2D-03

意図的代入法における最適代入値の理論的解析

福島 卓弥† 長谷川 拓† 中島 智晴†
†大阪府立大学大学院 人間社会システム科学研究科

1 はじめに

多くの機械学習手法では，データに欠損値がないことが前提となっている．欠損値がないことは機械学習において理想的な状況である．しかし，このような状況は実問題では起こりにくい．例えば医療診断では，測定機器のエラーや故障により値が得られない可能性があり，このような場合医師は測定値がないまま適切な診断をする必要に迫られる．

多くの欠損値に関する研究では，学習用データと未知データともに欠損があることが仮定される．一方で，実世界では学習時には十分な情報量が利用可能であるが，実際の場面では学習時ほど情報が得られないことが多く存在する．そのため本研究では，学習用データには欠損値がなく，未知データには欠損値がある場合を考えることにする．学習時に意図的代入法を用いることで，欠損値にロバストなモデルを獲得することを目的とする．

2 意図的代入法

本研究では，モデル学習時には欠損値が含まれない学習用データを用いることができるが，欠損値を含む可能性がある未知データを用いる状況を仮定する．また，未知データに対する予測時にどの特徴量で欠損が発生しうるかをあらかじめ知っているものとする．意図的代入法とは，これらの状況下において有効な，欠損値が生じる状況に対するモデル学習法である [1]．なお，欠損値は値がわからない，入手不可能などの理由により「値がない」ことを表し，値がゼロであることとは別であることに注意する．

学習用データを意図的代入法により修正する．意図的代入法では，学習用データに欠損がなく，値が利用可能であったとしてもそれを使用せずに強制的に値を置換する．また，意図的代入は常に行うものではなく，予め決められた確率に従って行うものとする．以下に，意図的代入法の手続きを示す．この手続きをミニバッチ方式に拡張してモデル学習に適用するには，手続きをバッチサイズだけ繰り返せばよい．

Step 1: 学習用データから入力ベクトルを一つ抽出する．

Step 2: 予め決められた確率に従って，欠損するとわかっている特徴量をある値と置換する．

Step 3: 修正された入力ベクトルと教師信号を用いてモデルを学習する．

3 意図的代入法における最適代入値

3.1 最適代入値の理論的解析

理論的解析にあたり実験設定に以下の仮定をおく．

仮定 1 真の関数 f がわかっている．

仮定 2 予測モデル g は真の関数 f を表現可能である．

仮定 3 各次元は独立であり，次元数は N である．また欠損は第 N 次元目にのみ確率 p_{mis} で起こる．

誤差を $\delta(f(\boldsymbol{x}), g(\boldsymbol{x})) = \{f(\boldsymbol{x}) - g(\boldsymbol{x})\}^2$ とする．このとき，誤差の期待値は

$$\begin{aligned}\mathbb{E}[\delta(f(\boldsymbol{x}), g(\boldsymbol{x}))] = \\ (1 - p_{\mathrm{mis}}) \int\cdots\int_{D_X} p(\boldsymbol{x})\{f(\boldsymbol{x}) - g(\boldsymbol{x})\}^2 dX \\ + p_{\mathrm{mis}} \int\cdots\int_{D_X} p(\boldsymbol{x})\{f(\boldsymbol{x}) - g(\psi^s(\boldsymbol{x}))\}^2 dX\end{aligned} \tag{1}$$

とかける．ここで，$\psi(\cdot)$ は欠損のある入力に対して，代入値を定める関数であり，$\boldsymbol{s}$ は欠損が起きているかどうかを表す確率変数の集合である．式 (1) から，誤差の期待値を最小化するためには $\{f(\boldsymbol{x}) - g(\psi^s(\boldsymbol{x}))\}^2$ を最小化する $\psi(\cdot)$ を求めればよい．$\psi(\cdot)$ を求めることができれば，意図的代入法において，その代入値が有効に働くことが実験的に示されている [1]．

仮定 1 は，常に成り立たないため，実際に誤差の期待値を最小化するような $\psi(\cdot)$ を算出することは不可能である．そこで本論文では，仮定 1 について考え，学習用データから $\psi(\cdot)$ を推定する手法を提案する．

3.2 最適代入値の推定手法

以下の手順で $\psi(\cdot)$ を推定する．表記を簡単にするため $N = 2$ とする．

Theoretical Analysis of Optimal Values for Intentional Substitution Methods
†Takuya FUKUSHIMA †Taku HASEGAWA †Tomoharu NAKASHIMA
†Graduate School of Humanities and Sustainable System Science, Osaka Prefecture University

Step 1: 欠損のない次元（1次元目）の定義域を n 等分する．このときの端点を $(x_{11}, \ldots, x_{1n})$ とする．

Step 2: 区間 $(x_{1i}, x_{1(i+1)}]$ に存在する T 個の学習用データについて $y_{ave} = \frac{1}{T}\sum_{t=1}^{T} y^t$ を求める．ここで y^t は学習用データ $\boldsymbol{x}^t$ の目的変数である．

Step 3: $t' = \arg\min_t (y_{ave} - y^t)^2$ とするとき，区間 $(x_{1i}, x_{1(i+1)}]$ における推定代入値を $x_2 = x_2^{t'}$ と定める．

Step 4: Step 2, 3 を全区間分（$n-1$ 回）行う．

4 数値実験

本章では，3.2 節で示した意図的代入法における最適代入値の推定手法の性能を調査する．

4.1 実験設定

実験では，2 次元の入力ベクトルに対して，未知入力における欠損が確率 p_{mis} で発生すると仮定する．意図的代入は確率 p_{sub} で行うものとする．一般性を失うことなく，欠損は最後の次元で発生すると設定する．学習モデルとして，3 層階層型ニューラルネットワークを用いる．機械学習のタスクを回帰とし，独立変数は多次元実数値ベクトル，従属変数を実数値とする．学習用データとして，定義域から一様乱数により入力を生成して独立変数とし，予め決められた関数の出力値を回帰モデル学習における教師信号（従属変数）とする．ニューラルネットワークは，以下のハイパーパラメータを用いて学習される．活性化関数はシグモイド関数を用いる．中間層のユニット数は 50 とする．重み調整の最適化手法には Adam を用いる．ベンチマーク関数には以下のものを用いる．

$\boldsymbol{f}$ （2 次元 2 次関数）:

$$f(\boldsymbol{x}) = (x_1 - x_2)^2, \qquad (-5 < x_1, x_2 < 5) \tag{2}$$

入力が独立かつ一様に生成されると仮定すると，欠損値がある場合に予測誤差を最小化するような代入値は以下となる [1].

$$\begin{aligned} &\psi_2'(x_1) \\ &= \arg\min_{x_2} \{ \int_{-5}^{5} \frac{1}{10}(x_1 - x_2)^2 \, dx_2 - (x_1 - x_2)^2 \}^2 \\ &= x_1 \pm \sqrt{x_1^2 + \frac{25}{3}} \end{aligned} \tag{3}$$

4.2 実験結果

図 1 に，x_1 を 22 分割した際の推定手法の結果を示す．細線は式 (3) で表される最適代入値，太線はその推定値を表す．また，推定代入値を用いて確率 $p_{sub} = 0.75$ で学習を行った際の予測誤差を図 2 に示す．推定代入値（Estimation）は，最適代入値（Theory, Theory random）と同等な性能で学習できていることがわかる．このことから，本推定手法で最適代入値を十分に表現可能であるといえる．真の関数 f が分からない状態において，推定代入値を用いることが欠損値を含むデータに対して有効に働くことが示された．

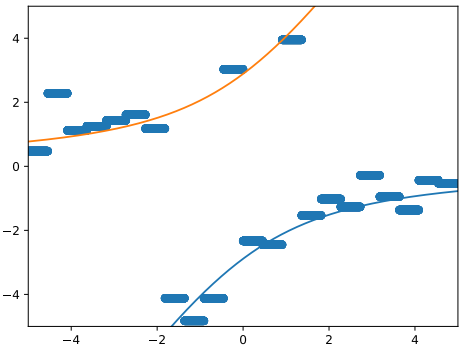

図 1: Estimation of the optimal substitution value

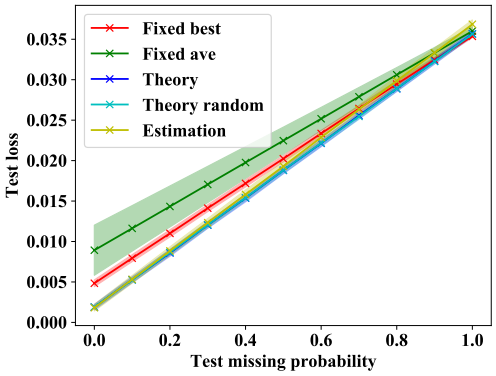

図 2: Comparison of the test error among difference methods of the intentional substitution ($p_{sub} = 0.75$)

5 おわりに

本研究では意図的代入法における最適代入値の推定手法を提案した．数値実験の結果から，最適代入値を推定することで未知データの欠損に対してロバストなモデルの妥当性を示すことができた．今後の課題として，ノイズの多い環境における欠損値の取り扱いに対して，本研究の知見を生かすことが考えられる．

参考文献

[1] 福島卓弥, 長谷川拓, 中島智晴. 欠損値のない学習用データが利用可能な場合における欠損値に対してロバストな学習法. 第 14 回コンピューテーショナル・インテリジェンス研究会論文集, pp. 47–53, 2018.