와 시뮬레이션 결과 데이터이다. 사용자가 입력한 데이터는 여러 가지 파라미터와 분자의 정보이고, 시뮬레이션 결과 데이터는 LCAODFTLab이 수행한 계산 결과이다. 계산 결과는 매우 다양한 정보가 있으므로 거기서 필요한 정보만 수집했다.

수행시간을 예측하기 위해서는, 시뮬레이션의 실패와 성공 여부를 떠나서 시뮬레이션이 얼마 동안 실행되고 있었는지를 알아내야 한다. 실패한 시뮬레이션은 주로 7초 미만의 시간으로 시뮬레이션이 끝났고, 성공한 시뮬레이션은 시간이 다양했다. 원자가 3개 미만인 가벼운 분자들로 이루어진 데이터라서 대체로 걸린 시간이 긴 편은 아니었다. 하지만 길게는 300초까지 걸렸다.

수집한 데이터는 대부분 분자를 이루고 있는 원자의 개수가 3개 미만인 분자들이다. 그 이유는 수집할 분자의 개수가 매우 많아서 선택적으로 수집했다. 그래서 우선 가벼운 분자들부터 수집했다. 약 15000개의 데이터를 수집했다. 그러나 이 데이터가 모두 유효한 데이터는 아니다. 일부분은 다양한 이유로 시뮬레이션에 실패한 데이터였다.

수행시간을 수집하고 분포를 살펴보니, 수행시간이 짧은 쪽에 쏠린 데이터 분포가 나타났다. 이러한 분포는 학습에 도움이 되지 않는다. 이러한 분포를 조정하기 위해, 수행시간에 로그를 취했다. 그 후 분포를 살펴보기 위해서 히스토그램을 그려보니 정규분포와 비슷한 모양을 가지게 되었다. 그리고 우리가 예측하고자 하는 값인 시간과 데이터 간의 상관관계를 살펴보았다. 분자를 이루는 원자의 종류와 개수 그리고 사용자가 입력한 파라미터와 시간에 대해서 상관관계를 구해보았지만, 높은 상관관계를 가지는 특징은 없었다.

수집한 데이터를 바탕으로 학습할 모델에 사용할 특징을 찾아보았다. 가지고 있는 데이터는 시뮬레이션에 사용한 몇 가지 파라미터와 분자를 이루고 있는 각 원자의 종류와 개수이다. 각 원자의 특징이 시뮬레이션에 영향을 미칠 것 같다는 생각으로 분자의 질량, 부피, 그리고 분자를 이루고 있는 총 원자의 수를 특징으로 사용했다. 그리고 상관관계를 통해 결과에 영향을 미칠만한 파라미터도 특징으로 사용했다.

분자를 이루는 원자의 종류와 개수와 파라미터는 시간에 큰 영향을 주지 못했다. 그래서 시뮬레이션을 수행하는 분자의 특성을 이용하여 특징을 추출했다. 사용한 특징은 원자의 공유 반지름과 판데르발스(van der Waals) 반지름이다. 공유 반지름은 공유 결합을 하는 같은 원자 사이의 거리의 절반이다. 그리고 판데르발스 반지름은 공유 결합을 하는 두 원자 간의 거리를 말한다. 공유 반지름과 판데르발스 반지름은 다른 특징들에 비해 유의미한 상관관계를 가졌다.

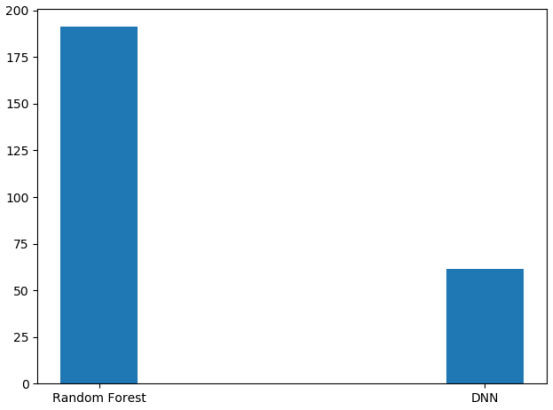
좋은 모델을 만들기 위해서는, 데이터를 가공해야 한다. 첫 번째로 중복데이터를 삭제했다. LCAODFTLab을 사용한 사용자들이 같은 분자를 가지고 파라미터만 조금씩 바꾸면서 시뮬레이션한 데이터들이 많았다. 이런 경우 미세한 조정은 결과에 미치는 영향이 크지 않았다. 그래서 같은 분자로 실행한 시뮬레이션은 같은 데이터로 보고 중복을 삭제했다. 두 번째로 이상치를 제거했다. 이상치는 수집한 데이터의 범위에서 많이 벗어난 아주 작은 값이나 큰 값을 말한다. 이상치는 학습에 방해가 되므로 제거하는 것이 좋다. 그리고 데이터를 모두 정규화했다. 데이터를 정규화하지 않으면 범위가 너무 넓어서 노이즈 데이터가 생성되거나 과적합이 일어날 확률이 높다. 그리고 값이 너무 커서 활성화 함수를 거쳐도 한쪽으로 데이터가 쏠릴 가능성이 있다. 그래서 정규화를 거쳐서 데이터의 범위를

축소하는 과정을 거쳤다.

2-1. Total Energy 예측

분자의 총 에너지(total energy)는 운동 에너지(kinetic energy)와 위치 에너지(potential energy)의 합이다. LCAODFTLab은 분자의 총 에너지 값을 반복적인 계산을 통해 알아낸다. 하지만 이러한 반복적인 계산은 비용이 발생한다. 그래서 우리는 머신러닝을 통해 사용자가 입력하는 파라미터와 분자의 정보를 통하여 반복적인 계산 없이 총 에너지를 예측해보았다.

알고리즘은 여러 머신러닝 알고리즘 중 성능이 좋은 편이라고 알려진 Random Forest와 딥러닝 알고리즘인 Deep Neural Network를 사용했다.[3] 딥러닝과 전통적인 머신러닝에 있어서 가장 큰 차이점은 데이터양에 따른 성능이다. 데이터 양에 따라 어느 알고리즘이 우수한 성능을 낼지 모르기 때문에 두 알고리즘의 정확도를 비교했다. 그리고 과대 적합(Overfitting)을 막기 위해 K-Fold 교차검증(Cross Validation)을 사용하여 예측했다.



(그림1) Random Forest와 DNN의
Total Energy 예측 RMSE 비교

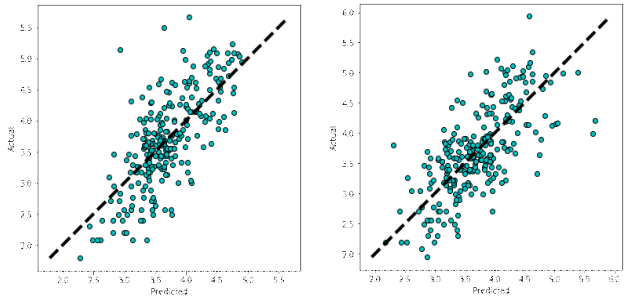
Random Forest를 사용하여 예측해본 결과 RMSE = 195.306, 결정계수는 0.94가 나왔다. 그리고 DNN을 사용하여 예측해본 결과 RMSE = 61.297, 결정계수는 0.99가 나왔다. 딥러닝으로 예측했을 때 더 좋은 결과가 나왔다.

정확도를 표현하는데 RMSE(Root Mean Square Error)와 결정계수(Coefficient of Determination)를 사용했다. RMSE는 오차를 제곱해서 평균한 값의 제곱근을 뜻한다. 추정값 혹은 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용하는 척도이다. 결정계수는 추정한 선형 모형이 주어진 자료에 적합한 정도를 재는 척도이다. 결정계수의 값은 0에서 1사이에 있으며, 종속변인과 독립변인 사이에 상관관계가 높을수록 1에 가까워진다.

2-2. 시간 예측

시뮬레이션이 얼마 동안 실행되는지 시간을 미리 알 수 있으면 불필요한 시간 낭비를 줄일 수 있다. 시뮬레이션의 수행시간은 얼마나 복잡한가에 따라 달라진다. 시뮬레이션이 얼마나 복잡한가는 사용자가 어떤 데이터와 파라미터를 입력했는가에 따라 달라진다. 그래서 우리는 사용자가 입력한 데이터를 통해 시뮬레이션의 수행시간을 예측해보고자 한다. 알고리즘은 마찬가지로 Random Forest와 딥러닝 알고리즘인 Deep Neural Network를 사용했다.

데이터 분석 단계에서 수행시간에 로그를 취한 값을 예측할 값으로 설정했다. 혹은 더 좋은 성능을 위해서 일정 간격으로 시간으로 구분하기도 한다. 하지만 여기서는 로그를 사용한 경우가 더 좋은 데이터 분포를 나타냈기 때문에 로그를 사용했다.



(그림2) Random Forest와 DNN의
시간 예측 결정계수 비교

Random Forest를 사용하여 예측해본 결과 그리 좋은 결과는 나오지 않았다. DNN도 마찬가지로 좋지 않았다. Random Forest로 만든 모델의 결정계수는 0.52이고 DNN으로 만들 모델의 결정계수는 0.42가 나왔다. 총 에너지와 다르게 좋은 결과를 얻지 못했다. 시간을 예측하려면 더 많은 데이터가 있어야 한다. 분자를 이루는 원자의 개수가 3개 미만인 데이터들의 시뮬레이션 시간은 대체로 짧으므로 데이터의 분포가 편향적이다. 따라서, 오래 걸리는 시뮬레이션의 수행시간을 예측하기 위해서는 분자의 개수가 많은 데이터가 더 필요하다.

2-3. Space Group 예측

스페이스 그룹(Space Group)은 물질 내부에 결정구조가 있을 때 결정의 외형에서 찾아볼 수 있는 대칭들의 조합과 격자의 대칭성을 조합해서 얻을 수 있는 대칭 군을 가리킨다. 결정 내의 원자 배열의 대칭인 공간 대칭을 나타내는 데 이용되며, 결정구조를 기술하는 데 필요하다. 우리는 OQMD(Open Quantum Materials Database) 데이터셋에 포함된 분자들의 스페이스 그룹을 예측해보고자 한다.

OQMD 데이터셋은 120,543개의 분자를 포함한다. 그리고 분자의 Crystal system, Final energy, Space group, volume, mass 등 다양한 특성을 포함한다.[2] 그래서 필요한 정보만 추출하여 CSV 파일로 만들었다. 새로 만든 데이터셋은 분자를 이루고 있는 원자와 개수 그리고 Space group만 포함한다.

Space group의 분포를 살펴보면 어떤 Space group은 약 30,000개가 있는 반면에 다른 Space group은 1개 밖에 없는 예도 있다. 한마디로 데이터가 한곳에 집중되어 있다. 이러한 데이터 분포는 모델이 올바르게 학습하는 것을 방해한다. 그래서 우리는 스페이스 그룹이 100개 미만인 데이터들을 모두 제외했다. 그리고 지나치게 많은 Space group에 속해있는 데이터들도 모두 제거했다.

마지막으로, 데이터셋에 있는 중복데이터를 제거했다. 어떤 분자의 Reduced formula가 중복이면, Final energy가 가장 낮은 분자를 남기고 나머지를 삭제했다. Reduced formula가 같아도 분자 구조가 다른 때도 있다. 이런 경우 Final energy가 가장 낮은 경우를 안정한 상태로 보고 안정한 상태의 Space group을 얻었다.

모델의 정확도 평가 척도는 Categorization Accuracy와 평균 F-1 Score를 사용했다. Categorization Accuracy는 모델이 분류를 얼마나 잘 맞췄는지에 대한 확률이다. 하지만 말 그대로 어디까지나 확률일 뿐이다. Categorization Accuracy로 모델의 성능을 평가하려면 데이터의 분포가 균형적이어야 한다. 하지만 데이터가 불균형적이라면 모델의 성능이 좋지 않음에도 불구하고, 좋다고 평가할 수도 있다. 이러한 문제를 해결하기 위해 F-1 Score를 사용한다. F-1 Score를 사용하면 불균형한 데이터에서 만들 모델의 성능을 평가할 수 있다.

DNN을 사용한 모델에서는 Categorization Accuracy = 0.5628, Mean F-1 Score = 0.4447의 결과를 얻었다. 그리고 Random Forest를 사용한 모델에서는 Categorization Accuracy = 0.5081, Mean F-1 Score = 0.4181의 결과를 얻었다.

3. 결론

EDISON 플랫폼의 LCAODFTLab 시뮬레이션 비용을 줄이기 위해서 결과와 수행시간을 예측했다. 총 에너지와 수행시간 그리고 스페이스 그룹을 예측하기 위해 데이터를 수집하고 모델을 만들었다. 기존에 EDISON 플랫폼을 사용한 사용자들의 데이터를 수집하여 모델의 학습과 검증에 사용했다. 사용자들이 입력한 파라미터와 분자의 정보를 가지고 특징을 추출하여 총 에너지와 수행시간 그리고 스페이스 그룹을 예측했다. 결과적으로 총 에너지는 비교적 예측을 잘했으나, 시간과 스페이스 그룹은 잘 예측하지 못했다.

총 에너지에 비해서 시간을 잘 예측하지 못한 이유는 데이터가 부족하다. 수집한 데이터 대부분이 짧은 시간에 시뮬레이션이 끝나서, 오랜 시간이 걸리는 시뮬레이션 수행시간을 예측하지 못했다. 시뮬레이션 비용을 줄일 수 있는 시간 예측을 위해서는 짧은 시간이 아닌 긴 시간이 걸리는 시뮬레이션을 예측할 수 있어야 한다. 왜냐하면, 주로 비용이 크게 발생하는 시뮬레이션은 오래 걸리기 때문이다. 시간이 오래 걸리는 시뮬레이션 데이터를 수집하여 시뮬레이션 비용을 효과적으로 줄일 수 있어야 한다.

스페이스 그룹도 비슷한 맥락이다. 그러나 스페이스 그룹을 잘 예측하기 위해서는 데이터 수집보다는 특징추출에 더 신경 써야 한다. 비교적 많은 데이터를 가지고도 정확도가 높지 않은 것을 보면, 지금 가지고 있는 데이터의 특징이 스페이스 그룹을 잘 구별하지 못하는 것을 알 수 있다.

앞으로 더 해야 할 일은 시뮬레이션 데이터를 더 효과적으로 수집하는 것이다. 사용자로부터 데이터를 수집하는 것은 한계가 있다. LCAODFTLab의 사용자가 많지 않을 뿐만 아니라, 수집한 데이터가 모두 유용하지 않다. 대부분이 중복데이터이다. 그래서 유용한 데이터를 수집하기 위해서는 직접 데이터를 생성해야 한다. 다양한 경우의 시뮬레이션을 만들면서 데이터를 수집하고, 수집한 데이터를 바탕으로 효과적인 모델을 만들어야 한다. 사용자는 이를 바탕으로 시뮬레이션을 수행하면서 들어가는 비용을 줄일 수 있을 것이다.

참고문헌

[1] Emilio Artacho E Anglada, O Diehuez, J D Gale, A Garcia, J Junquera, R M Martin, P Ordejon, J MPruneda, D Sanchez-Portal and J M Soler. "The SIESTA method. Developments and applicability" Journal of Physics: Condensed Matter, Vol.20, No.6, 2008.

[2] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl and Chris Wolverton. "The Open Quantum Materials Database (OQMD): assessing the accuravy of DFT formation energies" npj Computational Materials 1, Article number: 15010, 2015.

[3] Geron, A. "Hands-On Machine Learning with Scikit-Learn and Tensorflow" O'Reilly Media, 2017.