DPPML : 메타러닝을 활용한 데이터베이스 성능 예측*

염찬호 ○ 이지은 박상현†

연세대학교 컴퓨터과학과

{chanho0475, jieun199624, sanghyun}@yonsei.ac.kr

DPPML : Database Performance Prediction with Meta-Learning

Chanho Yeom○ Jieun Lee Sanghyun Park†

Dept. of Computer Science, Yonsei University

요 약

데이터베이스의 성능을 높이기 위해서는 데이터베이스에 존재하는 다양한 Knob을 조정해야 한다. Knob의 적절한 값을 찾아가는 과정을 데이터베이스 튜닝이라고 하고, 이를 위해서는 빠르게 성능을 예측할 수 있는 신뢰도가 높은 모델이 요구된다. 하지만 성능을 예측하는 모델을 학습시키는 데에는 많은 양의 데이터가 필요하다. 어느 워크로드 환경의 데이터베이스인지에 따라 데이터 패턴이 다양하기 때문에 기존에 학습한 모델일지라도 워크로드 환경이 다르면 모델을 새로 학습해야 하는 경우도 있다. 따라서 본 논문에서는 메타러닝을 적용하여 기존의 데이터셋으로부터 학습한 모델이 새로운 데이터셋에 대해서도 적은 샘플로 빠르게 학습 가능한 DPPML을 제안한다. 실험을 통해 새로운 데이터 셋에 대해 일반적인 사전학습 모델과 제안하는 모델간의 학습 수렴 속도와 예측 정확도를 비교하였고 다양한 데이터 샘플 개수에 대해 실험하여 모델의 우수성을 검증하였다.

1. 서 론

기술이 발전함에 따라 생산 및 활용되는 데이터의 양이 폭발적으로 증가하였고 그 빅데이터들을 활용해 사람들의 일상에 도움을 주는 스마트 도시화가 진행되고 있다[1,2]. 빅데이터를 관리하는 데이터베이스에 대한 필요성이 증가하였는데, 이런 빅데이터 환경에서 생성되는 비정형 데이터를 효율적으로 처리할 수 있는 데이터베이스가 사용되고 있다[3,4,5].

데이터베이스는 사용자가 설정할 수 있는 다양한 파라미터들이 존재하고 이를 Knob이라 지칭한다. 이 Knob들의 설정에 따라서 데이터베이스의 성능이 상이한데, 전통적으로는 DataBase Administration(DBA)가 원하는 성능을 낼 수 있도록 직접 Knob들을 설정하였다. 하지만 시간이 흐르면서 데이터베이스에 존재하는 Knob의 개수가 계속해서 증가하며[6] DBA가 튜닝하기에는 한계가 있다.

이러한 한계를 해결하기 위해서 데이터베이스 튜닝을 자동으로 추천해주는 여러 연구가 진행됐으며 데이터베이스 튜닝을 할 때 데이터베이스 성능을 예측하는 모델이 사용된다[6,7,8]. 예측 모델의 정확도가 떨어지면 데이터베이스 튜닝의 성능에 영향을 주기 때문에 정확도 높은 예측 모델이 필요하다. 또한, 예측 모델을 학습하기 위해서는 많은 양의 데이터가 필요하다. 데이터베이스를 사용하는 워크로드(Workload) 환경이 매우 다양하기 때문에, 매번 예측 모델을 학습하기 위해 많은 양의 학습 데이터를 준비하기에는 시간과 비용적인 한계가 따른다.

따라서, 본 논문은 새로운 워크로드 환경에 대해서 적은 샘플로도 빠르게 학습이 가능한 데이터베이스 성능 예측 모델 DPPML을 제안하고자 한다. DPPML은 메타러닝[9]을 적용하여 기존의 데이터셋으로부터 새로운 워크로드 환경에 빠르게 수렴할 수 있는 가중치를 찾는다.

본 논문은 일반적인 사전학습 모델과 제안하는 방법론 간의 새로운 워크로드 데이터에 대한 수렴 속도 및 다양한 데이터 샘플 개수에 대한 비교 실험을 진행하였다.

2. 모델 구조

본 논문은 메타러닝을 적용함으로써, 새로운 워크로드 환경에 대해서 적은 데이터셋으로도 빠르게 학습하여 데이터베이스의 성능을 예측할 수 있는 모델 DPPML을 제안한다. DPPML은 Knob과 워크로드 정보가 주어졌을 때 데이터베이스의 데이터 처리량(Rate)를 예측하는 것을 목표로 한다. 그

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2017-0-00477, (SW 스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발)과 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행된 연구임.

† 교신 저자: sanghyun@yonsei.ac.kr

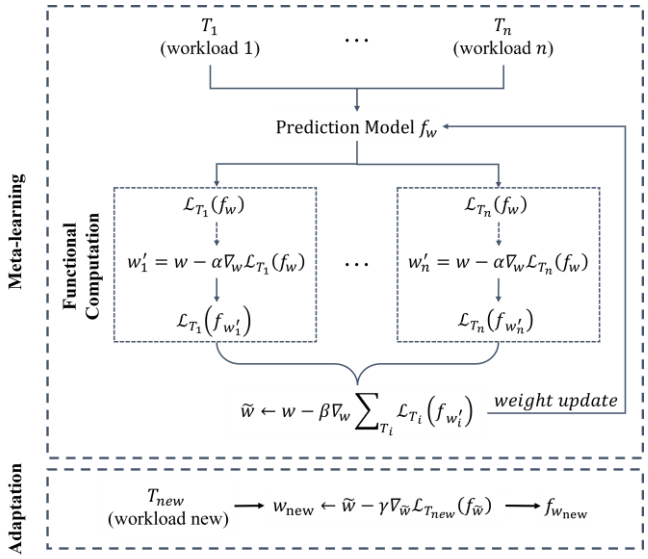

그림 1. DPPML 흐름도

리고 기존의 데이터셋을 통해 새로운 데이터에 빠르게 수렴할 수 있도록 학습하는 Meta-learning 과정과 실제로 새로운 데이터에 대해 성능을 확인해보는 Adaptation과정으로 이루어져 있으며 그 과정은 그림 1과 같다.

2.1 Meta-learning

Meta-learning 단계는 수식 (1)과 같이 기존의 n 개의 워크로드 데이터에 대한 경사하강법을 통해 수정된 가중치 w_i를 총 n개 구한다. α는 학습률(learning rate)이고, f_w는 가중치 w를 가지는 예측 모델, 그리고 i번째 워크로드 T_i의 데이터에 대한 f_w의 손실함수를 $\mathcal{L}_{T_j}(f_w)$라고 한다.

$$w_i = w - \alpha\nabla_w\mathcal{L}_{T_i}(f_w) \quad (1)$$

그 뒤 수식 (2)에 나타난 것처럼 수식(1)로 계산된 w_i에 대한 손실함수값의 총합으로 모델의 가중치 w를 업데이트 한다. β는 학습률이다.

$$\tilde{w} \leftarrow w - \beta\nabla_w \sum_{i=1}^{n} \mathcal{L}_{T_i}(f_{w_i}) \quad (2)$$

인공신경망 구조와 손실함수는 OANet[10]의 네트워크 구조와 손실함수를 사용하였다. 수식 (1)의 과정에서는 각 워크로드 데이터의 특징을 반영한 가중치 파라미터 w_i를 구한다. 그리고 w_i들을 활용해 수식 (2)에서 w를 업데이트 하기 때문에, 모델의 가중치를 각각의 워크로드에 대해 균형 잡힌 지점으로 업데이트 할 수 있다. Meta-learning 과정을 마친 모델은 Adaptation 과정에서 새로운 워크로드 데이터가 주어졌을 때 빠르게 수렴할 수 있다.

2.2 Adaptation

Adaptation과정에서는 Meta-learning과정에서 학습을 마친 모델 $f_{\tilde{w}}$을 새로운 워크로드 T_{new} 데이터로 수식 (3)과 같이 학습을 한다. γ는 학습률이다.

$$w_{new} \leftarrow \tilde{w} - \gamma\nabla_{\tilde{w}}\mathcal{L}_{T_{new}}(f_{\tilde{w}}) \quad (3)$$

Adaptation과정을 거친 모델인 $f_{w_{new}}$는 새로운 워크로드 환경에 맞게 데이터베이스 성능을 예측할 수 있다.

3. 실험 및 결과

3.1 실험 환경

본 논문에서는 Read-Write의 비율이 9:1, 1:1, 1:9 그리고 Update에 대해 value-size를 각기 다르게 한 16개의 RocksDB 워크로드 환경에서 db_bench[11]로 성능을 측정한 데이터로 실험을 진행했다. Meta-learning과 Adaptation에서 사용한 워크로드 환경 데이터셋은 표 1과 같다.

표 1. 워크로드 환경에 따른 데이터셋 구성

Meta-learning dataset			
RW 9:1 (1)	RW 1:1 (1)	RW 1:9 (1)	Update (1)
RW 9:1 (4)	RW 1:1 (4)	RW 1:9 (4)	Update (4)
RW 9:1 (64)	RW 1:1 (64)	RW 1:9 (64)	Update (64)
Adaptation dataset			
RW 9:1 (16)	RW 1:1 (16)	RW 1:9 (16)	Update (16)

3.2 실험 결과

그림 2는 Adaptation 과정에서 DPPML과 일반적인 사전 학습(Pretrained) 모델의 수렴 양상을 그래프로 표현한 것이

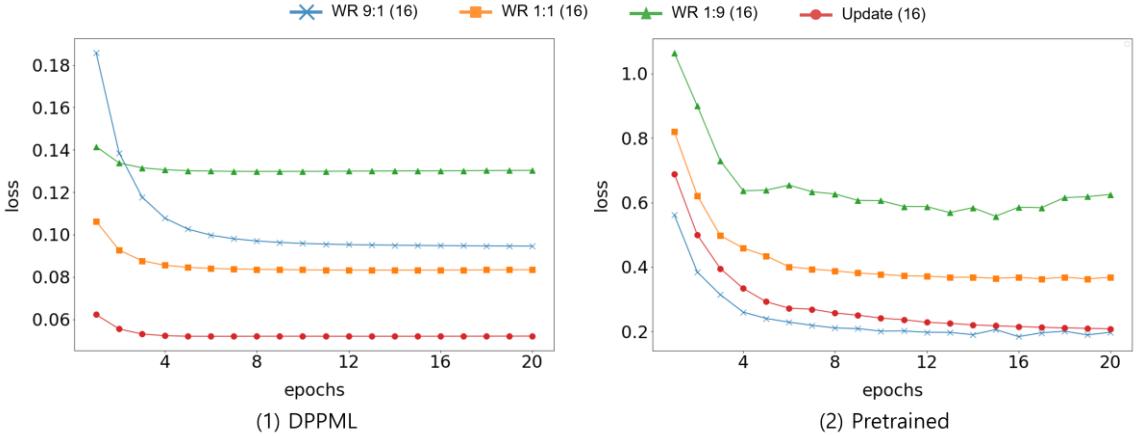

그림 2. Adaptation과정에서 DPPML과 Pretrained모델의 epoch당 loss(손실함수 값) 그래프

표 2. 데이터 샘플 개수에 따른 R2 score(↑) 비교

Model	# of Samples	RW 9:1 (16)	RW 1:1 (16)	RW 1:9 (16)	Update (16)
DPPML	10	0.6646	0.8598	0.8406	0.8995
	40	0.6970	0.8676	0.8734	0.9148
	70	**0.7103**	**0.8709**	0.8413	0.9120
	100	0.7025	0.8703	**0.8759**	**0.9151**
Pretrained	10	0.4698	0.4363	0.5746	0.6681
	40	0.4478	0.5877	0.5967	0.6046
	70	0.4711	0.6587	0.6674	0.6886
	100	**0.4981**	**0.6980**	**0.7382**	**0.7560**

다. [RW 9:1 (16)], [RW 1:1 (16)], [RW 1:9 (16)], [Update (16)] 4개의 워크로드에 대해 각각 실험을 진행하였다. 그림 2에서 DPPML은 평균 5 에폭에 수렴하고, Pretrained 모델은 평균 11 에폭에서 수렴하는 것을 확인할 수 있다. 이를 통해 DPPML이 더 빠르게 수렴하는 것을 검증하였다. 또한, loss 값을 비교해보았을 때 DPPML은 0.2 이하의 값을 가지고, Pretrained 모델은 0.2 이상의 값을 가진다. 이는 DPPML이 Meta-learning과정에서 워크로드 데이터셋들에 대해 균형 잡힌 가중치를 학습했기 때문에 새로운 워크로드 데이터셋에 대해서도 학습 정확도가 높은 것을 알 수 있다.

표 2는 Adaptation과정에서 사용한 데이터 샘플 개수를 바꿔가며 R2 score를 실험해본 결과이다. 모든 샘플 개수에 수에 대해 DPPML이 Pretrained 모델보다 R2 score가 높은 것을 확인할 수 있다. Pretrained 모델의 경우 샘플의 개수가 많아지면 성능이 좋아지는 경향을 확인할 수 있다. DPPML 또한 샘플의 개수가 많아지면 성능이 좋아지지만 샘플이 10개일 때도 100개와 비교했을 때 비슷한 수치를 보인다. DPPML은 Meta-learning과정에서 새로운 데이터에 대해서도 빠르게 수렴할 수 있는 가중치를 학습했다. 그렇기 때문에 적은 샘플 개수로도 수렴이 가능하게 되고 10개의 샘플로 실험한 DPPML의 성능이 100개의 샘플로 실험한 Pretrained 모델보다 월등히 좋은 성능을 낼 수 있다는 것을 알 수 있다.

4. 결 론

데이터베이스의 Knob 튜닝을 할 때 주어진 워크로드 환경이 매우 다양하여 그 때마다 학습에 필요한 데이터를 생성하는데 많은 시간과 비용이 든다는 문제점이 있다. 본 논문에서는 메타러닝을 적용하여 새로운 환경에서 사용하여야 할 경우에 적은 샘플의 데이터만으로도 빠르게 수렴하여 데이터 베이스의 성능을 예측할 수 있는 모델을 제안하였다. 그리고 성능을 확인하기 위해 일반적인 Pretrained 모델과 DPPML의 Adaptation 과정에서 loss의 크기와 수렴하는 속도를 비교하였다. 추가적으로 Adaptation dataset에서 사용하는 데이터 샘플의 개수를 바꿔가며 R2 score로 성능을 비교하여 제안하는 모델의 성능이 우수함을 검증하였다.

향후에는 DPPML 모델을 활용하여 데이터베이스 튜닝에 관한 연구를 진행할 예정이다.

참 고 문 헌

[1] Mama Nsangou Mouchili, et al., “Smart City Data Analysis”, Proceeding of the First International Conference on Data Science, E-Learning and Information Systems, Article No.33, Pages 1-6, 2018

[2] M. Dalla Cia et al., “Using Smart City Data in 5G Self-Organizing Networks”, in IEEE Internet of Things Journal, vol. 5, no. 2, pp. 645-654, 2018

[3] RocksDB, website. http://rocksdb.org/.

[4] Redis. https://redis.io

[5] MySQL. https://github.com/mysql

[6] Dana Van Aken et al., “Automatic Database Management System Tuning Through Large-scale Machine Learning”, SIGMOD '17 : Proceedings of the 2017 ACM International Conference on Management of Data, Pages 1009-1024, 2017

[7] 김휘군, 최원기, 최종환, 성한승, 박상현, “데이터베이스 성능 향상을 위한 기계학습 기반의 RocksDB 파라미터 분석 연구”, 2020 온라인 추계학술박표대회 논문집, 제27권, 제2호

[8] 서주연, 이지은, 김경훈, JIN HUIJUN, 박상현, “비선형 기계학습 기반의 Redis 파라미터 튜닝 연구”, 한국정보과학회 2021컴퓨터종합학술대회 논문집, p.69 - 71

[9] FINN, Chelsea; ABBEEL, Pieter; LEVINE, Sergey. Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. PMLR, 2017. p. 1126-1135.

[10] 염찬호, 이지은, 서주연, & 박상현. (2021). OANet: 데이터베이스 성능 예측을 위한 Ortho Attention Net. 한국정보과학회 학술발표논문집, 81-83.

[11] Cao, Z., Dong, S., Vemuri, S., & Du, D. H. (2020). Characterizing, Modeling, and Benchmarking {RocksDB}{Key-Value} Workloads at Facebook. In 18th USENIX Conference on File and Storage Technologies (FAST 20) (pp. 209-223).