# 修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission:01/10/2023 (MM/DD/YYYY)

| 専攻名（専門分野）<br>Department | 経営システム<br>工学専攻 | 氏 名<br>Name | 今福　太一<br>Taichi Imafuku | 指 導<br>教 員<br>Advisor | 後藤　正幸　　印<br>Seal |
|---|---|---|---|---|---|
| 研究指導名<br>Research guidance | 情報数理応用研究 | 学籍番号<br>Student ID number | 5221C009-5 | | |
| 研究題目<br>Title | 重要度サンプリングを用いた敵対的反事実回帰モデルの提案<br>Adversarial CounterFactual Regression with Importance Weighting | | | | |

## 1. 研究背景と目的

近年多くの分野で，個別レベルの因果効果である個別介入効果 (Individual Treatment Effect：ITE) を推定することに関心が集まっている．例えば，EC サイト上でクーポンを発行するユーザを選定する際，クーポンを発行することで購買金額が増えるユーザにクーポンを発行したい．しかし，あるユーザにクーポンを発行したときの結果と発行しなかったときの結果の両方を同時に観測することはできないため，その差分である個別介入効果は直接観測することができない．このとき，介入するときの結果と介入をしないときの結果に影響を与える変数である共変量を活用して個別介入効果を推定する．具体的には，共変量が同じ値で観測できなかった方の結果を持っているサンプルの結果で補完することによって個別介入効果を推定する．

しかし，類似した共変量を持つサンプルが対照群に存在しないとき，より一般化すると，介入を受ける群である介入群と介入を受けない群である対照群の共変量の分布が乖離しているとき，個別介入効果を正確に推定することができない．この問題に対処したモデルとして，CounterFactual Regression(CFR)[1] が知られている．CFR は，データを介入群と対照群の共変量が一致するような特徴空間に写像し，その空間内で回帰を行うことで個別介入効果を推定する．

しかし，共変量の分布の乖離が大きいとき，共変量の乖離が小さくなるような特徴空間を獲得することが難しくなるため，CFR の推定精度が大幅に悪化してしまう．共変量の分布に大きな乖離が生じるのは，共変量空間において，対照群のサンプルが比較的広く分布しているのに対し，介入群のサンプルが一部の領域にしか分布していないときである．例えば，EC サイトにおける推薦を因果推論における介入とみなす場合が想定される．一般的に EC サイトでは，人気商品がユーザに推薦される場合が多いため，EC サイトにおいて多数存在する不人気商品が推薦されるユーザは少数である．そのため，不人気商品をユーザに推薦したときの効果を推定することは難しい．このような状況で個別介入効果をより正確に推定するためには，介入群の共変量の分布を広げる，つまり対照群のいくつかのサンプルに介入する必要がある．しかし，介入にはコストが生じたり，EC サイトにおける推薦のようにそもそも介入できる数に制約が存在する場合がある．そのため，より個別介入効果の推定に寄与する対照群のサンプルに介入したいと考えられる．介入群のサンプルが一部の領域にしか分布せず，個別介入効果の推定が困難な状況に対処する方法は今まで研究されていない．しかし，この状況に対処すれば，今まで個別介入効果の推定が困難であった領域でより正確に個別介入効果を推定可能になり，その推定結果をもとにより良い意思決定が可能になる．

そこで本研究では，対照群のサンプルに介入することが可能な状況下で，個別介入効果の予測精度をより向上させる対照群のサンプルを特定する手法を提案する．具体的には，CFR において分布の乖離度を測る機構として Generative Adversarial Nets (GAN)[2] の識別器を利用し，学習後に得られる識別器の出力から対照群の各サンプルの重要度を算出し，その重要度が高いサンプルに対して介入を行う．これにより，より少ないコストで個別介入効果の推定精度を向上させることが可能になる．最後に実データを用いて提案手法の有効性を検証する．

## 2. 準備

### 2.1. 個別介入効果の定義

個別介入効果を定義する．サンプル $i$ に仮に介入するときの潜在的結果変数を $y_i^t$，仮に介入しないときの潜在的結果変数を $y_i^c$ とする．このときサンプル $i$ の個別介入効果 (ITE) は以下の式 (1) のように定義できる．

$$ITE_i = y_i^1 - y_i^0 \tag{1}$$

### 2.2. CounterFactual Regression(CFR)

介入群と対照群の共変量の分布が乖離していると，結果を補完するサンプルが存在しないために，個別介入効果を正確に推定できない．この問題に対処したのが CounterFactual Regression(CFR) である．CFR は共変量の乖離が小さく

なるような空間にデータを写像し，その空間で潜在的結果変数の予測を行う．

ここで，CFR の損失関数を定義する．サンプル $i$ の共変量を $\boldsymbol{x}_i$, 観測される潜在的結果変数を $y_i^{obs}$, 介入変数を $t_i$ とする．ただし, サンプル $i$ に介入するとき $t_i = 1$, 介入しないとき $t_i = 0$ である．特徴写像を $\Phi$, 潜在的結果変数を予測する関数を $h$, 介入群と対照群の共変量の特徴表現の分布をそれぞれ, $p_\Phi^{t=1}$,$p_\Phi^{t=0}$, 分布の乖離度を $\mathrm{IPM}(p_\Phi^{t=1}, p_\Phi^{t=0})$ と表す．モデルの学習では，$\mathrm{IPM}(p_\Phi^{t=1}, p_\Phi^{t=0})$ と潜在的結果変数の予測誤差 $L(h(\Phi(\boldsymbol{x}_i), t_i), y_i)$ を最小化する $\Phi$ と $h$ を求める．CFR の損失関数を式 (2) に表す．

$$\frac{1}{n}\sum_{i=1}^{n} L(h(\Phi(\boldsymbol{x}_i), t_i), y_i^{obs}) + \lambda\|h\| + \alpha \cdot \mathrm{IPM}(p_\Phi^{t=1}, p_\Phi^{t=0}) \tag{2}$$

ここで，$\lambda$ と $\alpha$ はそれぞれ，正則化項の重みと共変量の分布の乖離度の重みであり，これらはハイパーパラメータである．

### 2.3. 共変量シフトと重要度重み付き学習

標準的な教師あり学習では，訓練データとテストデータは入出力規則 (入力に対する出力の生成規則) と入力分布がそれぞれ同じ確率分布に従うという仮定が置かれている．しかし, 現実には訓練データとテストデータが同じ分布に従わない場合も多い．その1つに共変量シフトと呼ばれる状況がある．共変量シフトとは，入出力規則が訓練データとテストデータで変化しないが, 入力分布が訓練データとテストデータで変化するような状況である．具体的には，訓練データ $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$ が同時確率密度 $p(\boldsymbol{x}, y)$ を持つ確率分布に従い，テストデータ $\{\boldsymbol{x}_i'\}_{i'=1}^{n'}$ が確率密度 $\int p'(\boldsymbol{x}, y)dy$ を持つ確率分布に独立に従うとしたとき，共変量シフトは以下の式 (3) で表される．

$$p(\boldsymbol{x}) \neq p'(\boldsymbol{x}), \quad p(y|\boldsymbol{x}) = p'(y|\boldsymbol{x}) \tag{3}$$

$y$ を予測する関数を $f_{\boldsymbol{\theta}}(\boldsymbol{x})$，損失関数を $L(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))$ としたとき，汎化誤差は以下の式 (4) で近似される．

$$G = \iint L(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))p'(\boldsymbol{x}, y)d\boldsymbol{x}dy \tag{4}$$

共変量シフト下では，　$p(\boldsymbol{x}, y) \neq p'(\boldsymbol{x}, y)$ であるため, 訓練データの誤差を最小化しても汎化誤差 $G$ に対して最適なパラメータ $\boldsymbol{\theta}$ を求めることができない．この問題に対処した学習方法として, 重要度重み付き学習 [3] がある．重要度重み付き学習では，以下の式 (5) で表される重要度を用いることによって，式 (6) のように汎化誤差 $G$ を近似する．

$$w(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})} \tag{5}$$

$$\begin{aligned} G &= \iint L(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))p'(\boldsymbol{x}, y)d\boldsymbol{x}dy \\ &= \iint L(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))p'(y|\boldsymbol{x})p'(\boldsymbol{x})d\boldsymbol{x}dy \\ &= \iint L(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))w(\boldsymbol{x})p(\boldsymbol{x}, y)\boldsymbol{x}dy \\ &\approx \frac{1}{n}\sum_{i=1}^{n} L(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}))w(\boldsymbol{x}) \end{aligned} \tag{6}$$

### 2.4. Generative Adversarial Network(GAN)

#### 2.4.1. GAN のモデル概要

GAN は生成器と識別器の 2 つの機構からなる生成モデルである．生成器は入力ノイズから学習データに似たデータを生成する．そして，識別器に学習データか生成されたデータを入力し，識別器は入力されたデータが学習データなのか，生成されたデータなのかを識別する．生成器と識別器を交互に最適化することで，識別器が実際の学習データと区別することが難しいデータを生成器が生成できるようになる．GAN の損失関数は以下の式 (7) で表される．

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{7}$$

ここで，$\mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\cdot]$ は学習データの確率分布からサンプリングされた $\boldsymbol{x}$ のミニバッチの平均値を表し，$\mathbb{E}_{\boldsymbol{z}\sim p_{data}(\boldsymbol{z})}[\cdot]$ は生成器の確率分布からサンプリングされた入力ノイズのミニバッチの平均値を表す．また，$D(\cdot)$ は入力されたデータが学習データである確率を出力する．

#### 2.4.2. 最適な識別器

前述の通り, GAN は生成器が学習データの分布 $p_{data}(\boldsymbol{x})$ に一致するような分布 $p_g(\boldsymbol{x})$ を生成できるように最適化を行う．ここで，$p_{data}(\boldsymbol{x})$ と $p_g(\boldsymbol{x})$ の JS ダイバージェンスを考えると，以下の式 (8) のように表せる．

$$\begin{aligned} 2D_{JS} &= D_{KL}\left(p_{data}(\boldsymbol{x})\middle|\frac{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}{2}\right) \\ &\quad + D_{KL}\left(p_g(\boldsymbol{x})\middle|\frac{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}{2}\right) \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log\left\{\frac{2p_{data}(\boldsymbol{x})}{p_g(\boldsymbol{x}) + p_{data}(\boldsymbol{x})}\right\}\right] \\ &\quad + \mathbb{E}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})}\left[\log\left\{\frac{2p_g(\boldsymbol{x})}{p_g(\boldsymbol{x}) + p_{data}(\boldsymbol{x})}\right\}\right] \end{aligned} \tag{8}$$

JS ダイバージェンスは分布が近いほど値が小さくなるため，GAN は JS ダイバージェンスを最小化していると捉えることが可能である．そこで，式 (8) の期待値と式 (7) を比較すると，最適な識別器 $G$ は以下の式 (9) のように表される．

$$D^*(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_g(\boldsymbol{x}) + p_{data}(\boldsymbol{x})} \tag{9}$$

## 3. 提案手法

### 3.1. 提案手法の着想と概要

2.2 章で述べた通り，CFR は介入群と対照群の共変量の乖離を小さくする特徴表現を獲得することで，個別介入効果を推定する．しかし，共変量の乖離が大きいときに推定精度が大幅に悪化してしまう．ここで，介入群のサンプルの介入しなかった結果を得ることは難しいが，対照群のサンプルに介入した結果は介入することによって入手可能である．例えば，EC サイト上で一部のユーザにクーポンを発行していて，介入群の分布が偏っているときに，まだクーポンを発行していない対照群のサンプルにクーポンを発行することは可能である．そこで本研究では，対照群のサンプルに介入することが可能な状況下で，個別介入効果の予測精度をより向上させる対照群のサンプルを特定する手法の開発を目指す．ここで，式 (5) より，重要度は学習データに出現しにくく，テストデータに出現しやすいしやすいサンプルほど値が大きくなる．言い換えれば，テストデータの重要度は，もしそれが学習データとして入手できた際に，目的変数の推定精度の向上するにどれほど貢献するかを表すと言える．これを因果推論の問題に置き換えると，重要度は，介入群の確率密度 $p_t(\boldsymbol{x})$ と対照群の確率密度 $p_c(\boldsymbol{x})$ を使って，$p_t(\boldsymbol{x})/p_c(\boldsymbol{x})$ と表せる．これは，もし $\boldsymbol{x}$ が介入群のデータとして入手できた際に，個別介入効果の推定精度をどれほど向上させるかを表していると言える．したがって，CFR における重要度の算出が要求される．

重要度を算出する単純な方法として，訓練データとテストデータの入力密度 $p_t(\boldsymbol{x})$，$p_c(\boldsymbol{x})$ をそれぞれ推定して，その比を取る方法が考えられる．しかし，比を取ることにより，確率密度の推定誤差が増幅されてしまうことから，この方法は一般に精度が悪化してしまう [4]．そのため，密度比を直接推定する方法が好ましい．ここで，2.4.2 章で述べたように，GAN は識別器を用いた密度比の推定が検討できる．GAN の識別器に学習データの分布と生成データの分布の代わりに，介入群の共変量の分布と対照群の共変量の分布を入力し，個別介入効果の推定におけるデータの重要度を算出することが可能となる．また，GAN の識別器は，本来，学習データと生成データの識別が難しくなるように生成器を学習するために用いられるため，個別介入効果の推定問題に置き換えると，介入群のデータか，対照群のデータかを見分けることが難しくなるような特徴表現を獲得する目的で識別器を用いることができる．以上の理由から提案手法では，介入群と対照群の乖離度を測る機構として GAN の識別器を採用し，識別器から得られる密度比を用いて，対照群の中から新たに介入するサンプルを決定する．

### 3.2. 提案手法の定式化

共変量を $\boldsymbol{x}_i$，特徴写像による変換後の共変量を $\Phi(\boldsymbol{x})$ とし，介入変数を $t_i$，観測される潜在的結果変数を $y_i^{obs}$，介入群と対照群の共変量の分布をそれぞれ $p_t$，$p_c$ とする．また，潜在的結果変数を予測する関数を $h$，$y^{obs}$ と予測値の誤差を測る損失関数を $L$，GAN の識別器を $D$，共変量の分布の乖離度を調整するハイパーパラメータを $\alpha$ とする．このとき，提案手法の損失関数は式 (10) のように表される．

$$
\begin{aligned}
V = \min_{L,\Phi}\max_{D}\sum_{i=1}^{n} & L(h(\Phi(\boldsymbol{x}_i),t_i),y_i^{obs}) \\
& + \alpha\{\mathbb{E}_{\boldsymbol{x}\sim p_t}[\log D(\boldsymbol{x})] \\
& + \mathbb{E}_{\boldsymbol{x}\sim p_c}[\log(1-D(G(\boldsymbol{x})))]\}
\end{aligned}
\tag{10}
$$

ここで，$\mathbb{E}_{\boldsymbol{x}\sim p_t}[\cdot]$ は介入群からサンプリングされた共変量 $\boldsymbol{x}$ のミニバッチの平均値を表し，$\mathbb{E}_{\boldsymbol{x}\sim p_c}[\cdot]$ は対照群からサンプリングされた共変量 $\boldsymbol{x}$ のミニバッチの平均値を表す．また，提案手法のネットワーク構造のイメージを図 1 に示す．

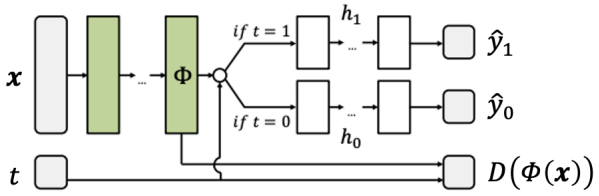


図 1. 提案手法のモデル構造のイメージ

式 (10) に基づいてモデルの最適化を行ったとき，最適な識別器は式 (11) で表される．

$$
D^*(\Phi(\boldsymbol{x})) = \frac{p_c(\Phi(\boldsymbol{x}))}{p_c(\Phi(\boldsymbol{x})) + p_t(\Phi(\boldsymbol{x}))} \tag{11}
$$

式 (11) の $D^*(\Phi(\boldsymbol{x}))$ を変形することによって，介入群と対照群の確率分布の密度比を以下の式 (12) によって表すことができる．

$$
v(\Phi(\boldsymbol{x})) = \frac{p_c(\Phi(\boldsymbol{x}))}{p_t(\Phi(\boldsymbol{x}))} = \frac{D^*(\Phi(\boldsymbol{x}))}{1-D^*(\Phi(\boldsymbol{x}))} \tag{12}
$$

この密度比 $v(\Phi(\boldsymbol{x}))$ が重要度重み付き学習における重要度を表すため，対照群の中でこの値が大きいものに介入する．新たに介入するサンプルを決定する回数を $N$ としたときの提案手法のアルゴリズムを以下に示す．

**Algorithm 1** 提案手法

損失関数 $V$ に基づきモデルを学習する
**for** $n$=1,...,$N$ **do**
・$v(\phi(\boldsymbol{x}))$ を算出する
・$v(\phi(\boldsymbol{x}))$ の値が大きい対照群のサンプルをいくつか選び，そのサンプルの潜在的結果変数 $y_1$ を獲得する
・損失関数 $V$ に基づきモデルを学習する
**end for**

## 4. 評価実験

提案手法の有効性を検証するために実データである IHDP データセットを用いた評価実験を行った．IHDP データセットは，幼児に対する特別な教育が将来のテストのスコアに与える影響について調査したデータセットであり，幼児の出生体重や世帯年収などを含む 25 個の共変量が存在し，データ数は 7,470 である．共変量の乖離を生じさせるために，上記のデータに対しロジスティック回帰を適用し，対照群の中から $\hat{p}(y|\boldsymbol{x}=1)$ が高いサンプルを 2,000 個削除した．また，実験条件は以下のように設定した．ニューラルネットワークの共変量を変換する層，潜在的結果変数を予測する層，識別器の隠れ層は全て 2 層とし，各層の次元数は 100 とした．介入群の対照群の共変量の分布の乖離度を調整するハイパーパラメータの値は 3 とし，正則化項の重みは 1 とした．比較手法として，ランダムサンプリングとカーネル密度推定を用いた．ランダムサンプリングは，対照群の中からランダムに介入対象を決定する方法である．カーネル密度推定は，対照群と介入群の確率密度をそれぞれカーネル密度推定を用いて推定し，その比を重要度として，その重要度が高いサンプルを介入対象とする方法である．また，毎回のサンプリングで介入するサンプルを 100 個選択した．評価指標として以下の式 (13) で表される PEHE を用いた．

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum\left\{(\hat{y}_i^0 - \hat{y}_i^1) - (y_i^1 - y_i^0)\right\}^2} \quad (13)$$

ここで，$\hat{y}_i^0, \hat{y}_i^1$ はそれぞれ，$y_i^1, y_i^0$ の予測値である．実験結果を以下の図 2 に示す．

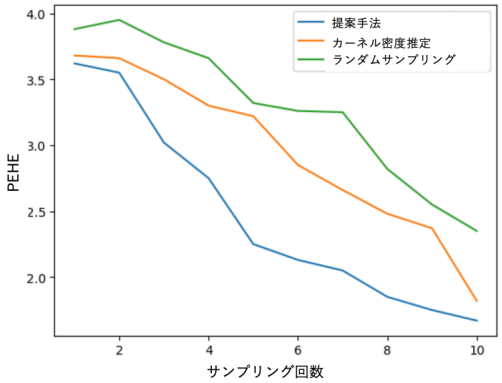


図 2. PEHE の推移

図 2 より，提案手法はランダムサンプリングよりも PEHE が低い．このことから，重要度を用いたサンプリング方法が個別介入効果の推定精度をより向上させるサンプルを特定する方法として優れていることが分かる．また，同じ密度比推定方法であるカーネル密度推定よりも PEHE が低いことから，提案手法は密度比推定手法としてより優れていることが分かる．

## 5. 考察

評価実験では，同じ密度比推定方法でも提案手法とカーネル密度推定の PEHE の差が大きかった．これは IHDP データセットの共変量が 25 次元と高次元であることが関係していると考えられる．カーネル密度推定は，各データ点を中心に極小のカーネルを無数に配置することで滑らかな分布推定を実現しているため，学習データが十分密集して存在していることを暗に仮定している．しかし，高次元ではデータ同士の距離が大きくなり，データの分布が疎になるため，分布推定の精度が低下する．一方，GAN はニューラルネットワークを活用しているため，識別器も高次元のデータに対応している．そのため，GAN の識別器を用いた提案手法は，データの複雑さや次元に対する頑健性が高く，共変量が高次元の IHDP データセットにおいても，より正確に学習が行えていると考えられる．また，ランダムサンプリングやカーネル密度推定によるサンプリングでは，PEHE の現象具合が比較的単調であるに対して，提案手法の PEHE は 2 回目のサンプリングと 5 回目のサンプリングの間で値が大きく減少している．これは，提案手法が比較的初期のサンプリングで予測精度を向上させる少数のサンプルを見つけることに成功していることを示している．

## 6. 結論と今後の課題

本研究では，対照群のサンプルに新たに介入できる状況下で，介入することで個別介入効果の予測精度をより大きく向上させるような対照群のサンプルを特定する手法を提案した．また，実データを用いた評価実験を通して提案手法の有効性を示した．提案手法では，共変量の密度比に基づいて新たに介入するサンプルを決定したため，潜在的結果変数は全く考慮していない．しかし，観測できていない潜在的結果変数の情報量は一律ではないと考えられるので，潜在的結果変数の情報量を考慮することでさらに予測精度を向上させることができる可能性がある．そのため，今後の課題としては，潜在的結果変数の情報量，分散も考慮した手法を開発することが挙げられる．

## 参考文献


[1] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, Vol. 27, , 2014.

[3] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, Vol. 90, No. 2, pp. 227–244, 2000.

[4] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.