A Gradient-based Adaptive Learning Framework for Efficient Personal Recommendation

Yue Ning\textsuperscript{1}  Yue Shi\textsuperscript{2}  Liangjie Hong\textsuperscript{2}  
Huzefa Rangwala\textsuperscript{3}  Naren Ramakrishnan\textsuperscript{1}

\textsuperscript{1}Virginia Tech  
\textsuperscript{2}Yahoo Research. Yue Shi is now with Facebook, Liangjie Hong is now with Etsy.  
\textsuperscript{3}George Mason University

August 27, 2017
Outline

Introduction
  Problem Challenges

The Proposed Framework

Applications
  Adaptive Logistic Regression
  Adaptive Gradient Boosting Decision Tree
  Adaptive Matrix Factorization

Experimental Evaluation
  Datasets & Metrics
  Comparison Methods
  Ranking Scores

Summary
Challenges in Personalized Recommender Systems

- Alleviate “average” experiences for users.
Challenges in Personalized Recommender Systems

- Alleviate “average” experiences for users.
- Lack of generic empirical frameworks for different models.
Challenges in Personalized Recommender Systems

- Alleviate “average” experiences for users.
- Lack of generic empirical frameworks for different models.
- Distributed model learning and less access of data.
Example of Personal Models

Figure: An example of global and personal models. Left figure showcases the nDCG score of users from global (y-axis) and personal (x-axis) models. (Right: MAP score).
Figure: System Framework. Component $C_1$ trains a global model. Component $C_2$ generates a hashtable based on users’ data distribution. Users request $t_u$ from $C_2$ and $C_1$ returns a subsequence of gradients $g^{(0:t_u)}$ to users.
Adaptation Mechanism

Global update →

$$\theta^{(T)} = \theta^{(0)} - \eta \sum_{t=1}^{T} g^{(t)}(\theta)$$

Local update →

$$\widetilde{\theta}_u = \theta^{(0)} - \eta_1 \sum_{t=1}^{t_u-1} g^{(t)}(\theta) - \eta_2 \sum_{t=t_u}^{T} g^{(t)}(\theta_u)$$

- $\theta$: the global model parameter.
- $\theta_u$: the personal model parameter.
- $u$: the index for one user.
- $t_u$: the index of global gradients for user $u$.
- $g^{(t)}(\theta)$: global gradients
- $g^{(t)}(\theta_u)$: personal gradients
How do we decide \( t_u \)?

- Group users into \( C \) groups based on their data sizes in descending order.
- Decide the position \( p_u = \frac{i}{C} \),
  - \( C \) is \# groups.
  - \( i \) is the group assignment for user \( u \).
  - the first group (\( i=1 \)) of users has the most data.
- Set \( t_u = \lfloor T \times p_u \rfloor \)
  - \( T \): total iterations in the global SGD algorithm
  - Users with the most data have the earliest stop for global gradients.
Adaptive Logistic Regression

Objective:

\[
\min_{\mathbf{w}} L(\mathbf{w}) = f(\mathbf{w}) + \lambda r(\mathbf{w}) \quad (1)
\]

- \( f(\mathbf{w}) \) is the negative log-likelihood.
- \( r(\mathbf{w}) \) is a regularization function.

Adaptation Procedure:

- Global update \( \rightarrow \)

\[
\tilde{\mathbf{w}}_{u}^{(0)} = \mathbf{w}^{(0)} - \eta_1 \sum_{t=1}^{t_u-1} g^{(t)}(\mathbf{w}) \quad (2)
\]

- Local update \( \rightarrow \)

\[
\tilde{\mathbf{w}}_{u}^{(T)} = \tilde{\mathbf{w}}_{u}^{(0)} - \eta_2 \sum_{t=1}^{T-t_u} g^{(t)}(\mathbf{w}_u) \quad (3)
\]
Adaptive Gradient Boosting Decision Tree

Objective:

\[ L^{(t)} = \sum_{d}^{N} l(y_d, F_d^{(t-1)} + \rho h^{(t)}) + \Omega(h^{(t)}) \]

\[ = \sum_{d}^{N} l(y_d, F_d^{(0)} + \rho h^{(0:t)}) + \Omega(h^{(t)}) \quad (4) \]

Adaptation Procedure:

\[ \tilde{F}_{u}^{(0)} = F^{(0)} + \rho h^{(0:t_u)} \quad (5) \]

\[ \tilde{F}_{u}^{(T)} = \tilde{F}_{u}^{(0)} + \rho h^{(t_u:T)} \quad (6) \]
Adaptive Matrix Factorization

Objective:

$$\min_{q^*, p^*, b^*} \sum_{u, i} (r_{ui} - \mu - b_u - b_i - q_u^T p_i)$$

$$+ \lambda (\|q_u\|^2 + \|p_i\|^2 + b_u^2 + b_i^2)$$ (7)

Adaptation Procedure:

$$\tilde{q}^{(0)}_u = q^{(0)}_u - \eta_1 \sum_{t=0}^{T-t_u} g^{(t)}(q_u)$$, $$\tilde{q}^{(T)}_u = \tilde{q}^{(0)}_u - \eta_2 \sum_{t=0}^{T-t_u} g^{(t)}(\tilde{q}_u)$$ (8)

$$\tilde{b}^{(0)}_u = b^{(0)}_u - \eta_1 \sum_{k=0}^{T-t_u} g^{(t)}(b_u)$$, $$\tilde{b}^{(T)}_u = \tilde{b}^{(0)}_u - \eta_2 \sum_{t=0}^{T-t_u} g^{(t)}(\tilde{b}_u)$$ (9)
Properties

- **Generality**: The framework is generic to a variety of machine learning models that can be optimized by gradient-based approaches.

- **Extensibility**: The framework is extensible to be used for more sophisticated use cases.

- **Scalability**: In this framework, the training process of a personal model for one user is independent of all the other users.
Datasets

Table: Dataset Statistics

<table>
<thead>
<tr>
<th>News Portal</th>
<th>Movie Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td># users</td>
<td>Netflix</td>
</tr>
<tr>
<td># features</td>
<td># users</td>
</tr>
<tr>
<td>54845</td>
<td>478920</td>
</tr>
<tr>
<td>351</td>
<td></td>
</tr>
<tr>
<td>2,378,918</td>
<td>sparsity</td>
</tr>
<tr>
<td>26,916,620</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
</tr>
<tr>
<td>534</td>
<td></td>
</tr>
</tbody>
</table>

- For LogReg and GBDT: News Portal dataset
- For Matrix Factorization: Movie rating datasets (Netflix, Movielens)
Metrics

- MAP: Mean Average Precision.
- MRR: Mean Reciprocal Rank.
- AUC: Area Under (ROC) Curve.
- nDCG: Normalized Discounted Cumulative Gain.
- RMSE: Root Mean Square Error
- MAE: Mean Absolute Error
Comparison Methods

Table: Objective functions for different methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>LogReg</th>
<th>GBDT</th>
<th>MF</th>
</tr>
</thead>
</table>
| Global | \[\sum_{d=1}^{N} f(w) + \lambda ||w||^2\] | \[\sum_d l(y_d, F_d^{(0)} + \rho h^{(0:t)}) + \Omega(h^{(t)})\] | \[
\sum_{u,i} (r_{ui} - \mu - b_u - b_i - q_u^T p_i) + \lambda (||q_u||^2 + ||p_i||^2 + b_u^2 + b_i^2)\]
| Local | \[\sum_{u}^N f(w_u) + \lambda ||w_u||^2\] | \[\sum_j l(y_j, F_j^{(0)} + \rho h^{(0:t)}) + \Omega(h^{(t)})\] | \[
\sum_{i \in N_u} (r_{ui} - \mu - \tilde{b}_u - \tilde{b}_i - \tilde{q}_u^T \tilde{p}_i) + \lambda (||\tilde{q}_u||^2 + ||\tilde{p}_i||^2 + \tilde{b}_u^2 + \tilde{b}_i^2)\]
| MTL | \[\sum_j f(w_u) + \frac{\lambda_1}{2} ||w_u - w||^2 + \frac{\lambda_2}{2} ||w_u||^2\] | - | \[\text{global} + \lambda_2 [(q_u - q)^2 + (p_i - p)^2 + (b_u - A_u)^2 + (b_i - A_i)^2]\] |

- Global: models are trained on all users’ data
- Local: models are learned locally on per user’s data
- MTL: users models are averaged by a global parameter.
AUC, MAP, MRR and nDCG scores on the test dataset with varying training epochs.

The proposed adaptive LogReg models achieve higher scores with fewer epochs.

Global models perform the worst.
## Ranking Performance - GBDT

Table: Performance comparison based on MAP, MRR, AUC and nDCG for GBDT. Each value is calculated from the average of 10 runs with standard deviation.

| #Trees | Global-GBDT |  |  |  |  |
|-------|-------------|-------------|-------------|-------------|
|       | MAP         | MRR         | AUC         | nDCG        |
| 20    | 0.2094(1e-3)| 0.3617(2e-3)| 0.6290(1e-3)| 0.5329(6e-4)|
| 50    | 0.2137(1e-3)| 0.3726(1e-3)| 0.6341(1e-3)| 0.5372(6e-4)|
| 100   | 0.2150(8e-3)| 0.3769(1e-3)| 0.6356(8e-4)| 0.5392(6e-4)|
| 200   | 0.2161(5e-4)| 0.3848(1e-3)| 0.6412(6e-4)| 0.5415(5e-4)|

| #Trees | Local-GBDT |  |  |  |  |
|-------|------------|-------------|-------------|-------------|
|       | MAP         | MRR         | AUC         | nDCG        |
| 20    | 0.2262(2e-3)| 0.4510(5e-3)| 0.6344(3e-3)| 0.5604(2e-3)|
| 50    | 0.2319(2e-3)| 0.4446(4e-3)| 0.6505(2e-3)| 0.5651(2e-3)|
| 100   | 0.2328(1e-3)| 0.4465(5e-3)| 0.6558(2e-3)| 0.5651(2e-3)|
| 200   | 0.2322(2e-3)| 0.4431(2e-3)| 0.6566(1e-3)| 0.5649(1e-3)|

| #Trees | Adaptive-GBDT |  |  |  |  |
|-------|---------------|-------------|-------------|-------------|
|       | MAP           | MRR         | AUC         | nDCG        |
| 20+50 | 0.2343(2e-3)  | 0.4474(4e-3)| 0.6555(2e-3)| 0.5661(2e-3)|
| 50+50 | 0.2325(2e-3)  | 0.4472(1e-4)| 0.6561(8e-4)| 0.5666(6e-4)|
| 10+100| 0.2329(2e-3)  | 0.4423(3e-3)| 0.6587(1e-3)| 0.5650(3e-3)|
MAP score for the groups of users with least data (Group 1) and most data (Group 7) for GBDT models.

- Adaptive-GBDT outperform both global and local GBDT models in terms of MAP for all groups of users.
Ranking Performance - LogReg vs GBDT

▶ AUC score for Global-GBDT, Local-GBDT, and Adaptive-GBDT with # of training samples from 20% to 100%.

▶ On average of AUC, Adaptive-GBDT performs better than other methods.

▶ With the increase of training samples, GBDT based methods tend to perform better while LogReg methods achieve relatively stable scores.
Results - MF

- RMSE and MAE on MovieLens(ML) and Netflix datasets.
- The quartile analysis of the group level RMSE and MAE for different MF models.
- Gold: Adaptive-MF

(a) ML-RMSE

(b) ML-MAE

(c) Netflix-RMSE

(d) Netflix-MAE
Summary

- Effectively and efficiently build personal models that lead to improved recommendation performance over either the global model or the local model.
- Adaptively learn personal models by exploiting the global gradients according to individuals characteristic.
- Our experiments demonstrate the usefulness of our framework across a wide scope, in terms of both model classes and application domains.
Thank you!

Q&A