Active Learning for ML-Enhanced Database Systems

Lin Ma, Bailu Ding, Sudipto Das, Adith Swaminathan
Carnegie Mellon University
Microsoft Research

lin.ma@cs.cmu.edu
Emerging ML-Enhanced Databases

Many academic contributions

- Query Run-time Prediction
- Query Optimization
- Index Recommendation
- Autonomous Administration

Challenge at deployments
ML-Enhanced Database Example

ML Model:
- [Ding, B., et al. SIGMOD 2019]

Model Input

Model Output
- P1 is cheaper than P2

Applications
- Query Optimizer
- Index Advisor
ML-Enhanced Database Example
Simulated Model Training and Deployment

Collect Data
- Standard Benchmarks and Available Workloads

Train
- Training Error: 2%
- Validation Error: 5%

Deploy (simulated)
- Error: 32%

What’s wrong?
Challenge: Data Distribution Shift

ML assumes same training-test distribution

Test data distribution varies heavily in production databases

Key barrier to productionize ML for databases
Solution: Collect More Data in Deployments

Insight: actively collect data for individual database deployments

• Acquire labels from replicas (b-instances) without impacting the normal operation

• The “target test data” is often derivable for a specific workload

Reduces 75% error by executing ~100 queries
Active Data Collection Platform

Production Database

Replica / B-Instance

ML Enhanced Component

Query Optimizer

Index Advisor

Target Test Data

Selected Unlabeled Data

New Labels

Make Prediction

Can be large

Plan Space

Budget and # iterations

New Labels

Can be large

Model

User

ML Model

ADCP

Can be large
Active Learning

AL strategy selects the best training data from a pool of unlabeled data

- Long and successful history in database crowdsourcing

Typical AL:

\[ w(x) \]

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
<th>x7</th>
<th>x8</th>
<th>x9</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Most common \( w(x) \): uncertainty

Model Input

- P1 is cheaper than P2: 70%
- P2 is cheaper than P1: 30%

Uncertainty: 30%

Model Output
Holistic AL Challenges

- Robust
  Noisy uncertainty signal under significant distribution shift

- Cost-sensitive
  Drastically different labeling costs, especially with index creations

- Batch-friendly
  Expensive model retraining
## Holistic AL Challenges

<table>
<thead>
<tr>
<th>AL Strategy</th>
<th>Robust</th>
<th>Cost-sensitive</th>
<th>Batch-friendly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertainty</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Hybrid</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBMAL</td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>ROUND</td>
<td>Y</td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>SIMILAR</td>
<td>Y</td>
<td></td>
<td>Y</td>
</tr>
</tbody>
</table>

Fertile area of future research

[Hass, D., et al. VLDB 2015]
Holistic Active Learner (HAL) for ADCP

**Biased sampling: robust**

<table>
<thead>
<tr>
<th>$w(x)$</th>
<th>0.9</th>
<th>0.8</th>
<th>0.7</th>
<th>0.6</th>
<th>0.5</th>
<th>0.4</th>
<th>0.3</th>
<th>0.2</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Cost weighting: cost-sensitive**

- Per “cost unit” uncertainty

**Redundancy rejection: batch friendly**
Evaluation

14 workloads include industrial standard benchmarks (e.g., TPC-DS) and customer workloads

- Hold out each workload as the target production database, and round robin
- 30K plans, 1M plan pairs

Multiple AL iterations with evenly split budget for each iteration

- Total budget of 150x average estimate plan cost

Different ML tasks, budget sizes, models, features, cost types, or no cost estimation
Baselines

Optimizer

Random

Uncertainty

Hybrid
  • Random + Uncertainty
  • [Hass, D., et al. VLDB 2015]
Results

Model Error Reduction

F1 Error

# Iteration

0.32

Random

Uncertainty

Hybrid

HAL

Optimizer

~100 Queries

Budget: 50x average query cost per iteration
Addressing the training/deployment distribution shift is crucial for ML-enhanced databases

A practical solution to actively collect training data during deployment using replicas and HAL

Fertile area of future research
- Better address the holistic AL challenges
- Better use the training data during deployments

lin.ma@cs.cmu.edu