근사 질의 처리를 위한 확률 회로 모델의 점진적 학습

김성수, 이태휘
한국전자통신연구원
sungsoo@etri.re.kr, taewhi@etri.re.kr

Incremental Learning of Probabilistic Circuit Models for Approximate Query Processing

Sungsoo Kim, Taewhi Lee
Electronics and Telecommunications Research Institute

요 약

머신러닝 모델에 대한 점진적 학습은 변화하는 데이터 환경에 적응하고 모델의 성능을 지속적으로 향상시키기 위해 필수적인 기술이다. 특히, 대규모 빅데이터를 분석하는 ML 모델 이용하는 시스템에서 그 중요성이 더욱 부각된다. 본 논문에서는 근사 질의 처리를 위한 확률 회로 기반 모델에 대한 점진적 학습 기법을 제안한다. 인스타카트 벤치마크 데이터셋을 활용한 실험에서, 제안된 방법이 근사 질의 처리 모델에 대한 점진적 학습을 선형시간에 처리함을 보여주었다.

I. 서론

근사 질의 처리 (*Approximate Query Processing*, AQP)는 COUNT, SUM, AVG 등과 같은 집계 질의의 결과를 얻기 위해 모든 데이터를 처리하는 대신, 일부 데이터를 요약, 샘플링하거나 머신러닝 모델을 구축하여 질의를 수행하는 기술이다 [1, 2]. 아래 SQL 질의는 인스타카트 벤치마크 데이터셋에서 평균값을 계산하는 근사 질의에 대한 예를 보여주고 있다.

```
SELECT APPROXIMATE SUM(reordered) FROM order_products
WHERE add_to_cart_order <= 4;
```

특히, 머신러닝 기법을 활용한 AQP 기법으로는 지도학습을 수행하는 질의기반 (*query-driven*) AQP와 데이터기반 (*data-driven*) AQP로 분류할 수 있다 [3, 4].

연구 동기. ML 모델의 점진적 학습은 변화하는 데이터 환경에 적응하고 모델의 성능을 지속적으로 향상시키기 위한 필수적인 기술이다. 제안 시스템인 TRAINDB [1]에서는 데이터기반 AQP 기법인 확률 회로 기반 모델의 점진적 학습 기법을 소개한다.

확률 회로 모델. 합-곱 네트워크 (*Sum-Product Network*; SPN)는 루트가 있는 비순환 방향 그래프 (DAG)를 기반으로 하는 확률 회로 모델 (probabilistic circuit models)이다. 말단 (terminal) 노드는 단변량 확률 분포를, 비말단 (non-terminal) 노드는 확률 함수의 가중 합 (sum)과 곱셈 (product)을 각각 나타내며, 베이즈 네트워크 (Bayesian network)와 유사하지만, SPN은 잠재 변수를 도입하여 복잡한 조건적 분포를 더욱 효율적으로 표현할 수 있다는 장점이 있다. SPN의 합 노드는 여러 자식 노드의 확률 분포를 가중합하여 새로운 확률 분포를 생성한다. 반면, 곱 노드는 여러 자식 노드의 확률 분포를 곱하여 새로운 결합 분포를 생성한다. 그리고, 리프 (*leaf*) 노드는 단변량 확률 분포를 나타낸다. 일반적으로 가우시안 분포, 베르누이 분포 등을 사용한다. 특히, 관계형 합-곱 네트워크 (*Relational Sum-Product Network*; RSPN)는 SPN의 확장된 형태로, 관계형 데이터를 모델링하는 데 사용된다 [4]. SPN은 변수 간의 조건적 독립 관계를 표현하는 데 사용되는 확률적 그래프 모델의 일종이지만, RSPN은 객체 간의 관계를 명시적으로 표현할 수 있어 관계형 데이터를 더욱 효율적으로 모델링할 수 있다. 특히, 근사 질의 처리에서는 데이터베이스내 테이블들의 데이터 분포를 학습해서 활용할 수 있다 [4].

II. 근사 질의 처리 모델의 점진적 학습

2.1 점진적 학습 처리 개요

제안 시스템인 TRAINDB는 근사 질의 처리를 위해 RSPN 모델을 구축한 후 질의 처리를 수행한다. 그림 1은 TRAINDB에서 수행하는 점진적 학습 주요 절차에 대한 개념도를 보여주고 있다. 기존 ML 모델에 대한 점진적 학습 처리 절차는 다음과 같다.

(1) 기존 ML 모델을 메모리에 로딩한다.
(2) 새로운 추가 데이터를 적용하여 점진적 학습을 수행한다.
(3) 점진적 학습이 완료된 모델로 RSPN 모델을 업데이트한다.
(4) 업데이트된 RSPN 모델을 통해 근사 질의 처리를 수행한다.

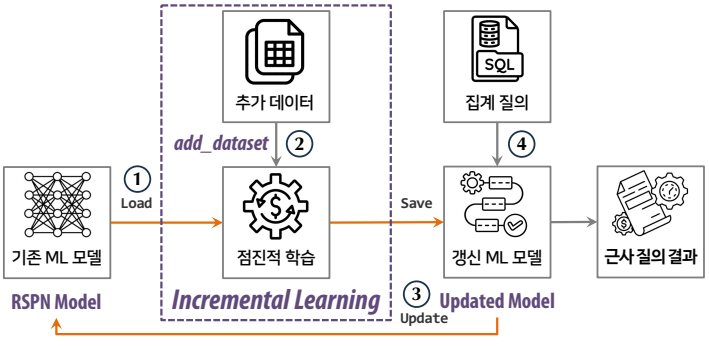

그림 1: 근사질의 처리 ML 모델에 대한 점진적 학습 처리 개념도

2.2 점진적 모델 업데이트 알고리즘

기존 데이터셋으로 학습된 RSPN의 트리 구조인 RSPN 모델을 활용하여, 새로운 데이터셋을 반영하여 점진적으로 모델을 업데이트한다 [4]. **Algorithm 1**은 근사질의 처리를 위해 구축된 RSPN 모델에 대한 점진적 모델 업데이트 절차를 기술하고 있다.

Algorithm 1: incremental_update (*node*, *tuple*)

```
   Input: RSPN node node, New dataset tuple
1  if node == leaf-node then
2  |  update_leaf_distribution(node, tuple)
3  end
4  else if node == sum-node then
5  |  nearest_child ← get_nearest_cluster(node, tuple)
6  |  adapt_weights(node, nearest_child)
7  |  incremental_update(nearest_child, tuple)
8  end
9  else if node == product-node then
10 |  for child in child_nodes do
11 |  |  tuple_proj ← project_to_child_scope(tuple)
12 |  |  incremental_update(child, tuple_proj)
13 |  end
14 end
```

이 알고리즘은 RSPN을 구성하고 있는 노드 유형(합, 곱, 리프 노드)에 따라 각각 처리를 진행하며, 합, 곱 노드에 대해서는 하위 트리를 incremental_update 함수를 재귀적으로 호출하여 처리한다. RSPN의 트리 루트 노드에서 부터 시작해서 하위 노드로 진행해 나가면서 처리한다. 합 노드일 경우(line 4), 새로 추가될 튜플이 어떤 자식 노드에 속하는지 식별하여, 어떤 가중치를 증가 또는 감소시켜야 할지 결정한다. 합 노드의 자식 노드는 학습 과정에서 k-means 클러스터링 알고리즘을 통해 찾은 행 클러스터를 나타내므로, 가장 가까운 클러스터 중심을 계산(line 5)하고 해당 가중치를 증감시킨 후(line 6), 튜플을 재귀적으로 이 하위트리로 전파 할 수 있다(line 7). 반대로 곱 노드는 열 집합을 분할한다(line 9). 따라서, 튜플을 자식 노드 중 하나로 전파하지 않고 분할하여 각 튜플 조각을 해당 자식 노드로 전파한다(line 10-11). 리프 노드에서는 튜플의 단일 컬럼 값에 따른 리프 분포를 업데이트 한다(line 2).

III. 실험 결과

TRAINDB의 근사질의 처리 모델에 대한 점진적 업데이트 알고리즘 성능 평가를 위해, 고객의 구매 행동 패턴 분석에 활용되는 인스타카트 (Instacart) 벤치마크 1GB 데이터셋을 이용하였다. 점진적 학습 성능 실험은 Intel Xeon W 3.5 GHz (8-core), 128Gb RAM 시스템 환경에서 진행했다. 실험을 위해, 초기 RSPN 모델은 order_products 테이블의 4개 integer 컬럼들 (*order_id, product_id, add_to_cart_order, reordered*)로 구성된 데이터셋[1]을 이용하여 구축했다. 이 근사 질의 처리 모델에 새로 추가되는 데이터를 2만개씩 증가시키며, RSPN 모델을 점진적으로 업데이트하는데 소요되는 시간을 측정했다. 그림 2는 order_products 테이블에 새로 추가되는 데이터셋 크기에 따른 TRAINDB 점진적 학습 성능 결과를 보여주고 있다. 제안한 점진적 RSPN 모델 업데이트 기법이 선형시간 (linear-time) 복잡도를 제공하여, 효과적으로 ML기반 근사 질의 처리에 적용할 수 있음을 확인할 수 있다.

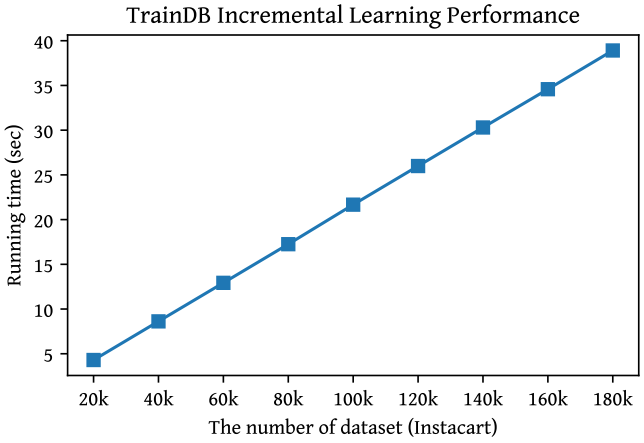

그림 2: 데이터셋 크기에 따른 TRAINDB 점진적 학습 성능

IV. 결론

본 논문에서는 끊임없이 변화하는 데이터 환경에서도 안정적인 성능을 유지하는 확률 회로 기반의 근사 질의 처리 모델을 위한 점진적 학습 방법을 제안했다. 제시된 방법은 대규모 데이터셋에 대한 점진적 학습을 가능하게 하여, 모델의 적응력을 높이고 정확도를 향상시킬 수 있다. 인스타카트 벤치마크 데이터셋을 활용한 실험 결과, 제안된 방법이 선형시간에 근사 질의 처리 모델을 선형시간에 업데이트할 수 있음을 확인할 수 있었다.

ACKNOWLEDGMENT

본 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00231, "빅데이터 대상의 빠른 질의 처리가 가능한 탐사 데이터 분석 지원 근사질의 DBMS 기술 개발")

참고 문헌

[1] S. Kim, C. S. Park, T. Lee, and K. Nam, "Constrained Approximate Query Processing with Error and Response Time-Bound Guarantees for Efficient Big Data Analytics," in *Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing, HPDC 2024, Pisa, Italy*, pp. 373–376, ACM, 2024.

[2] 김성수, 박춘서, 남택용, and 이태휘, "샘플링 데이터를 이용한 혼합 밀도 네트워크 모델기반 근사 질의 처리," *정보과학회 컴퓨팅의 실제 논문지*, vol. 28, no. 9, pp. 150–157, 2022.

[3] F. Savva, C. Anagnostopoulos, and P. Triantafillou, "ML-AQP: Query-Driven Approximate Query Processing based on Machine Learning," *CoRR*, vol. abs/2003.06613, 2020.

[4] B. Hilprecht, A. Schmidt, M. Kulessa, A. Molina, K. Kersting, and C. Binnig, "DeepDB: Learn from Data, not from Queries!," *Proc. VLDB Endow.*, vol. 13, no. 7, pp. 992–1005, 2020.

[1] https://www.kaggle.com/c/instacart-market-basket-analysis