

고비용 최적화를 위한 데이터 기반 진화 알고리즘 소개

김은찬, 최태종*

kie8898@jnu.ac.kr, *ctj17@jnu.ac.kr

Advances in Data-Driven Evolutionary Algorithms: A Research Trend Analysis

Kim Eun Chan and Tae Jong Choi*

Chonnam National Univ.

요약

이 논문에서는 고비용 최적화 문제를 해결하기 위해 주목받고 있는 데이터 기반 진화 알고리즘(Data-Driven Evolutionary Algorithms)에 대해 소개한다. 데이터 기반 진화 알고리즘은 최적화 과정에서 생성된 데이터를 바탕으로 대리 모델(Surrogate Model)을 구축하고, 이를 활용하여 최적화를 수행하는 기법이다. 데이터 기반 진화 알고리즘은 최적화 과정 중 새로운 데이터 확보 여부에 따라 온라인 데이터 기반 진화 알고리즘과 오프라인 데이터 기반 진화 알고리즘으로 나눌 수 있다. 본 논문에서는 이 두 가지 알고리즘을 소개하고, 각각의 한계와 향후 발전 가능성을 논의한다.

I. 서론

진화 알고리즘(Evolutionary Algorithm)이란 자연의 진화 과정을 모티브로 한 최적화 기법이다. 이는 다양한 후보 해(Candidate solutions)를 생성하고 반복적인 진화 과정을 통해 적합도가 높은 후보 해를 선택하여 변형하면서 최적해를 탐색한다. 진화 알고리즘은 공학 설계, 초매개변수 탐색, 금융 투자 전략 등 다양하고 복잡한 문제를 해결할 수 있는 접근법으로 평가받고 있다. 그러나 실세계 최적화 문제에서는 목적 함수와 제약 조건을 직접 계산하기 어렵고, 계산 비용이 높아 최적화 과정에서 사용하기 어려운 경우가 많다. 이러한 상황에서 데이터 기반 진화 알고리즘(Data-Driven Evolutionary Algorithms)이 최근 주목받고 있다.

데이터 기반 진화 알고리즘[1-4]은 수집된 데이터를 활용하여 대리 모델(Surrogate Model)을 구축하고, 이를 통해 최적화 과정을 가속화하거나 비용을 절감하는 방식으로, 계산 자원이 제한된 상황에서도 효율적인 최적화를 가능하게 한다. 이 알고리즘은 고비용의 최적화 함수를 직접 사용하는 대신 대리 모델을 활용해 평가 시간을 단축하고, 선택, 교차, 변이 과정을 반복하면서 최적해를 찾아가는 방식이다. 대규모 데이터와 복잡한 문제 구조를 처리할 수 있어 다양한 산업에서의 활용 가능성이 높으며, 맞춤형 솔루션 제공 측면에서 높이 평가받고 있다.

그러나 데이터 기반 진화 알고리즘에도 몇 가지 한계가 존재한다. 대리 모델의 정확성에 의존하는 특성상, 모델 선택 및 관리에서 발생하는 문제와 제한된 데이터로 인한 성능 저하 등이 도전 과제로 나타난다. 특히 온라인 및 오프라인 데이터 기반 접근 방식에서는 각각의 도전 과제가 존재하며, 이는 알고리즘을 실제로 적용할 때 중요한 고려 사항이다[2].

본 논문에서는 데이터 기반 진화 알고리즘의 두 가지 주요 접근법인 온라인 데이터 기반과 오프라인 데이터 기반 진화 알고리즘에 대해 살펴보고, 이들 각각의 도전 과제를 분석한다. 이를 통해 데이터 기반 진화 알고리즘이 현재 직면하고 있는 문제들을 파악하고, 이를 해결하기 위한 방향성을 논의한다.

II. 데이터 기반 진화 알고리즘의 최신 연구 동향

1. 온라인 데이터 기반 진화 알고리즘

온라인 데이터 기반 진화 알고리즘[2]은 실시간으로 수집되는 데이터를 활용하여 동적인 환경에서 고비용 최적화 문제를 해결하는 알고리즘이다. 이 알고리즘은 진화 알고리즘의 탐색 능력과 실시간 데이터 활용을 결합하여 최적의 솔루션을 도출하는 것을 목표로 한다. 그러나 이 과정에서 아래와 같은 고려사항이 발생한다[1-4].

(1) 대리 모델 선택과 사용

대리 모델이란 복잡한 문제를 머신러닝 또는 딥러닝 등을 통해 근사하는 모델을 말한다[3,4]. 대리 모델의 선택은 최적화 과정에서 중요한 단계로, 다양한 대리 모델이 존재하지만 명확한 지침이 부족해 주로 사용자 경험에 의존하게 된다[2]. 대리 모델이 선택된 후에는 이를 진화 알고리즘에 효과적으로 활용하는 것이 중요하다. 또한, 최적화 문제를 단일 목적 문제로 변환하거나 최적화 과정에서 복잡한 계산을 간소화하고 효율성을 높이는 데 기여할 수 있다[2]. 대리 모델의 성능을 향상시키기 위해 대리 모델 관리 전략을 수립하고, 모델을 지속적으로 업데이트하며, 모델의 불확실성을 평가하여 탐색 과정에서의 위험성을 줄이는 것이 중요하다[2]. 이를 통해 대리 모델이 최적화 과정에서 일관된 성능을 유지할 수 있어야 한다.

표 1 데이터 기반 진화 알고리즘 동작 과정

	동작 과정	설명
1	데이터 전처리	문제 해결에 필요한 데이터를 수집, 데이터 전처리
2	모델 학습	수집된 데이터를 바탕으로 대리 모델을 학습
3	초기 샘플 생성	학습된 정보를 바탕으로 초기 후보 해를 생성
4	적합도 평가	대리 모델을 사용해 각 후보 해의 적합도를 평가
5	선택	적합도가 높은 해를 선택해 다음 세대의 부모로 사용
6	교차 및 변이	선택된 해를 바탕으로 새로운 후보 해를 생성
7	반복	적합도 평가, 선택, 교차 및 변이 과정을 반복하여 최적해를 탐색
8	종료 및 최종해	수렴 기준을 만족하면 알고리즘을 종료하고, 최종 해를 도출

표 2 온라인 및 오프라인 데이터 기반 진화 알고리즘 비교

	분석	온라인 데이터 기반 진화 알고리즘	오프라인 데이터 기반 진화 알고리즘
1	데이터 수집 및 전처리	데이터가 실시간으로 수집되며 처리됨. 이는 동적인 환경에서 실시간 적용에 유리함.	데이터는 사전에 수집되고 고정된 데이터셋을 기반으로 처리됨. 안정적이지만 환경 변화에 대한 반응이 느림.
2	적응성	알고리즘이 환경 변화에 실시간으로 적응할 수 있으며, 새로운 패턴이나 변화를 빠르게 반영 가능.	환경 변화가 반영되지 않는 고정된 데이터로 훈련되므로 새로운 패턴이나 변화에 대한 적응이 느림.
3	효율성	연속적으로 데이터가 입력되기 때문에 높은 연산 자원이 필요.	고정된 데이터셋에 대해 일괄 처리하므로 연산 자원의 사용이 예측 가능하고 관리가 용이함.
4	일관성	실시간 데이터의 변동에 따라 결과가 변동될 수 있어 일관성이 떨어질 수 있음.	고정된 데이터셋으로 처리되므로 결과의 일관성이 높음.
5	실시간 반응성	실시간 분석 및 예측에 매우 유리.	사후 분석이나 배치 처리에 적합함.
6	데이터양	데이터양이 계속 증가할 수 있으며, 처리 성능에 큰 영향을 미침.	고정된 데이터양에 맞추어 처리되므로 데이터양에 대한 관리가 용이함.
7	장점	빠른 적응력과 실시간 대응 능력.	안정적이고 반복이 가능한 결과 제공
8	단점	높은 연산 자원 요구와 잠재적인 불안정성.	환경 변화에 대한 반응이 느림.
9	응용 분야	금융 거래, 실시간 추천 시스템 등	과학 연구, 이미지 인식, 변동이 적은 환경에서 활용

(2) 훈련 데이터 선택과 크기

훈련 데이터의 선택과 크기는 온라인 데이터 기반 알고리즘에서 중요한 단계이다. 대리 모델의 정확도와 진화 알고리즘의 탐색 능력을 향상시키기 위해서는 훈련 데이터를 지속적으로 업데이트하고, 적절한 크기로 유지하는 것이 필수적이다[1,2]. 데이터는 수렴성과 다양성을 고려하여 선택되며, 일반적으로 라틴 하이퍼큐브 샘플링(LHS)[5]과 같은 실험 설계 기법을 사용해 후보 솔루션을 생성한다. 이후, 최적의 후보 해를 선택하여 대리 모델을 재훈련하거나 업데이트한다. 그러나 훈련 데이터가 방대하면 계산 복잡성이 크게 증가하기 때문에, 적절한 데이터 크기를 선택하여 계산 효율성과 모델 정확도 간의 균형을 유지하는 것이 중요하다[2].

2. 오프라인 데이터 기반 진화 알고리즘

오프라인 데이터 기반 진화 알고리즘[2]은 과거에 수집된 데이터를 활용하여 정적인 환경에서 고비용 최적화 문제를 해결하는 알고리즘이다. 그러나 아래와 같은 고려사항이 발생한다[1-4].

(1) 대리 모델 신뢰성 확보

오프라인 데이터 기반 진화 알고리즘에서는 새로운 데이터를 활용해 대리 모델의 품질을 평가할 수 없기 때문에, 모델의 신뢰성을 확보하는 것이 큰 과제이다[1,2]. 이러한 이유로, 오프라인 데이터 기반 진화 알고리즘에서는 기존 데이터에 대한 신중한 분석과 더불어, 모델의 신뢰성을 평가하기 위한 추가적인 검증 절차가 필요하다[2]. 예를 들어, 앙상블 학습을 통해 다양한 유형 또는 동일 유형의 대리 모델을 병렬로 사용함으로써, 개별 모델의 편향이나 불확실성을 보완할 수 있으며, 교차 검증을 통해 모델의 정확도를 추정하고 최적의 모델을 선택하는 방법도 유용하다[2].

(2) 성능 검증

오프라인 데이터 기반 알고리즘의 또 다른 과제는 최적화 과정 중에 알고리즘이 도출한 솔루션의 성능을 검증하기 어렵다는 점이다[1,2]. 이는 실제 최적해를 알 수 없기 때문이다. 일부 연구에서는 알고리즘을 벤치마크 문제를 통해 간접적으로 검증하지만, 이러한 방법은 벤치마크 문제가 실제 문제와 충분히 유사하다는 가정에 기반한다[2]. 그러나 현실적으로는 최적화 문제에 대한 사전 정보가 부족하고, 적절한 벤치마크 문제를 선택하기 어렵기 때문에 알고리즘의 성능 평가가 왜곡될 위험성이 있다[2]. 따라서 문제 유형에 대한 성능을 종합적으로 평가할 수 있는 새로운 검증 방법이 필요하다[2].

III. 결론

본 논문에서는 최적화 문제 해결에 중요한 역할을 하는 데이터 기반 진화 알고리즘을 소개하였다. 온라인 및 오프라인 데이터 기반 접근법 각각의 도전 과제를 분석하고, 이러한 도전 과제들이 알고리즘의 성능에 미치는 영향을 살펴보았다. 온라인 데이터 기반 진화 알고리즘에서는 대리 모델의 선택과 사용, 훈련 데이터의 선택과 크기, 진화 알고리즘의 선택 등이 중요한 과제로 나타났다. 이 과정에서 다양한 머신러닝 기법과 데이터 관리 전략을 도입하여 효율성을 높일 수 있음을 확인할 수 있었다. 반면, 오프라인 데이터 기반 진화 알고리즘은 제한된 데이터로 인해 발생하는 문제들을 해결하는 데 중점을 두고 있다. 데이터 부족, 모델의 신뢰성 확보, 성능 검증의 어려움 등이 주요 과제로 제시되었으며, 이러한 문제들을 해결하기 위한 다양한 접근 방식도 논의되었다.

향후 연구에서는 대리 모델의 성능 향상과 데이터 관리 전략의 최적화에 중점을 두어야 한다. 이를 통해 데이터 기반 진화 알고리즘이 보다 넓은 범위의 현실 문제에 효과적으로 적용될 수 있기를 기대한다.

ACKNOWLEDGMENT

이 결과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2023-00214326, RS-2023-00242528)

참 고 문 헌

- [1] Jin, Yaochu, Handing Wang, and Chaoli Sun. Data-driven evolutionary optimization. Cham, Switzerland: Springer International Publishing, 2021.
- [2] Jin, Yaochu, et al. "Data-driven evolutionary optimization: An overview and case studies." IEEE Transactions on Evolutionary Computation 23.3 (2018): 442-458.
- [3] He, Chunlin, et al. "A review of surrogate-assisted evolutionary algorithms for expensive optimization problems." Expert Systems with Applications 217 (2023): 119495.
- [4] Liang, Jing, et al. "A survey of surrogate-assisted evolutionary algorithms for expensive optimization." Journal of Membrane Computing (2024): 1-20.
- [5] McKay, Michael D., Richard J. Beckman, and William J. Conover. "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code." Technometrics 42.1 (2000): 55-61.