# 데이터 사용 비율에 따른 LHS 기반 샘플링 효율 분석

김동혁, 윤재하, 전창재*
세종대학교

kimdonghyuk@sju.ac.kr, 613jay@sju.ac.kr, *cchun@sejong.ac.kr

# Efficiency Analysis of LHS-Based Sampling Across Different Data Usage Ratios

Donghyuk Kim, Jaeha Yoon, Chang-jae Chun*
Sejong Univ.

## 요 약

본 논문에서는 제한된 데이터 사용 환경에서 라틴 하이퍼큐브 샘플링(LHS) 기반 샘플 선택 전략의 효율성을 분석하였다. 공개 회귀 데이터셋을 대상으로, 기존 데이터 중 일부만을 선택해 학습하는 상황에서 LHS 기반 선택과 무작위 샘플링의 예측 성능을 비교하였다. 물리, 재료, 화학 분야 데이터셋을 사용한 실험 결과, LHS 기반 샘플 선택은 동일한 샘플 수 조건에서 전반적으로 안정적인 성능 개선을 보였으며, 25%의 샘플 사용 비율 구간까지 그 효과가 두드러졌다. 본 연구는 제한된 데이터 예산 환경에서 LHS 기반 접근법의 활용 가능성을 실증적으로 제시한다.

## Ⅰ. 서 론

실험 기반 데이터 수집이 요구되는 공학, 재료 및 제약 분야에서는 제한된 실험 예산 하에서 최대한의 정보를 확보하는 것이 핵심적인 과제로 대두되고 있다. 이러한 문제는 전통적인 실험계획법(Design of Experiments, DoE)의 주요 목적과도 직결되며, DoE는 제한된 실험 횟수 내에서 효율적인 데이터 수집을 가능하게 하는 방법론으로 발전해 왔다. 특히 입력 공간을 균일하게 탐색하기 위한 공간 충진 설계(space-filling design)는 실험 효율성을 높이기 위한 대표적인 접근 방식이다.

라틴 하이퍼큐브 샘플링(Latin Hypercube Sampling, LHS)은 이러한 공간 충진 설계를 대표하는 기법으로, 연속적인 입력 공간을 균일하게 커버할 수 있다는 장점으로 인해 널리 활용되어 왔다. 기존 연구에서는 LHS가 무작위 샘플링 대비 우수한 분포 특성을 가진다는 점이 보고되었으나[1], 대부분 생성된 LHS 샘플을 모두 사용하는 것을 전제로 분석이 이루어졌다. 그러나 실제 실험 환경에서는 모든 실험점을 수행하는 것이 비용 및 시간 측면에서 현실적이지 않은 경우가 많다.

특히 이미 수집된 실험 데이터로부터 일부 샘플만을 선택해 활용하는 상황에서, LHS 샘플을 어느 정도까지 사용하는 것이 충분한 성능을 확보하는지에 대한 체계적인 분석은 상대적으로 부족하다. 기존 연구들은 제한된 실험 예산 하에서의 실험 설계 효율성을 주로 이론적으로 논의해 왔으나[2][3], 샘플 사용 비율 관점에서 LHS의 성능을 정량적으로 비교한 실증 연구는 충분히 이루어지지 않았다. 즉, LHS의 분포적 장점이 실제로 사용되는 샘플 비율 관점에서도 유지되는지에 대한 검증이 필요하다.

본 논문에서는 이러한 DoE 기반 문제의식을 바탕으로, 제한된 데이터 예산 하에서 LHS 기반 샘플 선택의 효율성을 무작위 샘플링과 비교하여 정량적으로 평가한다. 실제 실험 환경을 반영한 공개 회귀 데이터셋을 사용하여, 이미 수집된 데이터 중 일부만을 선택해 학습하는 상황에서 입력 공간 대표성과 예측 성능 간의 관계를 분석한다.

## Ⅱ. 본론

### 2.1 실험 데이터셋 구성

| Dataset | Features | Samples |
|---|---|---|
| Yacht [4] | 6 | 309 |
| Concrete [4] | 8 | 1030 |
| Wine [4] | 11 | 1599 |

표1. 실험에 사용된 데이터셋 요약 (UCI ML repository)

본 연구에서는 연속형 입력–출력 관계를 갖는 공개 회귀 데이터셋을 사용하여 샘플 사용 비율에 따른 예측 성능 변화를 분석하였다. 실험에는 물리(Yacht), 재료(Concrete), 화학(Wine) 분야의 데이터셋을 사용하였으며, 서로 다른 도메인 특성을 통해 샘플 선택 전략의 일반적인 효율성을 비교하고자 하였다. 세 데이터셋은 각각 물리 기반 모델, 재료 배합 문제, 화학 성분 분석 문제를 대표하며, 제한된 데이터 환경에서 LHS 기반 샘플 선택의 효과를 평가하기에 적합하다.

### 2.2 LHS 기반 샘플 선택 방법

본 연구에서 LHS는 신규 실험점을 생성하기 위한 목적이 아니라, 기존 데이터로부터 입력 공간을 균일하게 대표하는 부분집합을 선택하기 위한 기준으로 활용되었다. 정규화된 입력 공간에서 LHS 목표점을 생성한 후, 각 목표점에 가장 가까운 실제 데이터를 선택함으로써 LHS 기반 샘플 집합을 구성하였다.

### 2.3 예측 모델 및 평가 지표

본 연구의 목적은 모델 성능의 극대화가 아니라, 동일한 단순 모델 하에서 샘플 선택 전략에 따른 성능 차이를 비교하는 데 있다. 이에 따라 비선형 고성능 모델 사용을 배제하고, 입력 변수의 2차 항을 포함한 선형 회귀 모델을 사용하였다. 이는 샘플 분포 차이에 따른 효과를 과도한 모델 표현력 없이 관찰하기 위함이다. 성능 평가는 테스트 데이터에 대한 평균 제곱근 오차(RMSE)를 기준으로 수행하였다. 각 샘플 사용 비율 $p$에 대해 무작위 샘플링과 LHS 기반 샘플 선택의 성능을 비교하였으며, 무작위 샘플링의 경우 반복 실험을 통해 평균 성능을 산

출하였다. 표에 제시된 값은 동일한 샘플 비율에서 무작위 샘플링 대비 LHS 기반 선택이 달성한 상대적 성능 개선 정도를 백분율로 나타낸 것이다. 랜덤 대비 LHS 개선율은 다음과 같이 정의된다.

$$\text{개선율}(\%) = \frac{RMSE_{Random}(p) - RMSE_{LHS}(p)}{RMSE_{Random}(p)} \times 100$$

해당 값이 양수인 경우, LHS 기반 샘플 선택이 동일한 샘플 수의 무작위 선택보다 더 낮은 예측 오차를 달성했음을 의미한다.

### 2.4 실험 설정

모든 데이터셋에 대해 전체 데이터를 학습 데이터와 테스트 데이터로 8:2 비율로 분할하였으며, 이 분할은 각 반복 실험마다 새롭게 수행되었다. 이에 따라 학습 데이터와 테스트 데이터는 매 반복마다 서로 다른 구성을 가지며, 특정 분할에 대한 의존을 피하도록 설계하였다. 각 반복 실험에서는 학습 데이터 내에서만 샘플 사용 비율을 조절하여 무작위 샘플링과 LHS 기반 샘플 선택을 수행하였다. 본 연구에서는 각 샘플 비율에 대해 총 100회의 반복 실험을 수행하였으며, 각 반복에서 계산된 RMSE의 평균값을 사용하여 두 샘플 선택 전략의 성능을 비교하였다.

### 2.5 실험 결과 및 분석

| $p$ | Yacht | Concrete | Wine | AVG |
|---|---|---|---|---|
| 5% | 19.51 | 3.12 | 55.45 | 26.03 |
| 10% | 7.33 | 26.38 | 44.37 | 26.03 |
| 15% | 19.59 | 13.62 | 30.03 | 21.08 |
| 20% | 32.8 | 8.25 | 20.02 | 20.35 |
| 25% | 19.37 | 5.94 | 15.45 | 13.59 |
| 30% | 0.74 | 4.13 | 12.03 | 5.64 |
| 35% | 1.58 | 3.18 | 8.78 | 4.51 |
| 40% | 3.38 | 2.88 | 7.01 | 4.42 |
| 45% | 3.42 | 2.21 | 6 | 3.88 |
| 50% | 3.36 | 1.84 | 4.54 | 3.25 |
| 55% | 3.19 | 1.74 | 3.89 | 2.94 |
| 60% | 3.38 | 1.34 | 3.2 | 2.64 |
| 65% | 2.6 | 1.31 | 2.57 | 2.16 |
| 70% | 2.27 | 0.91 | 2.1 | 1.76 |
| 75% | 1.91 | 0.83 | 1.65 | 1.46 |
| 80% | 1.46 | 0.73 | 1.14 | 1.11 |
| 85% | 1.14 | 0.51 | 0.79 | 0.82 |
| 90% | 0.75 | 0.37 | 0.36 | 0.49 |
| 95% | 0.38 | 0.15 | 0.07 | 0.2 |

표2. 샘플 사용 비율에 따른 무작위 샘플링 대비 LHS 기반 샘플 선택의 평균 개선율(%)

표 2는 샘플 사용 비율에 따른 무작위 샘플링 대비 LHS 기반 샘플 선택의 평균 개선율을 나타낸다. 전반적으로 LHS 기반 샘플 선택은 모든 샘플 비율 구간에서 양의 개선율을 보였으며, 특히 25% 구간까지 상대적으로 높은 개선율을 기록하여 제한된 데이터 환경에서 LHS의 효율이 가장 두드러지게 나타났다. 이후 25%를 초과하는 구간에서는 개선율이 점진적으로 감소하는 경향이 관찰되었는데, 이는 샘플 수가 증가함에 따라 무작위 샘플링과의 성능차이가 축소되며 LHS의 상대적 이점이 감소함을 의미한다. 데이터셋별로는 Concrete와 Wine 데이터셋에서 LHS의 효과가 비교적 일관되게 나타났다. 반면 Yacht 데이터셋은 표본 수가 상대적으로 적어 저비율 구간에서 변동성이 크게 나타났으나, 일정 수준 이상의 샘플이 확보된 이후에는 전반적으로 완만한 양의 개선율을 유지하는 경향을 보였다. 한편, 5-10%와 같은 매우 낮은 샘플 비율 구간에서는 데이터셋에 따라 개선율의 편차가 크게 나타났으며, 이는 학습에 필요한 최소 샘플 수가 확보되지 않은 경우에는 샘플 선택 전략보다 데이터 절대량의 영향이 지배적임을 의미한다.

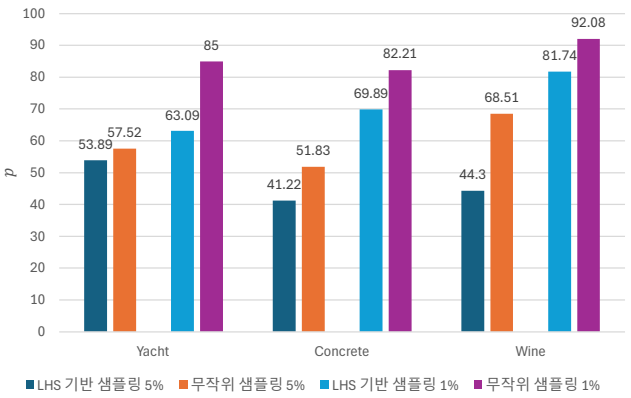


그림 1. 목표 성능 도달에 필요한 학습 데이터 사용 비율

추가적으로, 그림 1은 각 데이터셋에 대해 전체 학습 데이터를 사용했을 때 달성된 최종 RMSE를 기준 성능으로 설정하고, 해당 성능의 5% 및 1% 이내 범위에 도달하기 위해 필요한 학습 데이터 사용 비율을 비교한 결과를 보여준다. 모든 데이터셋에서 LHS 기반 샘플링은 동일 조건의 무작위 샘플링 대비 더 낮은 데이터 사용 비율로 목표 성능에 도달하였다. 특히 5% 기준에서는 Wine 데이터셋에서 약 24%p의 차이가 관찰되었으며, 1% 기준에서는 Yacht 데이터셋에서 약 22%p의 차이가 나타났다. 이는 평균 성능 비교를 넘어, 실제 목표 성능 도달 관점에서도 LHS 기반 샘플 선택이 데이터 효율성 측면에서 유의미한 이점을 가짐을 시사한다.

## Ⅲ. 결론

본 논문에서는 제한된 데이터 사용 환경에서 LHS 기반 샘플 선택 전략의 효율성을 분석하였다. 공개 회귀 데이터셋을 대상으로 샘플 사용 비율에 따른 예측 성능을 비교한 결과, LHS 기반 선택은 동일한 샘플 수의 무작위 샘플링 대비 전반적으로 안정적인 성능 개선을 보였으며, 특히 25%의 샘플 사용 비율 구간까지 효과가 두드러졌다. 이러한 결과는 제한된 데이터 환경에서 LHS가 효율적인 데이터 활용 전략으로 활용될 수 있음을 시사한다.


## ACKNOWLEDGMENT

본 연구는 2025년도 정부(과학기술정보통신부)의 재원으로 국가과학기술연구회 글로벌 TOP 전략연구단 지원사업(No.GTL25101-301)의 지원을 받아 수행되었습니다. 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업 연구 결과로 수행되었습니다. (IITP-2026-RS-2023-00254529).


## 참 고 문 헌


[1] 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.

[2] Joseph, V. Roshan. "Space-filling designs for computer experiments: A review." *Quality Engineering* 28.1 (2016): 28-35.

[3] Shahriari, Bobak, et al. "Taking the human out of the loop: A review of Bayesian optimization." *Proceedings of the IEEE* 104.1 (2015): 148-175.

[4] UCI Machine Learning Repository, Yacht Hydrodynamics Data Set [Link], Concrete Compressive Strength Data Set [Link], and Wine Quality Data Set [Link].