G-20

ニューラルネットワークを用いた ID-POS データの非会員顧客の性別と年齢の推定

Estimation of Gender and Age of Non-member Customers in ID-POS Data Using Neural Networks

井口 拓己 [1]　田井 紗瑛子 [2]　吉野 孝 [3]　貴志 祥江 [2]　坂本 明一 [2]　宮崎 裕之 [2]　大西 剛 [2]
Takumi Iguchi　Saeko Tai　Takashi Yoshino　Sachie Kishi　Akikazu Sakamoto　Hiroyuki Miyazaki　Takeshi Onishi

1. はじめに

近年，スーパーマーケットにおけるポイントカードの導入率は 82.6%であり，様々な小売店が独自のポイントカードを導入している [1]．ポイントカードを導入することにより，性別や年代などの顧客情報をレシートの情報と紐付けることが可能であり，そのようなデータを ID-POS データと呼ぶ．つまり，ID-POS データを生成するためには，性別や年代などの顧客情報を含むポイントカードが必要である．店舗が ID-POS データを生成・収集し，利活用することでマーケティングへの応用が期待できる．簡単な利用例として，商品購入者の年齢層を分析することで，特定の年代に合わせた商品の販売促進が可能である [2]．また，応用的な利用例として，ID-POS データの分析により，顧客を逃さない継続的な来店による客単価の増加などのマーケティングが可能である [3]．

しかし，ポイントカードの利用をメリットと感じず作成しない顧客や，取り出すのが面倒で提示しない顧客など，ID-POS データの中には非会員と分類される顧客が一定数存在する [4]．また，2019 年 10 月に実施されたキャッシュレス・ポイント還元事業の影響で，小売店独自のポイントカード以外の共通ポイントカード (楽天ポイント[*1]，Ponta[*2] など) を利用する顧客が増加した [5]．既に述べた通り，ID-POS データを生成するためには，性別や年代などの顧客情報を含むポイントカードが必要である．しかし，店舗は共通ポイントでは顧客情報を収集することができず，ID-POS データを生成することができない．そのため，ID-POS データの中には，年齢や性別などの顧客情報が欠損したデータが存在し，十分に活用できないといった問題がある．実際，本研究で使用するデータでは，A 店舗の 2021 年 9 月における非会員顧客のレシート枚数は，全レシート枚数中約 20%である．

そこで本研究では，ID-POS データの価値をさらに向上させるため，会員顧客のデータを利用して，顧客の性別や年齢の情報を推定することを目的とした分析を行う．

2. 関連研究

ID-POS データを分析し，マーケティングに活用した研究がある．土井らは，ニューラルネットワークを用いて，ID-POS データと調査モニタによるアンケート結果から，ライフスタイルを推定する研究を行った [6]．消費者のライフスタイルを理解することによって，より高度な One to One マーケティングの実現を狙った．中村らは，ID-POS データを活用し，店内のレイアウトの改善や品揃えの改善につながる客動線分析を行うためのシミュレーションシステムを開発した [7]．このシステムは顧客の購買行動をシミュレーションすることで購買経路を探索し，購買経路の結果を可視化することで客動線分析を机上でより簡潔に行うことを目的としている．ID-POS データを用いることで，マーケティングへの活用を狙った様々な分析や検証が可能であるが，ID-POS データのうち，顧客情報が含まれない ID-POS データは十分に利用できていない．本研究では，これらの研究やマーケティングにおいて，ID-POS データをより活用可能にすることを目的とした研究を行う．

ID-POS データの分析手法に機械学習を用いた研究がある．鈴木は，ID-POS データと AI，ディープラーニングの親和性を調査した [8]．この研究では，従来 ID-POS データが統計学的な手法で活用されていることを踏まえ，AI，ディープラーニングの活用により，ID-POS データがどのように活用できるかを提起している．辛らは，高級焼肉店の POS データに対して，機械学習 (クラスター分析) を行うことで，顧客の購買行動の特徴を分析した [9]．分析の結果，購買顧

レシートID	日付	時刻	商品名	個数	金額
A	8/29	9時	チョコ	3個	55円
			ガム	1個	110円
			クッキー	5個	30円
Z	9/20	23時	いちご	2個	500円
			メロン	1個	3000円

図 1: POS データ

レシートID	商品_1	…	商品_X	9時	…	23時	月曜日	…	日曜日
A	3	…	0	0	…	9	9	…	0
B	0	…	1	1	…	0	0	…	1
Z	2	…	1	0	…	3	0	…	3

図 2: 1 次元データ

レシート A	月曜_6時	…	月曜_23時	火曜_6時	…	火曜_23時	…	日曜_23時
商品_1	0	…	3	0	…	0	…	0
商品_2	0	…	1	0	…	0	…	0
商品_X	0	…	0	0	…	0	…	0

図 3: 2 次元データ

1 和歌山大学大学院 システム工学研究科，Graduate School of Systems Engineering, Wakayama University
2 株式会社オークワ，Okuwa Co., Ltd.
3 和歌山大学 システム工学部，Faculty of Systems Engineering, Wakayama University

[*1] https://pointcard.rakuten.co.jp/
[*2] https://point.recruit.co.jp/point/

表 1: ハイパーパラメータ

ハイパーパラメータ	設定値
学習回数	500 epochs
損失関数	Categorical Cross Entropy
活性化関数	ReLU(隠れ層)，Softmax(出力層)
バッチサイズ	32
最適化アルゴリズム	Adam

客数別の注文特徴や，顧客構成別の注文特徴を明らかにした．小紫らは，スーパーマーケットにおける ID-POS データに対して，機械学習 (ランダムフォレスト) を行い，ライフスタイルに関する質問結果と合わせた分析を行った [10]. この研究では，ID-POS データからライフスタイルの推定を目的に分析が行われた．これらの研究は ID-POS データを用いて機械学習を行うことで，マーケティング支援を行うことを目的としている．しかし，本研究では直接マーケティングを支援するのではなく，ID-POS データの価値を向上させることを目的としている．また，これらの研究を含む従来の研究は，本研究で提唱する 1 次元データの POS データを用いて分析を行っている．本研究では，ID-POS データを別の方法で加工したデータも用いて分析を行い，各機械学習モデルとの比較検証を行う．

3. 分析概要

3.1 使用データ

本研究で使用するデータは，株式会社オークワ[*3]で収集された，2021 年 8 月 21 日から 2021 年 12 月 20 日[*4]における，和歌山県内にある 2 店舗 (A 店，B 店) と愛知県の C 店，岐阜県の D 店の合計 4 店舗を使用する．

POS データとは，商品がレジで購入されるときのデータであり，商品名，日時，個数，金額，店舗名などの情報が含まれ，図 1 のような形式である．レシートと POS データの各行が対応しており，1 行が 1 つの商品における購買情報を表している．また，ID-POS データとは，顧客の会員情報が POS データと紐づき,『誰が』購入したのかが分かるデータである[*5].

本研究では，ID-POS データで利用可能なデータのうち商品名，日時，個数，店舗名を使用する．本研究で分析対象とする店舗では，商品は部門・AU・ライン・クラスという 4 つの順に階層構造で分類され，約 1,600 種類である[*6]. 本研究では，商品の分類がある程度の粒度であれば十分であることと，データが膨大になることも考慮し，ラインまでの分類 (292 種類) で商品を扱った．

[*3] https://www.okuwa.net/

[*4] 本研究では，8 月 21 日から 9 月 20 日までを 9 月，9 月 21 日から 10 月 20 日までを 10 月，10 月 21 日から 11 月 20 日までを 11 月，11 月 21 日から 12 月 20 日までを 12 月として扱う．

[*5] 匿名加工されたデータである．

[*6] 例えば牛肉の場合，ラインまでの分類では「輸入牛肉のしゃぶしゃぶ」となり，クラスでは「輸入牛肉のしゃぶしゃぶの各部位」となる．また，果物の場合，ラインまでの分類では「果物のぶどう」となり，クラスでは「果物のぶどうの各品種」となり，部門からクラスにかけて詳細な分類が行われる．

表 2: ID-POS データの内訳

店舗	月	性別 (人)		年齢層 (人)		
		男性	女性	ヤング	ミドル	シニア
A 店	9 月	17,881	66,457	2,906	33,166	48,014
	10 月	15,372	57,353	2,439	27,275	42,771
	11 月	16,699	61,336	2,423	29,329	46,022
	12 月	16,263	61,402	2,298	29,120	45,985
	平均	16,554	61,637	2,517	29,723	45,698
B 店	9 月	15,050	37,541	1,916	23,913	26,514
	10 月	14,805	38,004	1,729	24,160	26,720
	11 月	14,389	35,669	1,682	22,427	25,650
	12 月	13,764	33,114	1,474	21,076	24,097
	平均	14,502	36,082	1,700	22,894	25,745
C 店	9 月	4,920	12,586	341	8,400	8,657
	10 月	4,500	11,783	264	7,870	8,059
	11 月	4,488	11,720	278	7,732	8,098
	12 月	4,225	11,014	263	7,240	7,667
	平均	4,533	11,776	287	7,811	8,120
D 店	9 月	9,922	31,591	1,033	18,970	21,369
	10 月	9,595	31,426	920	18,236	21,666
	11 月	9,546	31,073	885	18,111	21,413
	12 月	9,275	30,223	902	17,308	21,078
	平均	9,585	31,078	935	18,156	21,382

3.2 データの加工

本研究では ID-POS データを 2 つの方法で加工したデータを用いて分析を行う．

1 つ目は，図 2 のように加工したデータで，本研究では「1 次元データ」と呼ぶ．1 次元データは，ID-POS データをレシートごとにまとめ，1 回の購買行動を 1 行で表現したデータである．横に商品・時間・曜日について，その購入個数を各カラムに並べた形式[*7]となっており，1 行が 1 度の購買行動を表している．

2 つ目は，図 3 のように加工したデータで，本研究では「2 次元データ」と呼ぶ．2 次元データは，ID-POS データをレシートごとにまとめ，1 回の購買行動を 1 つのテーブル形式で表現したデータである．具体的には，縦に商品，横に各曜日と時間を組み合わせ，その購入個数をテーブルに並べた形式[*8]となっており，そのテーブルデータを 1 か月のレシート枚数だけ並べた形式となっている．

1 次元データは 1 行，2 次元データは 1 つのテーブルが，1 枚のレシートを表しており，各機械学習モデルはこのレシートの情報を学習し，性別・年齢を予測する．

3.3 分析手法

本研究では，機械学習のうち K 近傍法，ランダムフォレスト，ニューラルネットワークの 3 種類を使用し，比較検証を行った．1 次元データは K 近傍法，ランダムフォレスト，

[*7] 図 2 の 1 行目を例にすると，レシート A はある顧客が 1 回の購買行動において，商品 1 を 3 個と他の商品を合わせて，月曜日の 23 時に 9 個，購入したことになる．

[*8] 図 3 のレシート A を例にすると，ある顧客が 1 回の購買行動において，商品 1 を月曜日の 23 時に 3 個購入し，他の商品についても同様に並べられている．

表 3: K 近傍法による精度の結果

店舗	テストデータ	性別精度	年齢層精度
A 店	10 月	0.507 (K=29)	0.393 (K=30)
	11 月	0.519 (K=29)	0.395 (K=30)
	12 月	0.514 (K=29)	0.394 (K=30)
	平均	0.513	0.394
B 店	10 月	0.527 (K=29)	0.348 (K=28)
	11 月	0.534 (K=29)	0.348 (K=30)
	12 月	0.535 (K=29)	0.349 (K=29)
	平均	0.532	0.348
C 店	10 月	0.551 (K=29)	0.401 (K=15)
	11 月	0.552 (K=25)	0.403 (K=13)
	12 月	0.551 (K=29)	0.408 (K=13)
	平均	0.551	0.404
D 店	10 月	0.581 (K=29)	0.404 (K=30)
	11 月	0.574 (K=27)	0.400 (K=29,30)
	12 月	0.577 (K=29)	0.403 (K=30)
	平均	0.577	0.402
全店平均	10 月	0.542	0.387
	11 月	0.545	0.388
	12 月	0.544	0.389
	平均	0.544	0.388

ニューラルネットワークで使用し，2 次元データはニューラルネットワークで使用した．ニューラルネットワークでは，画像のような平面のデータを学習させることが可能である．そこで本研究では，ID-POS データを単なるテーブル形式以外の形に加工することで，よりニューラルネットワークに適した方法で扱えると考え，2 次元データに加工し分析を行った．また，学習用データには 2021 年 9 月，テストデータには同年 10,11,12 月の 3 か月をそれぞれ使用する[*9]．ニューラルネットワークのハイパーパラメータは表 1 の通りで，20 回の学習の間に損失関数の値が改善されなくなった時点で学習を終了した．1 次元データに利用するニューラルネットワークは，隠れ層が 4 層からなり，2 次元データに利用するニューラルネットワークは，畳み込み層 2 層とプーリング層 1 層を含む，隠れ層が全部で 5 層からなる

[*9] 通常の機械学習では，訓練データ: 検証データ: テストデータ=8:1:1 などが使われる．今回は，9 月で学習し，テストデータとして 10 月，11 月，12 月を用いている．これは，人の生活は 1 週間や 1 か月の周期性などもあることを考慮したためである．

表 4: ランダムフォレストによる精度の結果

店舗	テストデータ	性別精度	年齢層精度
A 店	10 月	0.638	0.526
	11 月	0.655	0.536
	12 月	0.659	0.529
	平均	0.651	0.530
B 店	10 月	0.663	0.490
	11 月	0.670	0.490
	12 月	0.667	0.492
	平均	0.667	0.491
C 店	10 月	0.676	0.492
	11 月	0.677	0.490
	12 月	0.671	0.499
	平均	0.675	0.494
D 店	10 月	0.687	0.510
	11 月	0.679	0.509
	12 月	0.673	0.499
	平均	0.680	0.506
全店平均	10 月	0.666	0.505
	11 月	0.670	0.506
	12 月	0.668	0.505
	平均	0.668	0.505

構築とした[*10]．

3.4 予測する顧客情報

予測する顧客の情報は,「性別」「年齢層の」2 つである．

- 性別：男性，女性の 2 つのカテゴリに分類
- 年齢層：ヤング層 (10〜20 代)，ミドル層 (30〜50 代)，シニア層 (60 代〜) の 3 つのカテゴリに分類

店舗の特性上，顧客の性別と年齢層にそれぞれ偏りが生じるため，事前に学習データを男女比と年齢層比がそれぞれ平等となるようにアンダーサンプリングを行い調整した[*11]．そのため，表 2 における各店舗の 9 月に示す通り，学習には性別は男女合わせて，A 店は約 35,800 件，B 店は約 30,100 件，C 店は約 9,800 件，D 店は約 19,800 件を使用し，年齢層は各年齢層合わせて，A 店は約 8,700 件，B 店は約 5,700 件，C 店は約 940 件，D 店は約 3,000 件を使用した．また，テストには各店舗の平均に示す通り性別は男女合わせて，A 店は約 84,200 件，B 店は約 50,600 件，C 店は約 16,300 件，D 店は約 40,700 件を使用し，年齢層は各年齢層合わせて，A 店は約 77,900 件，B 店は約 50,300 件，C 店は約 16,200 件，D 店は約 40,500 件を使用した．全店舗において，男性よりも女性の方が多く，年齢層については，シニア層が最も多く，ミドル層，ヤング層の順に少ない．また，店舗ごとで比較すると，顧客数は A 店が最も多く，B 店，D 店，C 店の順に少ない．

[*10] 様々なモデルを試した結果，今回はテストデータにおける精度が最も良かったものを示す．

[*11] 男性: 女性=1:1，ヤング層: ミドル層: シニア層=1:1:1

表 5: ニューラルネットワークによる精度の結果

店舗	テストデータ	性別精度	年齢層精度
A 店	10 月	0.607	0.502
	11 月	0.578	0.447
	12 月	0.679	0.557
	平均	0.621	0.502
B 店	10 月	0.645	0.322
	11 月	0.663	0.334
	12 月	0.652	0.315
	平均	0.653	0.324
C 店	10 月	0.703	0.441
	11 月	0.710	0.444
	12 月	0.693	0.439
	平均	0.702	0.441
D 店	10 月	0.686	0.442
	11 月	0.513	0.378
	12 月	0.678	0.436
	平均	0.626	0.419
全店平均	10 月	0.660	0.427
	11 月	0.616	0.401
	12 月	0.676	0.437
	平均	0.651	0.421

表 6: 2 次元データによる精度

店舗	テストデータ	性別精度	年齢層精度
A 店	10 月	0.733	0.487
	11 月	0.691	0.450
	12 月	0.699	0.444
	平均	0.708	0.460
B 店	10 月	0.718	0.491
	11 月	0.699	0.529
	12 月	0.677	0.491
	平均	0.698	0.504
C 店	10 月	0.612	0.433
	11 月	0.577	0.409
	12 月	0.609	0.425
	平均	0.599	0.422
D 店	10 月	0.710	0.487
	11 月	0.751	0.475
	12 月	0.730	0.483
	平均	0.730	0.482
平均	10 月	0.693	0.475
	11 月	0.680	0.466
	12 月	0.679	0.461
	平均	0.684	0.467

4. 分析結果

4.1 1 次元データの学習による予測精度

1 次元データを用いた各機械学習における精度を，表 3 〜表 5 に示す．K 近傍法は，性別が 50.7%〜58.1%(平均：54.4%)，年齢層が 34.8%〜40.8%(平均：38.8%) であった[*12]．また，ランダムフォレストは，性別が 63.8%〜68.7%(平均：66.8%)，年齢層が 49.0%〜53.6%(平均：50.5%) であった．さらに，ニューラルネットワークは，性別が 51.3%〜71.0%(平均：65.1%)，年齢層が 31.5%〜55.7%(平均：42.1%) であった．各店舗各月における全ての平均を比較すると，性別と年齢のどちらに関してもランダムフォレストが最も良い精度を示し，最大値を比較するとニューラルネットワークが最も良い精度を示した．しかし，ニューラルネットワークは最小値が他の機械学習モデルと比較すると低く，特に B 店の年齢層に関しては，3 カテゴリをランダムに分類した場合である 33%と同等の精度となり，安定した精度を示すことはできなかった．

4.2 2 次元データの学習による予測精度

2 次元データを用いたニューラルネットワークにおける精度を，表 6 に示す．性別が 57.7%〜75.1%(平均：68.4%)，年齢層が 40.9%〜52.9%(平均：46.7%) であった．1 次元データでのランダムフォレスト・ニューラルネットワークと比較すると，性別の精度は高いが，年齢層の精度はやや低くなった．しかし，性別に着目すると，C 店が他店舗と比較して低い精度を示したため，平均値が低くなったと考えられる．C 店を除く A,B,D 店では 67.7%〜75.1%(平均：71.2%) となり，特に D 店においては 3 か月とも 70%を超え，他の機械学習モデルより高い精度を示した．一方で，年齢層に着目すると，C 店を除いても 50%弱であり，1 次元データを用いたランダムフォレストの方が精度は高い結果となった．C 店の精度が低く示されたのは 2 次元データでのニューラルネットワークでのみ見られた傾向であるが，表 2 に示した通り，学習データ数が最も少なく，最も多い A 店と比較すると，約 1/3 であるため，精度が低くなったと考えられる．

4.3 学習データを用いた精度

表 7 に，各機械学習の訓練用データである 9 月での精度を示す[*13]．K 近傍法，1 次元データでのニューラルネットワークがどちらも高くない精度を示したのに対し，ランダムフォレストと 2 次元データでのニューラルネットワークは高い精度を示したが，予測精度は低いため過学習であることがわかった．1 次元データでのニューラルネットワークは，4.1 節で示した通り，性別についてはランダムフォレストと同等の精度を示したが，学習データでの精度が低いため，十分に学習できていないと考えられる．

5. おわりに

本研究では，マーケティングへの活用を狙った様々な分析や検証が可能である ID-POS データにおいて，年齢や性別などの顧客情報が欠損したデータが存在することに着目し，ID-POS データの利用価値を更に高めることを目的に，各機械学習モデルを用いて顧客の性別や年齢の情報を予測した．予測の結果，ランダムフォレストと 2 次元データを用いたニューラルネットワークが他の機械学習モデルと比

[*12] K 近傍法は K を 1 から 30 までとした．

[*13] KNN は各月で最も精度が高いときの K の値が異なるため，10 月でのテスト時に最も高精度であった K での値を示す．

表 7: 訓練用データ (9 月) での精度

モデル	店舗	性別精度	年齢層精度
K 近傍法	A 店	0.628	0.518
	B 店	0.625	0.475
	C 店	0.626	0.510
	D 店	0.638	0.500
ランダムフォレスト	A 店	0.977	0.979
	B 店	0.982	0.988
	C 店	0.996	0.997
	D 店	0.991	0.994
ニューラルネットワーク	A 店	0.666	0.609
	B 店	0.678	0.591
	C 店	0.727	0.849
	D 店	0.713	0.696
2 次元データにおけるニューラルネットワーク	A 店	0.941	0.826
	B 店	0.961	0.854
	C 店	0.993	0.910
	D 店	0.982	0.860

較して高い精度を示した.

本研究では，学習データの前処理において，性別と年齢層で公平性を保つため，ダウンサンプリングを行った．そのため，学習時にデータを十分に利用できなかったことから，全体としてどのモデルにおいても実用可能レベルでの高い精度を示すことができなかったと考えられる．今後は，交差検証やダウンサンプリング時にバギングを行うアンサンブル学習等を検討し，データを失わずにより高精度なモデルの構築を目指す.

参考文献

[1] 全国スーパーマーケット協会：2023 年版スーパーマーケット白書，入手先：<http://www.super.or.jp/wp-content/uploads/2023/02/NSAJ-Supermarket-hakusho2023.pdf> (参照：2023-06-13).

[2] True Data：ID-POS データ分析の基礎と活用例，入手先:<https://www.truedata.co.jp/idpos/> (参照:2023-06-13).

[3] 電通リテールマーケティング：ID-POS データ分析の本質 1 ＜基礎と活用事例＞，入手先：<https://dentsu-rm.co.jp/markettopics/258/> (参照：2023-06-13).

[4] 読売新聞オンライン：キャッシュレスなのに!客もバイトもイライラ会計，入手先：<https://www.yomiuri.co.jp/fukayomi/ichiran/20180724-OYT8T50015/4/> (参照：2023-06-13).

[5] ニッセイ基礎研究所：キャッシュレスを学ぼう (7)-共通ポイントサービス，入手先：<https://www.nli-research.co.jp/report/detail/id=64978?pno=2&site=nli> (参照：2023-06-13).

[6] 土井千章，片桐雅二，太田賢，重野寛：購入商品レベルでの購買行動に着目したライフスタイルの推定，情報処理学会論文誌，Vol.58，No.2，pp.298–307 (2017).

[7] 中村綾乃，吉野孝，松山浩士，貴志祥江，大西剛：客動線分析のための ID-POS データを用いたエージェントシミュレーションシステムの提案，情報処理学会論文誌，Vol.63, No.1, pp.56–65 (2022).

[8] 鈴木聖一：ID-POS 分析と AI，仮説検証に AI をどう適用し，実践に活用するか，日本オペレーションズ・リサーチ学会，Vol.66, No.1, pp.25–32 (2021).

[9] 辛郷孝，菅愛子，山下泰央，高橋大志：機械学習を用いた高級焼肉店における顧客購買データ分析及びフィンテック応用に関する研究，人工知能学会第二種研究会資料，2019 巻，BI-011 号，pp.6–7 (2019).

[10] 小柴等，石垣司，竹中毅，櫻井瑛一，本村 陽一：行動履歴データとライフスタイル調査にもとづく顧客モデル構築技術，電気学会論文誌 C(電子・情報・システム部門誌)，133 巻，9 号，pp.1787–1795 (2013).