Redis 파라미터 분류 및 단계적 베이지안 최적화를 통한 파라미터 튜닝 연구

동국대학교 통계학과 조성운

SW STOR LAB
Software Technology Advanced Research

과제명: IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발

과제번호: 2017-0-00477

과학기술정보통신부
Ministry of Science and ICT

연세대학교
YONSEI UNIVERSITY

IITP 정보통신기술진흥센터
Institute for Information & communications Technology Promotion

목차

연구 목적 및 이론적 배경

파라미터 튜닝(Parameter Tuning)

- 데이터베이스에서 제공하는 파라미터 값을 조율하여, 최적의 성능을 도출하는 과정

파라미터 튜닝의 어려움

- 전체 시스템의 모든 부분을 제어하기 때문에 각 파라미터의 영향을 고려하는 것은 어렵다.
- 파라미터끼리 독립적이지 않고, 연관성이 존재할 수 있어 종속성을 고려해야한다.

→ **사용자가 직접 값을 설정하는 것은 현실적으로 불가능하다.**

연구 목적 및 이론적 배경

BO(Bayesian Opimization)

- 베이지안 이론을 기반으로 사전 데이터를 반영하여, 목적함수를 최적화하는 기법.
- 입력값을 받는 미지의 목적함수를 상정하여, 함수값을 최대, 최소로 만드는 최적의 입력값 집합을 찾는다.

Surrogate Model
f

현재까지 탐색된 데이터로 목적 함수 추정

argmax Acquisition function

다음 단계에 입력할 최적의 데이터를 추천

사전 데이터를 반영하여 목적함수 값을 최대로 하는 최적의 파라미터 집합을 얻는다.

연구 목적 및 이론적 배경

파라미터 개수에 비례하여 차원이 증가한다.

- 탐색 공간이 증가하여, 공간 복잡도와 시간 복잡도가 높아진다.
- 과적합 문제가 발생하여, 필요한 데이터 셋 양이 증가한다.
- Surrogate Model과 목적함수의 예측 정확도가 떨어진다.

연구 목적 및 이론적 배경

PBO(Phased Bayesian Optimization)

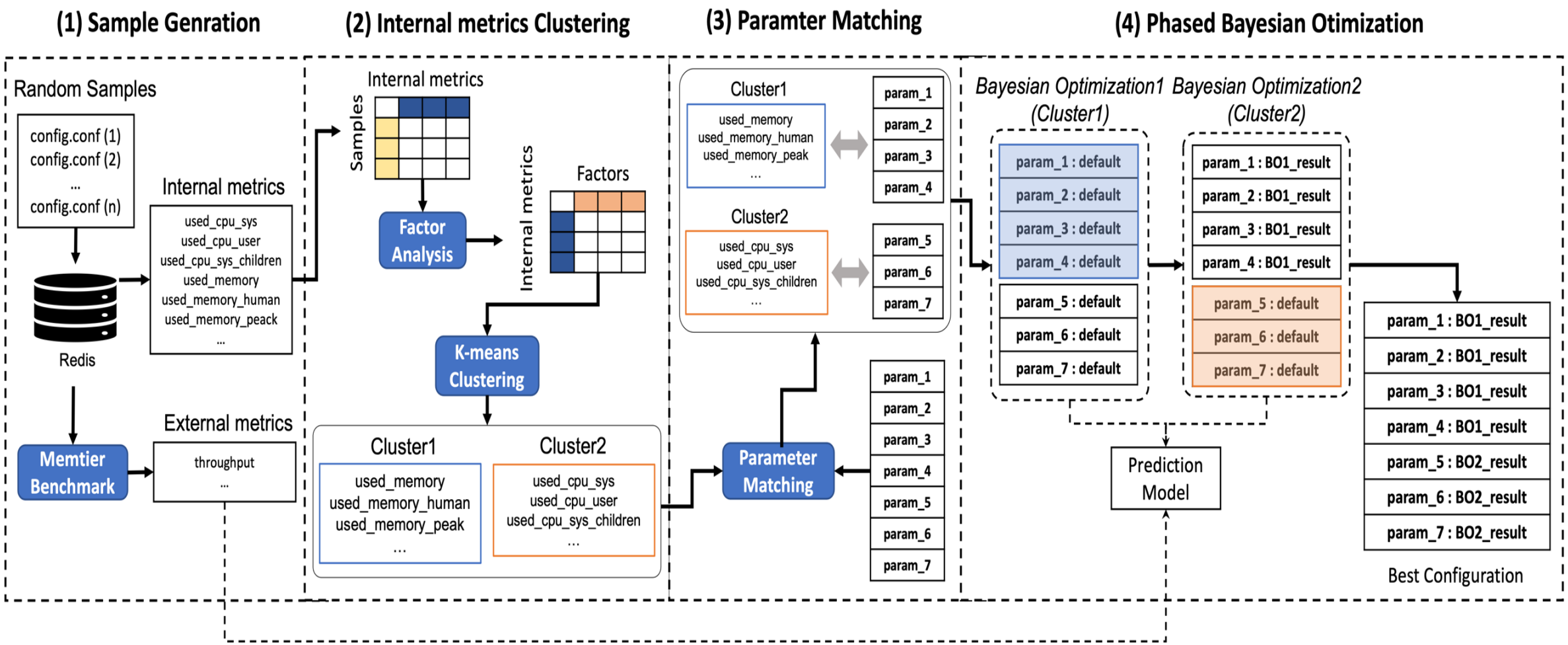
- 통계적 기법과 기계학습을 통해 파라미터를 분류한 후 단계적으로 BO 진행

→ 통계적 기법과 BO를 사용해 탐색 공간을 줄이며, 시간, 공간 복잡도를 낮춘다.

→ 과적합 문제를 방지하고, 학습 시 필요한 데이터 셋 양을 줄인다.

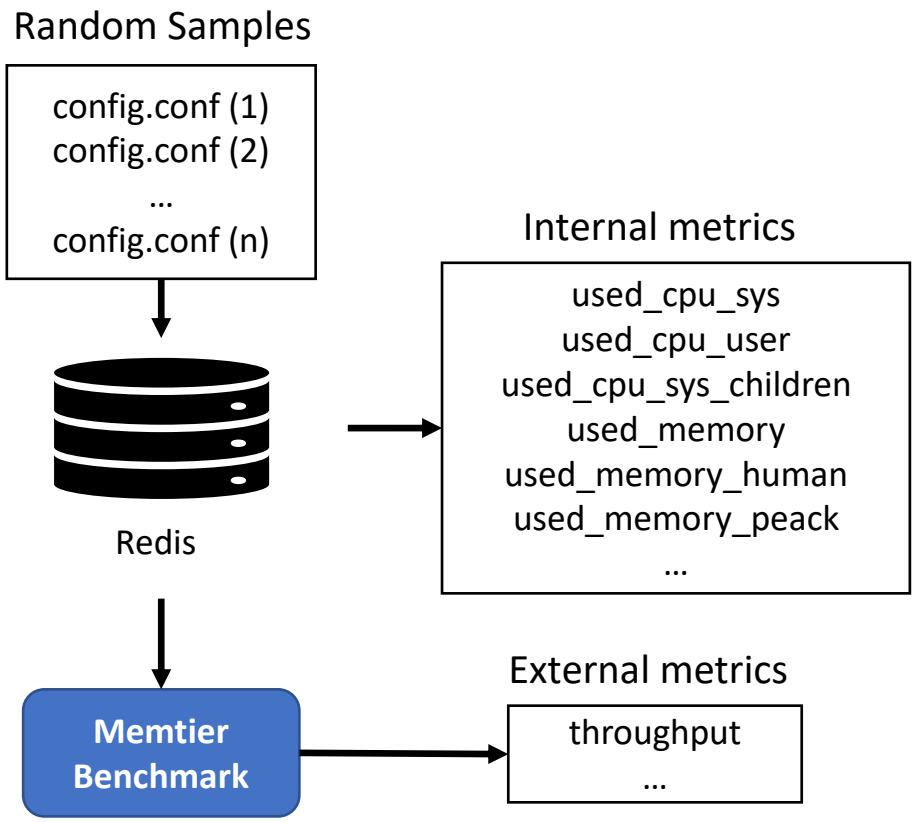
→ Surrogate Model과 목적함수의 예측 정확도를 높인다.

제안하는 모델


(1) Sample Genration
(2) Internal metrics Clustering
(3) Paramter Matching
(4) Phased Bayesian Otimization
Random Samples
config.conf (1)
config.conf (2)
...
config.conf (n)
Redis
Internal metrics
used_cpu_sys
used_cpu_user
used_cpu_sys_children
used_memory
used_memory_human
used_memory_peack
...
Memtier Benchmark
External metrics
throughput
...
Internal metrics
Samples
Factor Analysis
Internal metrics
Factors
K-means Clustering
Cluster1
used_memory
used_memory_human
used_memory_peak
...
Cluster2
used_cpu_sys
used_cpu_user
used_cpu_sys_children
...
Parameter Matching
Cluster1
used_memory
used_memory_human
used_memory_peak
...
Cluster2
used_cpu_sys
used_cpu_user
used_cpu_sys_children
...
param_1
param_2
param_3
param_4
param_5
param_6
param_7
Bayesian Optimization1 (Cluster1)
Bayesian Optimization2 (Cluster2)
param_1 : default
param_2 : default
param_3 : default
param_4 : default
param_5 : default
param_6 : default
param_7 : default
param_1 : BO1_result
param_2 : BO1_result
param_3 : BO1_result
param_4 : BO1_result
param_5 : default
param_6 : default
param_7 : default
Prediction Model
param_1 : BO1_result
param_2 : BO1_result
param_3 : BO1_result
param_4 : BO1_result
param_5 : BO2_result
param_6 : BO2_result
param_7 : BO2_result
Best Configuration

제안하는 모델

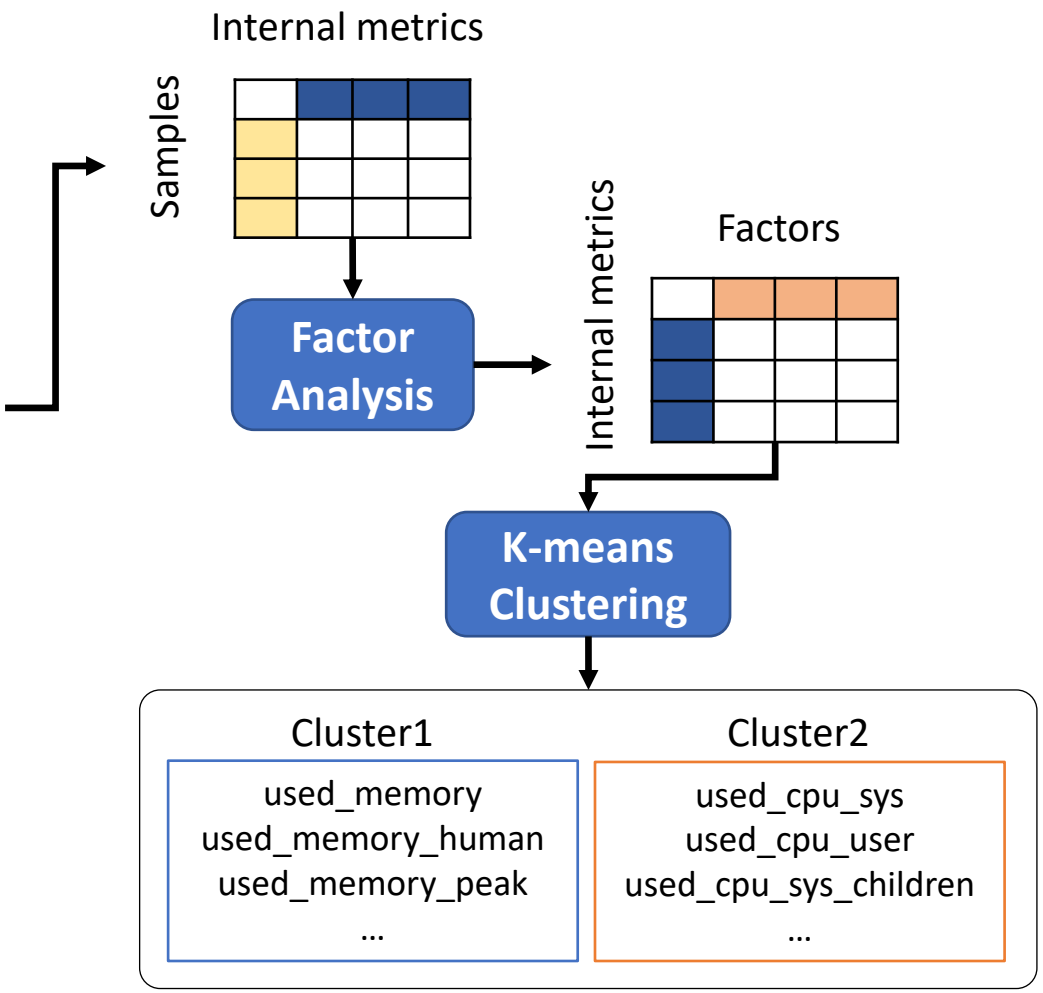
(1) Sample Generation



- 파라미터들에 랜덤 값 할당 후 Redis Configuration 파일 생성
- Info 명령어를 통해 Internal metrics 추출
- Memtier Benchmark를 통해 External metircs 추출

제안하는 모델

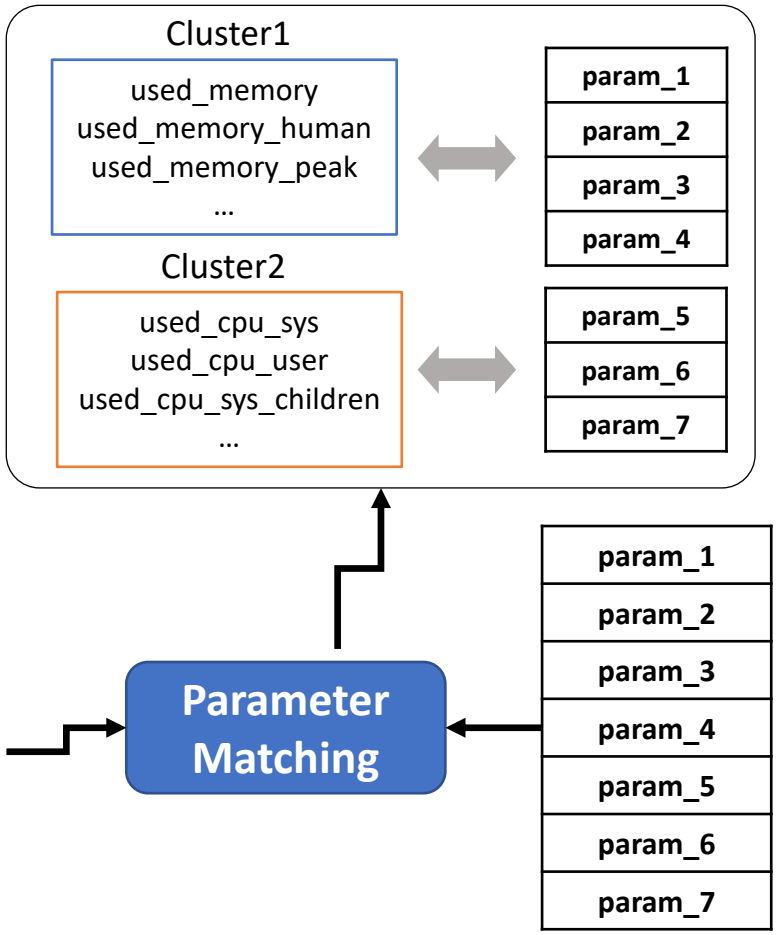
(2) Internal metrics Clustering



- 요인 분석을 진행해 내부 metrics 간의 공통 요인을 파악
- K 평균 군집화를 통해 내부 metrics 분류

제안하는 모델

(3) Paramter Matching



$$\gamma_{IMK} = \frac{cov(Param, IMK)}{\sigma_{param}\sigma_{IMK}}$$

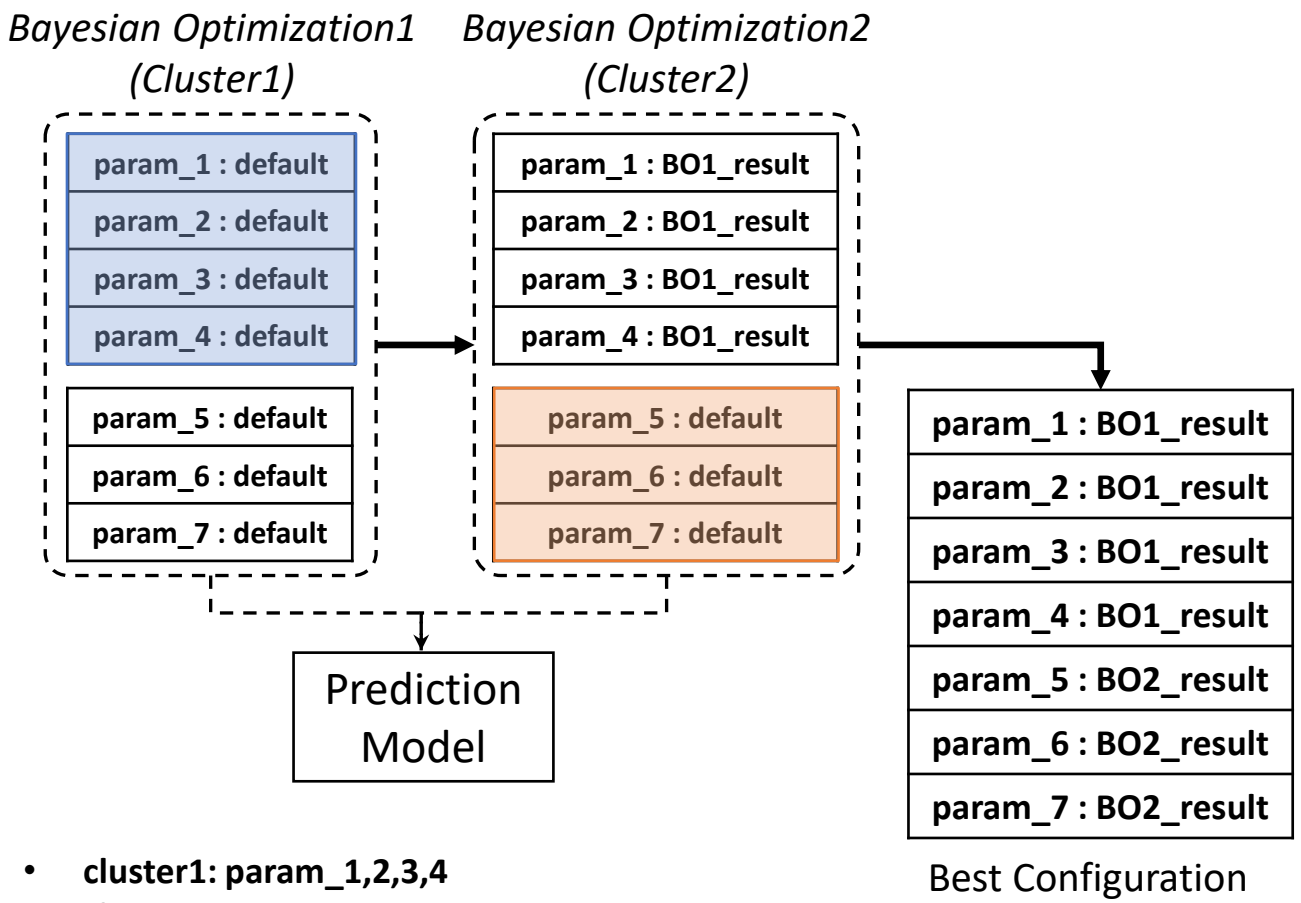
파라미터와 내부 metrics 간의 상관계수를 계산

$$\frac{|\gamma_{IM1}| + |\gamma_{IM2}| + \cdots + |\gamma_{IMN}|}{Total\ number\ of\ Internal\ Metrics}$$

군집과 파라미터와의 연관성 계산

제안하는 모델

(4) Phased Bayesian Optimization



- **cluster1: param_1,2,3,4**
- **cluster2: param_5,6,7**

- 파라미터 군집별로 BO를 단계적으로 진행한다.
- 진행중인 BO의 군집 대상에 포함되지 않은 경우 고정된 채 진행된다.

실험 및 결과 분석

실험 환경

OS	CentOS Linux release 7.6.1810 (Core)
CPU	Intel® Core™ i7-6700K CPU @ 4.00GHz
RAM	16384 MB
Redis Version	6.2.1

실험 및 결과 분석

비교 실험 진행

- PBO 를 통해 진행한 결과, 분류되지 않은 파라미터로 BO 를 진행한 결과, Redis의 default 성능을 단위 시간당 처리량 값을 기준으로 비교

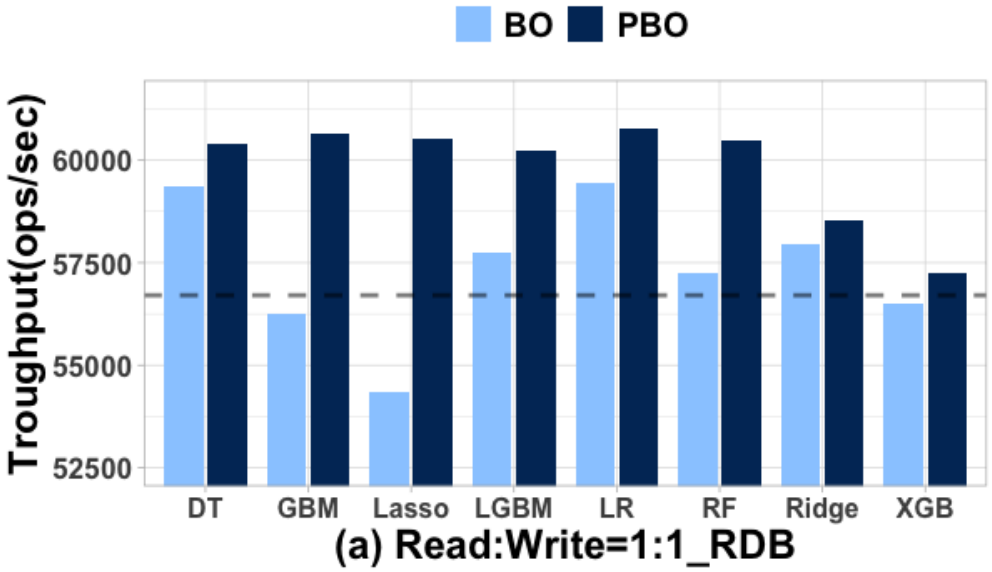
- Redis의 지속성 기법 RDB, AOF 방식별로 두 가지 워크로드 Read-Write(1:1), Write-Only 구분하여 진행

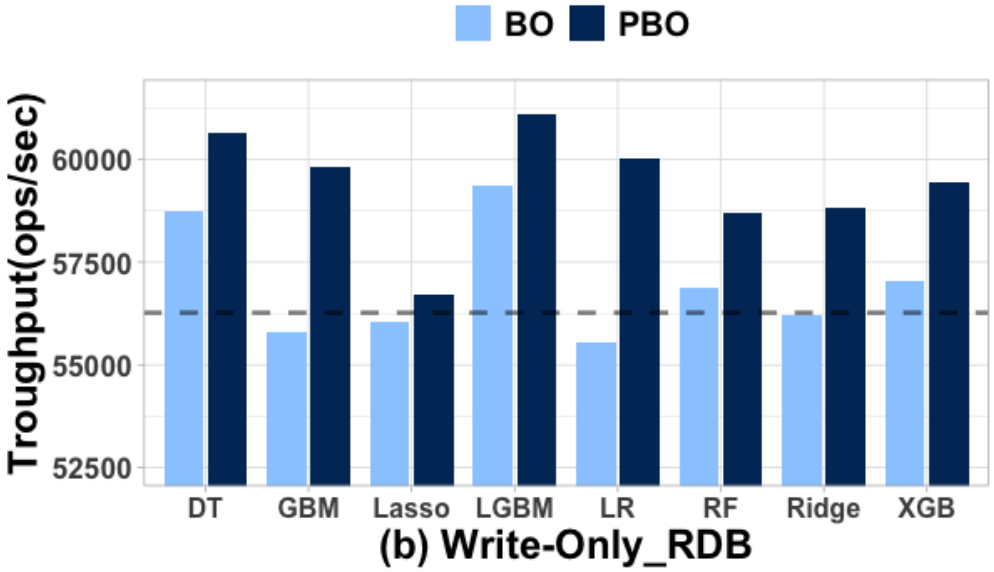
- BO의 예측 모델로 DT(Decision Tree), GBM(Gradient Boosting Machine), Lasso, LGBM(Light GBM), LR(Linear Regression), RF(Random Forest), Ridge, XGB(XGBoost) 8가지 회귀 모델로 성능 평가

- 지속성 기법별로 워크로드의 평균 값을 계산해 회귀 모델을 비교

실험 및 결과 분석

RDB 비교 결과 (PBO, BO, default)

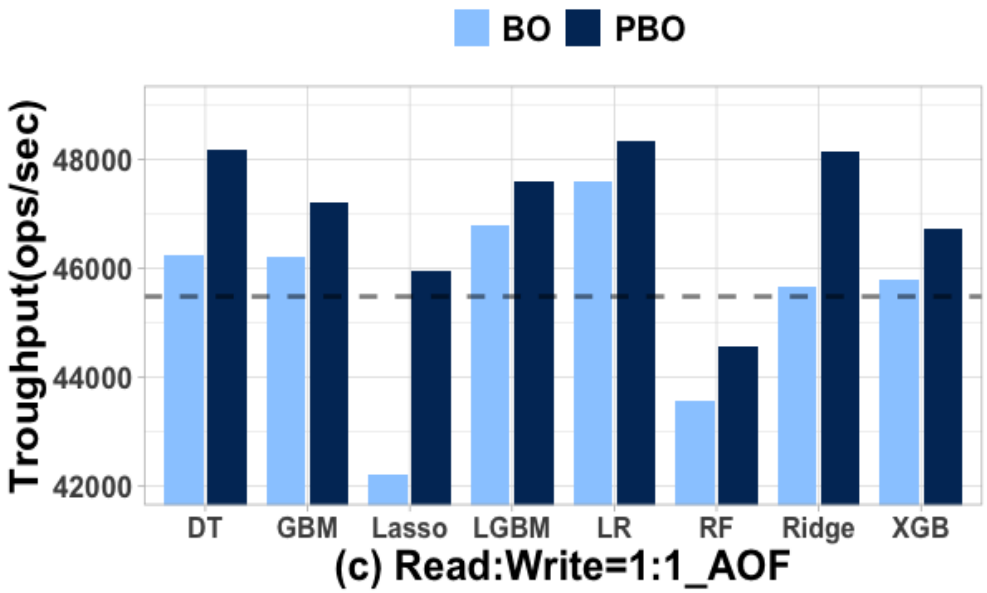


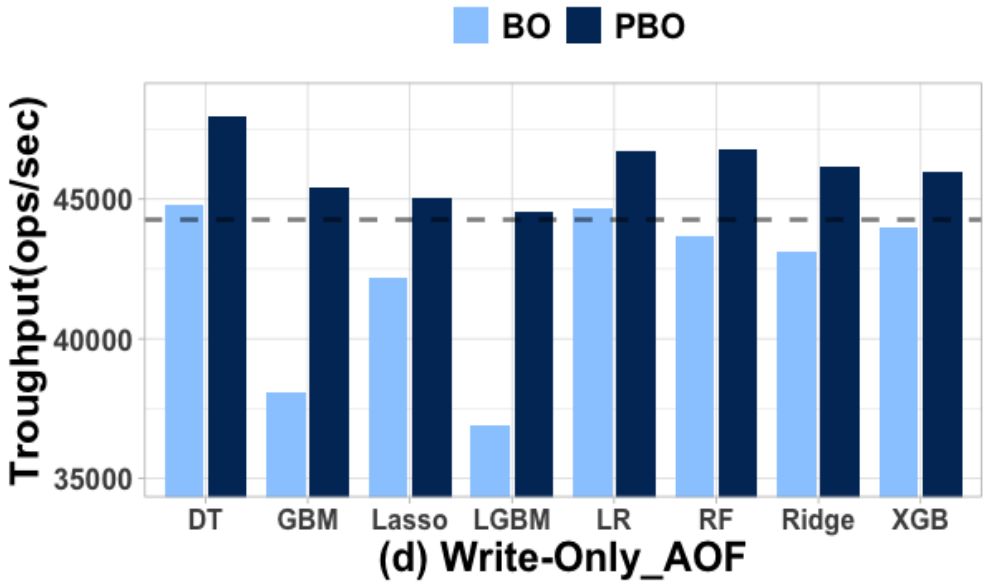


PBO가 BO와 default 성능 보다 모두 높은 성능을 보였다.

실험 및 결과 분석

AOF 비교 결과 (PBO, BO, default)





PBO가 Read:Write 워크로드에서 RF를 제외한 모든 경우에서 BO와 default 성능보다 높은 성능을 보였다.

실험 및 결과 분석

RDB, AOF 비교 결과 (Read-Write(1:1), Write-Only 평균)

RDB

	Read-Write(1:1)	Write-Only	Average
DT	60401	60648	60524.5
GBM	60629	59807	60218
Lasso	60538	56719	58628.5
LGBM	60218	61105	**60661.5**
LR	60766	60036	60401
RF	60467	58705	59586
Ridge	58515	58831	58673
XGB	57254	59432	58343

AOF

	Read-Write(1:1)	Write-Only	Average
DT	48181	47974	**48077.5**
GBM	47223	45391	46307
Lasso	45938	45073	45505.5
LGBM	47599	44520	46059.5
LR	48321	46739	47530
RF	44560	46795	45677.5
Ridge	48138	46167	47152.5
XGB	46738	45988	46363

RDB에서는 LGBM, AOF에서는 DT 에서 가장 높은 성능을 보였다.

실험 및 결과 분석

실험 결과

- PBO 방법이, 분류되지 않은 파라미터로 BO를 진행한 결과와 Redis의 default 성능보다 높았다. (AOF의 Read-write(1:1)에서 RF 제외)

- 모델 비교 결과 RDB에서는 LGBM, AOF에서는 DT 모델에서 가장 높은 성능을 보였다.

결론

- PBO 방식이 분류하지 않고 BO를 진행한 경우와, default 설정값보다 높은 성능을 보인다.
- 8가지 회귀 모델 중 LGBM과 DT에서 가장 높은 성능이 나타났다.

참고문헌

- https://redis.io/
- Van Aken, Dana, et al. "Automatic database management system tuning through large-scale machine learning.", Proceedings of the 2017 ACM International Conference on Management of Data, 2017, pp.1009-1024
- Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. "Practical bayesian optimization of machine learning algorithms.", Advances in neural information processing systems, 2012, 25
- Reunanen, Juha. "Overfitting in making comparisons between variable selection methods.", Journal of Machine Learning Research 3, 1371-1382, 2003
- Yong-Lak Choi, Byungkwon Yoon, and Kiwon Chong. "Database Management System Parameter Tuning Processes for Improving Database System Performance.", The Journal of Korean Institute of CALS/EC, vol. 7, no. 1, pp. 107-127, 2002
- Juyeon Seo, Jieun Lee, et al. "A Study on Redis Parameter Tuning Based on Non-linear Machine Learning.", The Korean Institute of Information Scientists and Engineers, 2021, 69-71
- Alabed, Sami, and Eiko Yoneki. "High-Dimensional Bayesian Optimization with Multi-Task Learning for RocksDB.", Proceedings of the 1st Workshop on Machine Learning andn Systems, 2021, pp. 111-119
- Memtier-Benchmark. https://github.com/RedisLabs/memtier_benchmark
- Yong, An Gie, and Sean Pearce. "A beginner’s guide to factor analysis: Focusing on exploratory factor analysis.", Tutorials in quantitative methods for psychology 9.2, 2013, 79-94
- Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm.", Pattern recognition 36.2, 2003, 451-461

감사합니다