Personalized Treatment Selection using Causal Heterogeneity

Kinjal Basu
Sr. Staff Researcher @ LinkedIn AI

INFORMS 2021

Joint work with Ye Tu, Preetam Nandy, Cyrus DiCiccio, Romil Bansal, Padmini Jaikumar & Shaunak Chatterjee
Overview

1. Introduction
2. Problem Setup
3. Methodology
4. Results
Example Problem: Notification Systems

- Notifications are an important driver for member visits and engagement.

- Sending more notifications can increase visits, but it also has negative consequences where members can disable the system.

- **Challenging Business Problem**: How to identify the optimal number of relevant notifications that we should send to our member?
Heterogeneity of Treatment Effects

Randomized experimentation (A/B testing) is widely used in the internet industry to measure the metric impact obtained by different treatment variants.

Treatment example: different number of notifications that members receive.

The effect of a given treatment can be heterogeneous across experimental units.
Personalized Treatment Selection

**Global allocation:** identify the treatment variant that performs the best in the entire population and ramp that variant to everyone.

A **personalized approach** for treatment selection can greatly improve upon the usual global selection strategy.
**Major Contributions**

A **general framework** for identifying the optimal cohorts and allocating the optimal treatment per cohort.

- **Framework of solutions**: With guidance on which one to pick and when

- **Technical novelty**
  - Merging tree algorithm - Selects optimal cohorts
  - Multiple cooperative stochastic approximation - Selecting optimal treatments

- **Real-world application**
  - Building the serving infrastructure
  - Strong, positive results from a large scale industrial application
Problem Set-up
Notations and Objective

Let $k = 0$ denote the main success metric
- Sessions in the notifications example

Other constraints metrics $k = 1, 2, \ldots, K$
- Clicks
- Disables, etc

Maximize Sessions
S.t. \hspace{1cm} \text{Clicks} > c
\hspace{1cm} \text{Disables} < d

Problem variable (x) :
- Numbers of notifications we sent to members
Notations and Objective

Objective 1: Find optimal cohorts

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J$</td>
<td>Total number of treatment variants or choices.</td>
</tr>
<tr>
<td>$K$</td>
<td>Total number of guardrail metrics</td>
</tr>
<tr>
<td>$C_i$</td>
<td>$i$-th cohort (the smallest cohort would be an individual member) for $i = 1, \ldots, n$.</td>
</tr>
</tbody>
</table>
Notations and Objective

Objective 1: Find optimal cohorts

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J$</td>
<td>Total number of treatment variants or choices.</td>
</tr>
<tr>
<td>$K$</td>
<td>Total number of guardrail metrics</td>
</tr>
<tr>
<td>$C_i$</td>
<td>$i$-th cohort (the smallest cohort would be a individual member) for $i = 1, \ldots, n.$</td>
</tr>
<tr>
<td>$U_k$</td>
<td>Vectorized version of $U_{i,j}^k$, which is the causal effect in metric $k$ by variant $j$ in cohort $C_i$.</td>
</tr>
<tr>
<td>$\mu_k$</td>
<td>Mean of $U_k$</td>
</tr>
<tr>
<td>$\sigma^2_k I$</td>
<td>Variance of $U_k$</td>
</tr>
</tbody>
</table>
## Notations and Objective

### Objective 1: Find optimal cohorts

### Objective 2: Allocate the optimal treatment to the cohorts

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J$</td>
<td>Total number of treatment variants or choices.</td>
</tr>
<tr>
<td>$K$</td>
<td>Total number of guardrail metrics</td>
</tr>
<tr>
<td>$C_i$</td>
<td>$i$-th cohort (the smallest cohort would be an individual member) for $i = 1, \ldots, n.$</td>
</tr>
<tr>
<td>$U_k$</td>
<td>Vectorized version of $U_{i,j}^k$, which is the causal effect in metric $k$ by variant $j$ in cohort $C_i$.</td>
</tr>
<tr>
<td>$\mu_k$</td>
<td>Mean of $U_k$</td>
</tr>
<tr>
<td>$\sigma_k^2$</td>
<td>Variance of $U_k$</td>
</tr>
<tr>
<td>$x$</td>
<td>The assignment vector.</td>
</tr>
</tbody>
</table>
Notations and Objective

Objective 1: Find optimal cohorts

Objective 2: Allocate the optimal treatment to the cohorts

Formally, we wish to get the optimal \( x^* \) by solving the following:

Maximize \( x^T U_0 \)

subject to \( x^T U_k \leq c_k \) for \( k = 1, \ldots, K \).

\[ \sum_j x_{i,j} = 1 \quad \forall i, \quad 0 \leq x \leq 1 \]

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J )</td>
<td>Total number of treatment variants or choices.</td>
</tr>
<tr>
<td>( K )</td>
<td>Total number of guardrail metrics</td>
</tr>
<tr>
<td>( C_i )</td>
<td>( i )-th cohort (the smallest cohort would be an individual member) for ( i = 1, \ldots, n ).</td>
</tr>
<tr>
<td>( U_k )</td>
<td>Vectorized version of ( U_{i,j}^k ), which is the causal effect in metric ( k ) by variant ( j ) in cohort ( C_i ).</td>
</tr>
<tr>
<td>( \mu_k )</td>
<td>Mean of ( U_k )</td>
</tr>
<tr>
<td>( \sigma_k^2 I )</td>
<td>Variance of ( U_k )</td>
</tr>
<tr>
<td>( x )</td>
<td>The assignment vector.</td>
</tr>
</tbody>
</table>
Problem Breakdown

1. Identify member cohorts $C_1, \ldots, C_n$ using data from randomized experiments to estimate the causal effect $U_k$ for each cohort

2. Optimally allocate treatment variants $x^*$ to each member cohort by solving the optimization problem.
Methodology
Cohort-Level Heterogeneity

We use the recursive partitioning technique from Athey and Imbens [1] to identify the heterogeneous cohorts.

Regression Tree (CART)
- Predict Y
- Splitting Objective: MSE(Y)

Causal Tree
- Estimate treatment effect $\tau$
  $E(Y(1)) - E(Y(0))$
- Splitting Objective: $\text{MSE}(\tau) + \text{Variance regularizer}$
Multiple treatments and metrics

Causal tree can only handle one objective metric and a binary treatment definition at a time.

- One option could be merging the $J (K + 1)$ tree models into one single cohort assignment.

- Simply merging all the trees would fragment the cohorts into very small subsets with extremely noisy estimations.

- We avoid this unwanted noise by carefully exploiting the within cohort homogeneity of the treatment effect by Algorithm 1.
Merging Trees - Algorithm 1

We sequentially merge the cohort sets $S_j,k = \{ C_1^{j,k}, \ldots, C_n^{j,k} \}$

For each treatment $j$ and each metric $k$, we retain the estimated treatment effect and its variance from the original cohort.

Since each $S_j,k$ is exhaustive, this provides estimates of treatment effect and its variance for all sub-partitions.

---

**Algorithm 1 Merging Trees**

**Input:** $L$ cohorts sets: $\{\{C_i^l\}_{i=1}^{n_l} | l = 1, \ldots, L\}$ and corresponding treatment effects and variances $\{((U(C), \sigma^2(C)) | C \in \{C_i^l\}_{i=1}^{n_l} | l = 1, \ldots, L\}$

**Output:** $S_{out}$ and $T_{out}$

1. Set $S_{out} = \{C_i^l\}_{i=1}^{n_l}$ and $T_{out} = \{(U_1(C), \sigma^2_1(C)) | C \in S_{out}\}$
2. for $\ell = 2, \ldots, L$ do
3.    for $A \in S_{out}$ do
4.      for $B \in \{C_i^l\}_{i=1}^{n_l}$ do
5.        $C = A \cap B$
6.        if $C \neq \emptyset$ then
7.          $S_{out} = S_{out} \cup \{C\}$
8.          $T_{out} = T_{out} \cup \{(U_m(C), \sigma^2_m(C)) | m = 1, \ldots, \ell\}$
9.        end if
10.    end for
11.  end for
12. $S_{out} = S_{out} \setminus \{A\}$
13. $T_{out} = S_{out} \setminus \{(U_m(A), \sigma^2_m(A)) | m = 1, \ldots, \ell - 1\}$
14. end for

where

$$U_m(C), \sigma^2_m(C) = \begin{cases} U_m(A), \sigma^2_m(A) & \text{for } m \leq \ell - 1 \\ U_m(B), \sigma^2_m(B) & \text{for } m = \ell \end{cases}$$
Optimization Solution

**Stochastic Optimization:** the problem is stochastic since both the objective function and the constraints are not deterministic but are coming from a particular distribution (e.g., Gaussian).

Maximize $x^T U_0$

subject to $x^T U_k \leq c_k$ for $k = 1, \ldots, K.$

$\sum_j x_{i,j} = 1 \ \forall i, \quad 0 \leq x \leq 1$

Maximize $f(x) = \mathbb{E}(x^T U_0)$

subject to $g_k(x) := \mathbb{E}(x^T U_k - c_k) \leq 0, \quad k = 1, \ldots, K.$

$\sum_j x_{ij} = 1 \ \forall i, \quad 0 \leq x \leq 1$
Multiple Cooperative Stochastic Approximation

At each step $t$ it starts by estimating the constraint function.

$$\hat{G}_{k,t} = \frac{1}{L} \sum_{\ell=1}^{L} G_k(x_t, U_{k,\ell}).$$

- If feasible, chooses the gradient to be the gradient of the \textbf{objective}.

- Otherwise, from the set of violated constraints, it chooses a constraint \textbf{at random} and use the gradient of \textbf{that constraint}.

\textbf{Algorithm 2 Multiple Cooperative Stochastic Approximation}

1: Input: Initial $x_1 \in \mathcal{X}$, Tolerances $\{\eta_k\}_t$, $\{\gamma_t\}_t$, Iterations $N$
2: for $t = 1, \ldots, N$ do
3:   Estimate $\hat{G}_{k,t}$ for all $k \in 1, \ldots, K$ using (4).
4:   if $\hat{G}_{j,t} \leq \eta_{j,t}$ for all $j$ then
5:     Set $h_t = F'(x_t, U_{0,t})$
6:   else
7:     Randomly select $k^*$ from $\{k : \hat{G}_{k,t} > \eta_{k,t}\}$
8:     Set $h_t = G'_{k^*}(x_t, U_{k^*,t})$
9:   end if
10:  Compute $x_{t+1} = P_{x_t}(y_t h_t)$
11: end for
12: Define $\mathcal{B} = \{1 \leq t \leq N : \hat{G}_{k,t} \leq \eta_{k,t} \ \forall k \in \{1, \ldots, K\}\}$
13: return $\hat{x} := \frac{\sum_{t \in \mathcal{B}} x_t y_t}{\sum_{t \in \mathcal{B}} y_t}$
Overall Algorithm

- Run Randomized Experiments to collect data across various treatments and metrics
- Generate a cohort-level causal effects for the different parameters.
- Solve the Stochastic Optimization to generate the final treatment allocation to each cohort
Results
Notification Systems Results

- Metrics of Interest:

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sessions (Objective)</td>
<td>Number of visits to the LinkedIn site/app</td>
</tr>
<tr>
<td>Notification Sends</td>
<td>Volume of notifications sent to members</td>
</tr>
<tr>
<td>Notification CTR</td>
<td>Click through rate on notifications</td>
</tr>
<tr>
<td>Total Disables</td>
<td>Number of total disables on notifications</td>
</tr>
</tbody>
</table>

Table 3: Metrics of Interest for Personalized Capping

Problem:

Maximize Sessions
S.t. Send < s
    CTR > c
    Disables < d
Notification System Results

- **Baseline**: Heuristic Cap A and B are based on a cohort definition where members are grouped into four segments according to their visit frequency.

- **Personalized cap treatment** showed significant positive impact on Sessions, while the impact on the constraint metrics remained within acceptable bounds. It also outperforms both heuristic solutions.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>ATE % Personalized Cap</th>
<th>ATE % Heuristic Cap A</th>
<th>ATE % Heuristic Cap B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sessions</td>
<td>+1.39%</td>
<td>+1.31%</td>
<td>+0.54%</td>
</tr>
<tr>
<td>Notification Sends</td>
<td>+1.64%</td>
<td>+6.62%</td>
<td>+3.07%</td>
</tr>
<tr>
<td>Notification CTR</td>
<td>-1.24%</td>
<td>-1.73%</td>
<td>-1.18%</td>
</tr>
<tr>
<td>Total Disables</td>
<td>Neutral</td>
<td>+9.23%</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

*Table 4: Notification Cap Experiment Results*
Discussions
Future work

A few non-trivial, but likely impactful extensions for future consideration include:

(1) Designing a more cost-efficient data collection framework or leveraging observational data to achieve the same performance would be beneficial.

(2) Users can potentially move in and out of cohorts. Extending this framework to incorporate the dynamic nature of cohorts could be an interesting research topic.

(3) Future work on generating one single optimal cohort definition based on effects from multiple treatments with various metrics of interests could further improve the method.
Reference


Thank you
Appendix
Member-level Heterogeneity

To estimate the heterogeneous causal effects at a member level, some of the options include:

(a) **Causal Forest**: The Causal Forest Algorithm is an extension of the Causal Tree which was inspired by Random Forest Algorithm and use ensemble learning to incorporate results from multiple tree models.

(b) **Two-Model Approach**: This is a baseline method (commonly applied in uplift modeling domain) that models the causal effect at a member level through the difference of the predicted response in the treatment and control models.
System Architecture

The general engineering architecture consists of two major components:

- One for heterogeneous causal effect estimations
- The other for the optimization module.
Simulation Analysis

We leverage simcausal R package [23] to generate simulation datasets under self-defined causal **Directed Acyclic Graphs** (DAG).

- $A_j$ as the treatment variables
- $Y_k$ are the metrics (or response variables)
- $U_y$ as a latent variable impacts $Y_k$
- $H_m$ as the heterogeneous variables

We simulate heterogeneity by introducing **interaction terms** between $A_j$ and $H_m$ on $Y_k$. 

![Directed Acyclic Graph](image-url)
Evaluation of Simulation

We consider the normalized mean of individualized treatment effect (ITE) for metric $k$ at optimal $x^*$ as

$$
\tau(x^*_k) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{3} (Y_{j,i,k} - Y_{0,i,k}) z_{i,j}^*,
$$

$(Y_{j,i,k} - Y_{0,i,k})$ represents the individualized treatment effect. We normalize the ITE by the control group mean $\mu$ to make results comparable across different simulated datasets.
Comparing all variants

(1) **HT.ST** : A heuristic cohort-level solution paired with stochastic optimization.

(2) **CT.ST** : Cohort-level estimations using Causal Tree model paired with stochastic optimization.

(3) **CF.DT** : Member-level estimations using the Causal Forest model [30] paired with deterministic optimization.

(4) **TM.DT** : Member-level estimations using a “Two-Model” approach (i.e., build two Random Forest [5] models) paired with deterministic optimization.

(5) **Global** : A best global allocation as baseline.
Analysis Results - Exist a global best

First scenario: Aligning the effect on the objective with that of the constraint metrics.

Benefit of the stochastic optimization:

- the cohort-level solutions paired with stochastic optimization ($HT.ST$ and $CT.ST$) perform almost at parity with the oracle global best solution $Global$.
- However, the member-level estimations paired with deterministic optimization ($CF.DT$ and $TM.DT$) show worse performance due to the high variance.

(a) Evaluation on the objective metric $Y_0$
(b) Evaluation on the constraint metric $Y_1$
(c) Evaluation on the constraint metric $Y_2
Analysis Results - No global best

**Second scenario:** the objective metrics move possibly in the opposite direction to some constraint metrics for some treatment.

**Benefit of heterogeneity estimation and personalization:**
- All the proposed approaches perform better than the *Global* solution.
- With low noise levels, member-level solutions (*CF.DT* and *TM.DT*) perform better than the cohort-level solution (*HT.ST, CT.ST*). Along with an increase in the noise level, *CT.ST* quickly starts to catch up and can outperform the member-level solutions.
Reproducibility
We share example scripts for conduct simulation analysis in examining the proposed methods and stochastic optimization algorithms in the following Github link: https://github.com/tuye0305/prophet.