

# 修士論文概要書

Master's Thesis Summary

Date of submission: 01/10/2024 (MM/DD/YYYY)

専攻名 (専門分野) Department	経営システム 工学専攻	氏名 Name	竹内 瑞生 Mizuki Takeuchi	指導 教員 Advisor	後藤 正幸 印 Seal
研究指導名 Research guidance	情報数理応用研究	学籍番号 Student ID number	5222C019-7		
研究題目 Title	Self- and Semi-supervised Learning に基づく行動履歴データに対する分析モデルに関する研究 A Study on Analysis Model for Action History Data Based on Self- and Semi-supervised Learning				

## 1. はじめに

伝統的なマーケティングの領域では、顧客の住所や年齢、性別といった属性情報をもとに形成したセグメントを活用するセグメントマーケティングが行われている。近日においては、インターネット上で展開される EC サイトでセグメントマーケティングが利用されることがある。日本国内における EC サイトの市場規模は年々増加傾向にあり、EC サイトの多くで新規顧客獲得のために、クーポンを配布するといった施策が行われている。このとき、セグメントを活用することで施策を行うユーザの抽出が可能となり、より有用な施策立案が可能となる。

また、多くの EC サイトにおいて、住所や性別、年齢といった属性情報を登録することで会員ユーザとして利用する方法と、属性情報を登録せず非会員ユーザとして利用する方法の 2 つの利用方法がある。会員ユーザに対して施策を行う場合、登録されている属性情報からセグメントを形成することで、特定の属性を持つユーザをターゲットにすることができる。一方で、非会員ユーザに対しては、属性情報を得ることができないため、属性をもとにしたセグメントを形成できない。そこで機械学習モデルを用いて、ユーザの購買や閲覧といった行動履歴から非会員ユーザの属性予測を行い、セグメント形成に利用するセグメント拡張が解決策として考えられる [1]。このとき、会員ユーザよりも多く存在する非会員ユーザのデータを予測に活用する半教師有り学習の学習アプローチを用いることで、高精度な予測が可能と考えられる。

半教師有り学習を行う手法はこれまでに数多く提案されているが、その 1 つに Value Imputation and Mask Estimation (VIME) [2] がある。VIME は表形式データに特化した事前タスクを解くことで自己教師有り学習を行う機構を含む半教師有り学習モデルである。そのため、表形式データの 1 種としてみなすことができる行動履歴データについても、この手法が適用できると考えられる。しかし、一般的な表形式データと比較して EC ユーザの行動履歴データは特徴量中の空値割合が高くスパースであるとい

う特徴があり、そのまま適用を行った場合、モデルの学習が不安定になり十分な性能を発揮できない恐れがある。

そこで本研究では、非会員ユーザに対する属性情報のより正確な予測を可能にするため、従来手法である VIME に対して、スパース性を考慮したモデルを提案する。また、実際のユーザ行動履歴データを用いた評価実験によって、提案手法の有効性を示す。

## 2. 準備

この章では、本研究で対象とする問題設定と、従来手法である VIME について説明する。

### 2.1. 対象とする問題設定

属性情報のある会員データよりも属性情報のない非会員データの方が一般に多く存在するため、本研究では正解ラベルが付与されている教師有りデータに加えて、正解ラベルが付与されていない教師無しデータが潤沢に利用可能であるとする。また、この状況は半教師有り学習の領域において一般的に設定されている状況と同様である。

### 2.2. VIME

#### 2.2.1. 概要

自己教師有り学習とは、データセットの特徴量を利用した事前タスクとその正解ラベルを機械的に生成し、モデルの適用対象の問題に最適化させる前に事前タスクを解くことで、適用対象の問題に対する汎化性能の向上を図る機械学習モデルのフレームワークである。VIME はニューラルネットワークを利用した機械学習モデルであるが、表形式データに対して有効に機能する自己教師有り学習を取り入れている。VIME の学習は大きく分けて 2 段階からなっており、自己教師有り学習によってエンコーダーから獲得できる特徴量表現を最適化したのちに、半教師有り学習によって対象とする問題に向けた予測器の最適化が行われる。モデルがテストデータから予測ラベルを出力するイメージを図 1 に示す。

#### 2.2.2. Pretext Generator

VIME では、2 つの学習段階の両方に Pretext Generator という機構が含まれている。この機構においては、

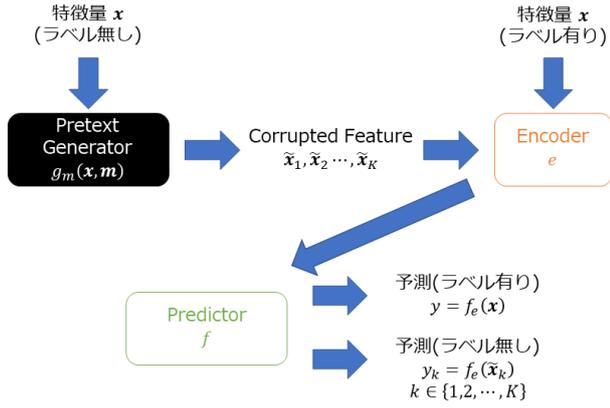


図 1. モデルの概要

$1 - m$ ,  $m$  の 2 つマスクベクトルとデータ  $x$ ,  $\bar{x}$  とのアダマール積を取り, これらの要素和を取ることによって, 元のデータの特徴量のうち一部をランダムに別の特徴量に置き換えた Corrupted Feature が生成される. 図 2 に Pretext Generator の仕組みを示す.

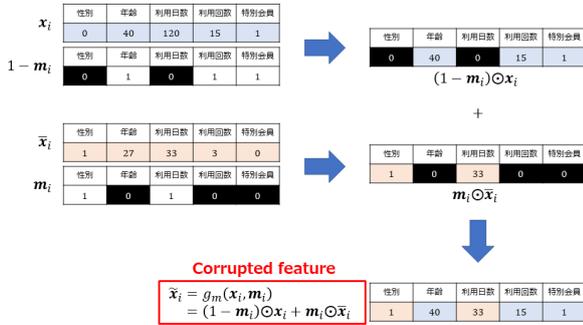


図 2. Pretext Generator における Corrupted Feature の生成過程

データの特徴量を  $x$  としたとき, マスクベクトル  $m$  と Pretext Generator  $g_m(x, m)$  により定義される Corrupted feature  $\tilde{x}$  は以下の式 (1)–(3) で表される.

$$\tilde{x} = g_m(x, m) = m \odot \bar{x} + (1 - m) \odot x \quad (1)$$

$$m = [m_1, m_2, \dots, m_d]^T \in \{0, 1\}^d \quad (2)$$

$$m_j \sim \text{Bern}(p_m), \quad j \in \{1, 2, \dots, d\} \quad (3)$$

ここで,  $\odot$  はアダマール積,  $\bar{x}$  は経験的分布よりサンプリングされる特徴量ベクトル,  $p_m$  はベルヌーイ分布の確率を表わすハイパーパラメータである.

### 2.2.3. エンコーダーの自己教師有り学習

1 段階目の学習として, Corrupted Feature  $\tilde{x}$  から元の特徴量  $x$  とマスクベクトル  $m$  を復元するタスクによって, 潜在表現  $z$  を出力するエンコーダー  $e$  の最適化を行う.

推定される特徴量ベクトルとマスクベクトルをそれぞれ  $\hat{x}$ ,  $\hat{m}$  としたとき, 元の特徴量ベクトルとマスクベクトルとのそれぞれの復元誤差, および自己教師有り学習のための誤差関数は以下の式 (4)–(6) で表される. そしてエンコー

ダー  $e$  は誤差逆伝播法によって式 (6) を最小化することで最適化される.

$$l_r(x, \hat{x}) = \frac{1}{d} \left[ \sum_{j=1}^d (x_j - (s \circ e)(\tilde{x}))^2 \right] \quad (4)$$

$$l_m(m, \hat{m}) = -\frac{1}{d} \left[ \sum_{j=1}^d m_j \log [(s_m \circ e)_j(\tilde{x})] + (1 - m_j) \log [1 - (s_m \circ e)_j(\tilde{x})] \right] \quad (5)$$

$$\min_{e, s_m, s_r} \mathbb{E} [l_m(m, \hat{m}) + \alpha \cdot l_r(x, \hat{x})] \quad (6)$$

ここで,  $s_r, s_m$  はエンコーダーから出力される潜在表現  $z$  から特徴量ベクトルとマスクベクトルをそれぞれ復元するための関数,  $\circ$  は合成関数,  $\alpha$  はハイパーパラメータである. また, 自己教師有り学習におけるタスクのイメージを図 3 に示す.

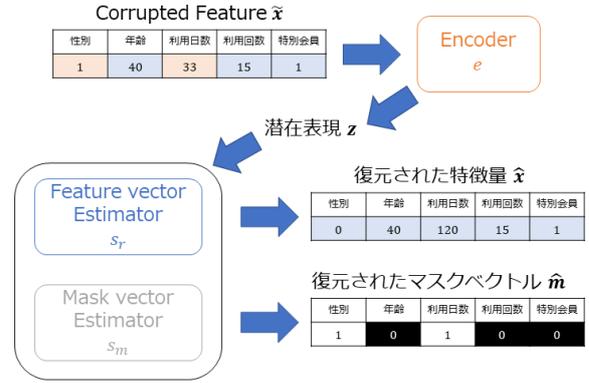


図 3. 自己教師有り学習のイメージ

この学習によって, エンコーダーが出力する潜在表現  $z$  が最適化され, 実際のタスクにおいて予測が行いやすくデータの特徴が的確に反映された潜在表現を得られるようになる.

### 2.2.4. 予測器の半教師有り学習

2 段階目の学習として, 回帰や分類といった実際のタスクに対して最適化するために, 予測器の半教師有り学習を行う. 実際のタスクの正解ラベルを  $y$  としたとき, 教師有り学習と教師無し学習それぞれの誤差関数  $\mathcal{L}_s, \mathcal{L}_u$ , およびそれらを用いて定義される最終的な誤差関数  $\mathcal{L}_{final}$  は以下の式 (7)–(9) で表される.

$$\mathcal{L}_s = \mathbb{E} [l_s(y, (f \circ e)(x))] \quad (7)$$

$$\mathcal{L}_u = \frac{1}{N_b K} \sum_{i=1}^{N_b} \sum_{k=1}^K [((f \circ e)(\tilde{x}_{i,k}) - (f \circ e)(x_i))^2] \quad (8)$$

$$\mathcal{L}_{final} = \mathcal{L}_s + \beta \cdot \mathcal{L}_u \quad (9)$$

ここで,  $f$  は実際のタスクに最適化する予測器,  $l_s$  は事前に決められた教師有り学習のための損失関数,  $N_b$  はバッチサイズ,  $\beta$  はハイパーパラメータである. 本論文では,  $l_s$  には交差エントロピーを用いる. また, ラベルの無

いデータに対して、Pretext Generator を利用してデータ拡張を行うことで、学習に利用するデータ数を増加させている。式 (8) の  $K$  はこのデータ拡張において、ラベル無しデータ 1 件に対して生成される Corrupted Feature の件数である。

ラベル有りデータによる教師有り学習は、ラベルとの誤差を計算することによって、予測精度の向上に寄与する。ラベル無しデータによる教師無し学習は、データ拡張を行ったデータ同士での予測ラベルが同じになるようにすることで、予測の一貫性を担保すると共に、未知のデータに対する汎化性能の向上に寄与する。

### 3. 提案手法

#### 3.1. 提案の着想

VIME は表形式データに対して提案された自己教師有り学習と半教師有り学習を行う手法である。その提案論文である [2] を含めて、評価実験において行動履歴データに適用した例が無いため、有効性の検証を行ったところ、VIME が持つパラメータの中でも Pretext Generator におけるベルヌーイ分布のパラメータ  $p_m$  がテストデータに対する汎化性能を特に大きく左右することが明らかとなった。このパラメータは大きければ大きいほど、Pretext Generator において特徴量の置き換えが発生する確率が高くなる。また、同じ条件での実験でも、テストデータに対する精度のばらつきが大きくなることも分かった。これらの結果は、行動履歴データのスパース性に起因するものと考えられる。

一般的な表形式データと比較して行動履歴データはスパースであるため、Pretext Generator において行われる特徴量の置換により値が変化する可能性が低くなる。これは事前タスクの難易度が高くなることでエンコーダーが出力する潜在表現がうまく最適化されないだけでなく、データ拡張の際に特徴量の置換が有効に機能せず元のデータと同じようなデータが多く複製され、有用なデータ拡張が行えないためである。また、既に値が入っている部分が置換される場合に空値になる可能性が高くなり、もともと高かったスパース性がさらに高くなることにも繋がる。その結果として 1 インスタンスが持っていた情報量の喪失を引き起こし、学習が不安定になり、実験結果のばらつきが大きくなることに繋がった可能性が考えられる。

以上の課題を解決するため、本研究では VIME を改良したモデルとして、Pretext Generator における特徴量操作に関して、行動履歴データのスパース性を考慮した操作を加えてモデルの学習に利用するモデルを提案する。具体的には、学習データの空値でない特徴量の割合を求め、ランダムに特徴量にバイナリノイズを付与することで、一部の特徴量の値を 0 または 1 に置換する。これにより Pretext Generator における特徴量の置換の発生が確実になることが期待され、データ拡張の有効性の向上、エンコーダーと予測器の学習が適切に行われることを促す。ここで、提案手

法における Pretext Generator のイメージを図 4 に示す。

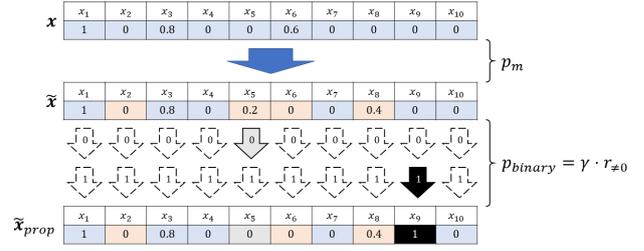


図 4. 提案手法における Pretext Generator のイメージ

#### 3.2. Pretext Generator の拡張

教師有りデータと教師無しデータの両方を含めた全学習データの全特徴量における、空値ではない特徴量の割合を  $r_{\neq 0}$  とする。各特徴量におけるバイナリノイズの付与確率  $p_{binary}$  を、ハイパーパラメータである  $\gamma$  を用いて以下の式 (10) により定義する。

$$p_{binary} = \gamma \cdot r_{\neq 0} \quad (10)$$

すなわち、提案手法の Pretext Generator では、これまでの Pretext Generator で行っていた特徴量置換を行ったあと、各特徴量を  $p_{binary}$  の確率で特徴量を 0 または 1 に上書きする処理を加える。そして、このように生成された新たな  $\tilde{x}_{prop}$  を自己教師有り学習と半教師有り学習のそれぞれの段階において用いる。

### 4. 評価実験

提案手法の有効性を確認するため、評価実験を行った。

#### 4.1. 実験設定

データセットには Movielens 1M Dataset [3] を行動履歴データと見立てて用い、各ユーザが評価した各映画の評価値をもとに、ユーザの年代区分 (7 クラス) と性別 (2 クラス) を予測する 2 つのタスクを各モデルに解かせる。各アイテムへの評価値は 1 から 5 までの整数値として与えられているが、特徴量の値域が  $[0, 1]$  になるように正規化した。比較手法として、ロジスティック回帰, XGBoost [4], 多層パーセプトロン, 従来の VIME を用いる。また、評価指標には Accuracy を用いる。テストデータのサイズを  $N_{test}$ , 正解ラベルと予測ラベルが一致したデータの件数を  $N_{y_i=\hat{y}_i}$  とすると、Accuracy は以下の式 (11) のように定式化される。

$$Accuracy = \frac{N_{y_i=\hat{y}_i}}{N_{test}} \quad (11)$$

学習用データセットは 5,000 件を教師無しデータ、500 件を教師有りデータとして扱い、VIME を除く 3 つの従来手法では教師有りデータのみを用いて学習を行った。

提案手法において、ベルヌーイ分布のパラメータを  $p_m = 0.3$ , 教師無しデータに対するデータ拡張の回数を  $K = 10$ , バッチサイズを  $N_b = 100$ , 式 (6), (9), (10) で用いられるハイパーパラメータを  $\alpha = 2, \beta = 1, \gamma = 0.2$  と設定した。

## 4.2. 結果

表1と表2は、各手法で10回ずつ実験を行って得られた Accuracy の結果を示したものである。これらの表から、評価指標の平均値について提案手法が最も高くなっており、それがある程度小さい標準偏差で達成されていることが分かる。

表 1. 年代 (7 クラス) 分類の実験結果

手法	Accuracy (平均 ± 標準偏差) [%]
ロジスティック回帰	40.9 ± 2.3
XGBoost	37.3 ± 1.7
多層パーセプトロン	40.1 ± 2.4
VIME ( $p_m = 0.3$ )	40.1 ± 4.5
提案手法	41.3 ± 2.3

表 2. 性別 (2 クラス) 分類の実験結果

手法	Accuracy (平均 ± 標準偏差) [%]
ロジスティック回帰	76.8 ± 2.1
XGBoost	75.0 ± 1.9
多層パーセプトロン	76.6 ± 2.4
VIME ( $p_m = 0.3$ )	77.7 ± 2.0
提案手法	77.8 ± 1.8

## 5. 考察

### 5.1. モデルの動作特性に関する考察

本節においては、年代分類タスクにおける1回分の実験結果を取り上げ、予測の特徴について詳しく調査を行った。テストデータの正解ラベルと各手法のテストデータに対する予測ラベルの各クラスの件数の割合を図5に示す。

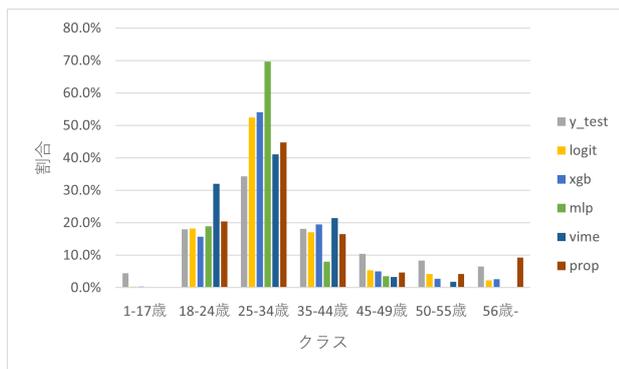


図 5. テストデータと各手法の予測ラベルの各クラスごとのデータ割合

図5から、従来のVIMEが18-24歳のクラスに多くのデータを分類していること、教師有り学習モデル3種が25-34歳に多くのデータを分類していることが分かり、従来手法はデータセットにおける割合が多いクラスに多くのデータを分類していることが考察される。一方で、提案手法はそれらのクラスに対して過度に多くのデータを分類することなく、教師有りデータのクラスの割合と大きく変わらない予測クラスの分類割合で高い精度を達成している。

このことから、Pretext Generator においてスパース性を考慮した機構を追加することで、事前に行う自己教師有り学習が行動履歴データに特化され、クラス間のデータ量の差に影響を受けにくい適切な分類が可能になった。加えて、半教師有り学習におけるデータ拡張においては、汎化性能の向上に有用に作用するデータの生成が安定したことで、結果の分散の低減効果が得られたことも併せて考察できる。

### 5.2. 波及効果に対する考察

本研究においては上限値と下限値が定まっている評価値データを用いてユーザ属性を予測する評価実験を行った。ハイパーパラメータの流用が容易になるという観点から、提案手法の適用に際しては事前にデータの正規化を行うべきである。よって、本研究において扱ったような上限値と下限値が明示的に定まっているデータに対しては提案手法を積極的に利用することができると考えられる。

一方で、閲覧や購買の回数といった、上限値の定まっていないデータに対する提案手法の適用に関しては、適切な正規化の方法という観点から課題として残った。学習データから各特徴量に対して機械的に、かつ特徴量が持つ性質を残すことができるような正規化を設定することができれば、提案手法は様々な行動履歴データに対して汎用的に利用可能なモデルになるものと考察される。

## 6. 結論と今後の課題

本論文ではユーザ行動履歴データを適用対象としてVIMEを改良したモデルを提案し、評価実験を通してその有効性を示した。提案手法では、スパース性の高い行動履歴データに対しても安定した予測を可能にするために、行動履歴データに特化した機能を追加し利用した。これにより、より正確で安定性の高い属性予測が可能になり、効果的なセグメントの絞り込みが可能になった。

今後の課題として、その他のドメインに対する実験や、それに伴うデータに対する事前処理の設定方法の立案が挙げられる。

### 参考文献

- [1] 齋藤優太, 安井翔太, 株式会社ホクソエム. 施策デザインのための機械学習入門〜データ分析技術のビジネス活用における正しい考え方. 技術評論社, 2021.
- [2] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 11033-11043, 2020.
- [3] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, Vol. 5, No. 4, pp. 1-19, 2015.
- [4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794, 2016.