同一顧客の購入数が少ない商品群を対象とした購買履歴に基づく商品特性分析モデル

山極　綾子[†a)]　　雲居　玄道[†]　　後藤　正幸[†]

An Analytical Model Based on Purchase History for Products with Low Multiple Purchases from Each Customer

Ayako YAMAGIWA[†a)], Gendo KUMOI[†], and Masayuki GOTO[†]

あらまし　近年，購買履歴データに基づく商品や顧客分析を行い，マーケティング施策に活用する研究が多くなされている．例えば商品分析方法について，顧客と商品の共起関係に基づき商品の特徴量を分析するモデルがある．しかし，同一顧客による被購買数が少ない商品群には，従来研究されてきたモデルを適用することが難しい．本研究ではこのような特徴をもつ事例として生花 EC サイト上の購買履歴データを対象とし，商品ごとに購買有無を識別する二値分類器を学習して，推定される偏回帰係数に基づく商品特性分析手法を提案する．具体的には，購買履歴データが商品，購入用途，顧客属性の組み合わせで与えられることに着目し，商品購買有無を目的変数，その他の情報を特徴量とする二値分類器を学習する．そこで得られた各特徴量の偏回帰係数を，購入用途と顧客属性が商品の購買有無に与える影響と捉え，商品特性を表す指標とみなすことで商品間の類似度を評価する．なお，それら関係性のモデル化には，特徴量間の交互作用を考慮可能な Factorization Machine で学習される回帰係数を用いる．最後に，提案手法を実データに適用し，その有効性を示す．

キーワード　機械学習，購買履歴，マーケティング，Factorization Machine，商品分析，類似度

1. まえがき

近年，インターネット技術の発達と PC やスマートフォンなどの情報端末の普及により，多くの購買行動が EC サイトを通して行われるようになった [1]．更に，それら EC サイト上の購買行動はログデータとして各企業に蓄積されており，その大量のデータをビジネスへ活用する重要性が高まっている [2]．特に BtoC 企業において，商品のラインナップを常に見直し，市場ニーズに合致させることは，顧客満足度を高め継続的な売り上げを確保するために欠かすことができない要素の 1 つである．そのためには，現在展開している商品が市場ニーズに合致しているかどうかなど，商品に関する分析が必要となる．自社が展開する商品が市場からどのように認識されているかを把握することにより，新しい商品の企画や，顧客への商品推薦へ活用することが可能となる．従来，購買履歴データが容易に得られなかった時代には，それら商品分析において属性情報などを用いて商品のグルーピングを行うなどの単純な方法を選択する必要があった．近年では，購買履歴データや閲覧履歴データが様々に活用されるようになり，顧客の嗜好を反映した商品分析を行い各商品間の類似度などの情報を得ることによって，データ駆動型の商品開発や顧客への商品推薦が実用化されている [3]．後者の代表的な手法である Item2Vec [4] は近年，EC サイト上での音楽や映画，ゲーム，日用品など様々な商品群の購買履歴に適用され，その有効性が示されている [5]〜[11]．一方で，この手法を含め，近年研究されている手法の多くは，同一顧客によって購入された商品の情報を用いて分析を行うモデルであり，1 人あたりの購入数が少ない商品群に対しては適用が難しい．

そこで，商品の分析を同一顧客との共起関係からではなく，各商品が購入される状況という観点から行うことを考える．ここで，二値分類器で推測される偏回帰係数は，目的変数と特徴量の関係性を表しているこ

[†] 早稲田大学，東京都
Waseda University, Tokyo, 169–8555 Japan
a) E-mail: saxophone-0105@ruri.waseda.jp
DOI:10.14923/transinfj.2021DEP0006

と [12] に注目する．実際に，顧客の商品への評価値を予測する回帰モデルで推定される偏回帰係数を用いて，解釈性のある推薦モデルを提案した藤井らの研究がある [13]．つまり，偏回帰係数を目的変数と特徴量間の関係性を示すものとして扱えることが示されている．そこで本研究では，各商品について購買有無を予測する識別器を学習し，推定された偏回帰係数を商品の特徴ベクトルとみなすことで商品分析に用いることを考える．このことにより，必要なデータは各商品の購買時の状況を表すデータ（商品属性や顧客属性，イベント情報，キャンペーン情報など）のみとなり，1人当たりの購入数が少ないデータに対しても適用が可能となる．ここで，商品の購買有無を識別する二値分類器を学習するために，ある商品を購買したデータを正例，その他の商品を購買するデータを負例とする学習データセットを作成する必要がある．その際，データセット作成において正例とする商品以外のデータを全て負例としてしまうと，そのデータ数が極端に偏ってしまうという問題が生じる．本研究では，この問題に対し，サンプリング対象の負例を適切にアンダーサンプリングするとともに，正例と負例の商品の購買有無が顧客の嗜好に依存する学習データセットを構成することで，二値分類器の識別境界に顧客の嗜好を反映できるような手法を提案する．具体的には，分析対象の商品が購入されるプロセスを考慮し，購買有無の分類器に顧客の嗜好が反映されるような学習データセットを生成する方法を示す．

本論文の提案手法は三つのステップで構成されている．初めに，二値分類器の学習を行うために適切な負例のアンダーサンプリング法を構成して学習用データセットを作成する．次に，各商品について，購買有無を識別する二値分類器を学習する．ここでは，対象データの特徴を適切に表現できる二値分類器を用いる必要があり，本研究では交互作用を考慮可能な Factorization Machine [14] を適用する．最後に，推定された偏回帰係数を用いて各商品をベクトル表現し，類似度分析を行う．各分類器は個別に学習されているが，それらの偏回帰係数は各特徴量と目的変数である商品の関係性を示すものであり，直接の比較が可能である．更に本研究では，提案手法を実際のデータに適用することで，提案手法の有効性を示し得られた結果について考察を行う．なお対象データには，1 人当たりの購入数の少ない商品群として，生花 EC サイト上での購買履歴データを用いている．

2. 準　　備

本章では分析対象データについて述べたあと，提案手法で用いる二値分類器について説明する．

2.1 分析対象データ

本研究では対象事例として，生花 EC サイト A 社から提供された購買履歴データを用いる．まず，A 社のビジネスモデルを以下の図 1 に示す．

顧客は生花 EC サイト上で購入する商品を選び，その注文情報とともに受領者の情報を入力する．その後，A 社に加盟するリアル店舗のうち，受領者に商品を届けるために最も適切な店舗を A 社が選択し，その店舗に注文情報を送る．最後に，リアル店舗が受領者に商品を届けることで，購買行動が終了する．

A 社の扱う生花商品のほとんどは，“母の日” や “誕生日” などのイベントの際に贈答用として購入されている．そのため各商品の製作についてもイベントに基づいて行われており，全ての商品は A 社が商品に付与するカテゴリーを用いて管理が行われている．カテゴリーは購入用途となり得るイベントに紐づき作成されており，“母の日” や “父の日”，“お中元” など全国的に同時期に発生するイベントや “お誕生日” や “結婚祝い” など顧客個々のタイミングで発生するイベントに紐づくものに加え，“ペット用” など購入用途そのものを表すものがある．なお，全ての商品に単一のカテゴリーが付与されているため，「イベントの対象とならない消費」はない．また，被購入数上位 1,000 商品を対象とした場合カテゴリー種類数は 41，各カテゴリーの商品数平均値は 24.4，分散は 1822.7 であった．

また，購買行動のきっかけとなるイベントは主に年に 1 回のみ発生するものが多く，多くの顧客は年に 1，2 回程度の購買に留まっている．同一顧客が 1 年間に購入した商品数を図 2 に示す．

ここでは，顧客 ID を用いて，同一顧客による購買数を算出している．図 2 より，88.29% の顧客が年に 2 回の以下の購入に留まっており，かつ 99.5% 以上の顧

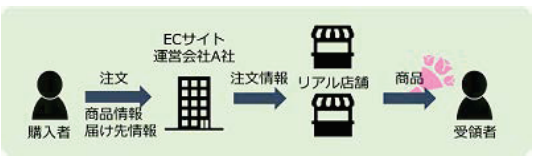

図 1　生花 EC サイト A 社のビジネスモデル
Fig. 1　Business model of the company A.

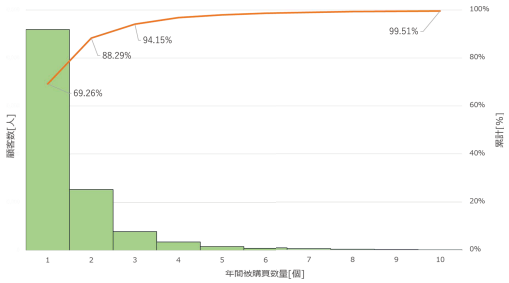

図 2　年間被購買数量別顧客数

Fig. 2　The number of customers by each purchase volume during 1 year.

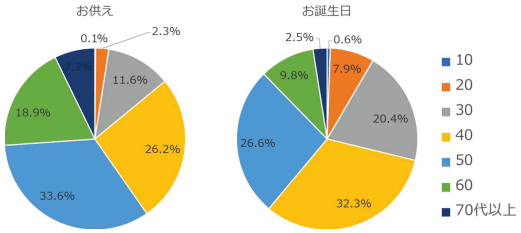

図 3　購買用途と顧客属性（年代）の関係

Fig. 3　The relationship between purpose of purchase and age.

客は，最大でも年間 10 個の購入しか行っていないことがわかる．また，本対象事例は贈答用商品であり，毎回の購買において購入目的が異なっていることが多い．そのため，複数の商品を購入している場合でもその連続する購買行動における嗜好が類似している可能性は低い．以上の特徴より，同一顧客による商品の共起関係に基づき商品の関係性をモデル化しようとする従来の手法を直接適用することは難しい．

次に，購入用途と顧客属性の関係性を図 3 に示す．左の円グラフは購買用途が “お供え” である購買，右は “お誕生日” 用途の購買について，それぞれの購買顧客の年代分布を示している．例えば “お供え” 用途では最も購買している顧客の年代は 50 代であるが，“お誕生日” 用途では 40 代が最も多くなっており，顧客属性の 1 つである年代傾向は購買用途によって異なっている．つまり，用途と顧客属性の間には相関があり，それらを適切に表現することが，商品特性のモデル化に重要であると考えられる．

2.2　関 連 研 究

近年，マーケティングの分野において機械学習の活用が進んでいる [2], [15]．Gerrikagoitia ら [2] によれば，企業側が機械学習によって実現したいトレンドとして，Interactive & media-rich，Real-time automation，Customer-journey focus や Personalization など様々なものがある．本研究は，購買履歴を用いた顧客の嗜好をモデル化を目的としており，上記のトレンドのうち Personalization にアプローチするものである．

同一の属性や購入用途をもつ顧客であってもその嗜好は異なることが一般的であり，購買履歴データによる顧客個々の嗜好分析には正解となるデータは存在しない．つまり，顧客の嗜好を分析しようとする機械学習手法の多くは，教師なし学習に位置づけられる．教師なし学習における伝統的な手法として，k-means 法によるクラスタリングや，主成分分析，要因分析による次元削減などがある [2]．これらの手法においては商品自体，若しくは商品を購入した顧客の属性情報を用いて，商品の分析を行うことができる．一方，近年発展している教師なし学習の手法として，トピックモデル [16] や行列分解を用いた手法 [17]，分散表現を学習する手法 [4] がある．例えばトピックモデルについて，各顧客と購入された商品の共起関係を潜在クラスを用いてモデル化することで，顧客の購買行動や商品を分析する研究がなされている [18]〜[20]．行列分解を用いる手法としては，顧客の商品購買有無や評価値情報を表す行列を近似する複数の低次元行列の出力を，商品や顧客の分析に用いる研究が行われている [21], [22]．また，分散表現を用いた研究も多く行われており，同一顧客による商品購買情報を用いる Item2Vec [4] とその応用研究 [5]〜[11] や，購買順序を考慮し商品の分散表現を推定する手法 [23]〜[25] など，様々な問題への適用が進んでいる．

従来研究されている手法の問題点として，同一顧客による購入数が不十分な場合にパラメータの学習ができないことや，モデルにより推定されるパラメータがもつ意味を解釈できない点がある．そのため，従来手法を適用できない事例が存在する．更にパラメータの解釈が困難であるため，得られた分散表現と人間の知見を合わせて活用し分析を行うことや，得られたパラメータを新しい商品開発に活用することは難しい．ここで，パラメータを解釈可能なモデルとして回帰モデルが挙げられる．つまり，回帰モデルで推定された係数を用いて商品の特徴ベクトルを学習することができれば，得られた特徴ベクトルの解釈が可能となる．実際に，藤井らの研究 [13] では，回帰モデルで商品の評価値予測を行い，推定された偏回帰係数を用いて顧客への推薦理由を解釈している．藤井らの研究においては，特徴量に実際に観測可能なものを用いることにより得られる回帰係数の解釈を可能としており，例えば

あるユーザにアイテムを推薦した理由について，別の類似したユーザが購入したためであることを示すことが可能であることが示されている．

更に，回帰モデルは商品が同一顧客に購入されている必要がないため，これまで分析が行えなかった商品群に対しても適用することが可能となる．

3. 提案手法

本章では提案手法の着想と，具体的なステップについて説明する．

3.1 着　　想

従来研究されている購買履歴データに基づく商品分析手法は，同一顧客による商品の共起関係から商品特性や商品間の類似度を評価しようとするものである．しかし本研究で扱う生花 EC サイト上での購買のように，1 人当たりの購入数が極端に少ない商品群では，同一顧客によって購入される商品のペア（商品の共起）というデータも非常に少なくなる．そのため，このような商品群に対しては，同一顧客による商品の購入という共起データを用いる従来の手法を直接適用することができない．そこで本研究では，回帰モデルで学習される各特徴量の偏回帰係数は目的変数との関係性を表すことに着目し，その推定値からベクトル表現を得て，商品の特徴を表現することを考える．具体的には，購買履歴データから得られる顧客属性などの情報を特徴量とし，各商品の購買有無を識別する二値分類器を学習した際に得られる偏回帰係数を用いる．ここで，得られる偏回帰係数は，目的変数である商品を購入する傾向をもつ顧客属性や状況を表していると考えられるため，それら要素を抽出したものを商品のベクトル表現とみなす．つまり，ベクトル表現が類似している商品は，その商品が購買されやすい条件という観点から類似しているとみなすことができる．実際に，佐和 [12] は，「係数そのものが特徴量の重要度を表すわけではないが，目的変数との関係性を示すものである」と述べており，回帰モデルにより得られる偏回帰係数を用いて，特徴量間の関係性を解釈する研究も行われている [13]．提案手法により，従来手法では分析が難しかった商品群に対しても，購買履歴データに基づく商品分析が可能となり，現状は担当者の経験にのみ行われる商品設計などにおいて得られた知見を活用することができる．

3.2 Factorization Machine

本研究では，購買履歴データに二値分類器を適用し，商品特性の推定を行うことを考える．ここで，特徴量の間には交互作用が存在することが事前分析より明らかになっている．そこで，本研究で用いる二値分類器として，比較的少ないパラメータ数で入力データ中の特徴量間の交互作用を考慮することができ，高い予測性能を示す予測モデルとして知られている Factorization Machine [14] (以下，FM) を用いる．

いま，目的変数を $y \in \{-1, 1\}$，特徴量ベクトルを $\boldsymbol{x} = (x_1, x_2, ..., x_d)$ としたデータが与えられているとし，w_0 をバイアス項，$\boldsymbol{w} = (w_1, w_2, ..., w_d)$ を各特徴量の偏回帰係数のベクトルとする．特徴量間の交互作用をその積により表現することができる交互作用行列を $\boldsymbol{V} = (\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top, ..., \boldsymbol{v}_d^\top)^\top \in \mathbb{R}^{d \times k}$ とし，その要素を交互作用ベクトル $\boldsymbol{v}_i = (v_{i1}, v_{i2}, ..., v_{ik})(k \ll d)$ と定義する．ここで，特徴量について，量的変数であったり，アンケートデータのように，各特徴量に対し当てはまる場合は 1，そうではない場合は 0 といったデータであれば，そのまま入力データを用いることができる．一方，本研究の対象事例のようにカテゴリー変数を特徴量として扱う場合，各変数ごとに 1hot ベクトル化する必要がある．例えば “顧客性別” という特徴量に対し “女性 (F)”，“男性 (M)”，“その他 (O)” のそれぞれ異なる値をもっている場合には，1hot ベクトル化（ダミー変数）し，いずれかの値に該当する箇所に 1 が割り当てられる．そのため，本研究では特徴量 x_i は 0, 1 のいずれかを取る二値変数となっている．

特徴量の種類ごとの 1hot ベクトルを連接したものが FM における特徴量ベクトル $\boldsymbol{x}$ となる．また目的変数 y について，本研究では商品 A を購買したデータと，それ以外の商品を購買したデータを識別する分類器の学習を行う．そのため，目的変数とする購入商品についても同様に変形し，y_n^A をある商品 A を購入したとき 1，それ以外の商品を購入したとき -1 となる変数とする．いま，上述で定義した変数を用いると，$\boldsymbol{x}$ が与えられたとき，FM は次の式 (1) で表される．

$$f(\boldsymbol{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_i x_j \qquad (1)$$

ただし，第 3 項の $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ は式 (2) で定義されるベクトルの内積を示す．

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \sum_{l=1}^{k} v_{il} v_{jl} \qquad (2)$$

FM は重回帰モデルに対して式 (1) の第 3 項を加える

ことで，特徴量間の交互作用を表現している．ここで，FM の交互作用は，$d \times k (k \ll d)$ の交互作用行列と呼ばれる行列の積で表現されている．そのため，多項式回帰モデルに比べ比較的少ないパラメータ数で交互作用を表現することができ，FM は交互作用が存在する問題に対して重回帰モデルよりも予測精度の向上が期待できる．

3.3 学習ステップ

次に，提案手法の学習ステップを以下に示す．

（1） 学習用データセットの作成
（2） FM による分類器の学習
（3） 係数を用いた商品のベクトル表現と類似度分析

3.3.1 学習用データセットの作成

商品 A の購買有無を識別する二値分類器を学習するために，商品 A の購買データを正例，その他の商品の購買データを負例とするデータセットを作成する．ここで，サイト上に蓄積されている全てのデータを対象とした場合，負例データは「商品 A を購入していない全てのデータ」となり，データ数が極端に偏ってしまう．このような正例と負例の数が偏った学習データに対し二値分類器を適用した場合，適切に係数の推定を行うことができない．そのようなアンバランスな学習データから得られた係数は，真にデータの特徴を表すものとはいえない [26]．そのため，適切に負例データをアンダーサンプリングし，正例と負例の数を一致させる必要がある．

また，負例のサンプリング方法によって，推定される係数が異なることに注意する必要がある．つまり，適切な負例の選択を行うことで，二値分類器の係数に顧客の嗜好を反映することが可能であり，結果として係数を用いた商品特性の評価が可能になると考えられる．本研究では，A 社による商品制作の際に，全ての商品に紐づけられているどのイベント用の商品であるかを示す情報を用いて，同じイベントを対象として制作された商品の購買履歴データから負例データのサンプリングを行う方法を提案する．なお，適切な負例の選択について，ランダムに選択，同一顧客からの購買履歴情報，EC サイト上での閲覧履歴情報を用いた場合それぞれに対し同様の分析を行い結果が変化することを確認する．更に得られた結果に基づき，提案手法による負例選択が適切であったことを示す．

3.3.2 分類器の学習

本研究で用いる対象データの事前分析から，特徴量間に交互作用が存在することが明らかになっている．そこで本研究では，交互作用行列の積により交互作用を表現することで，少ないパラメータで交互作用を評価することが可能な FM [14] を二値分類器として用いることとする．ある商品 A の購買有無を推定する分類器について，目的変数 $y^A(\boldsymbol{x})$ をある商品 A を購買した場合に 1，他商品を購買した場合に -1 を取る変数，特徴量 $\boldsymbol{x}^A$ を購入用途などの購買に関する情報と顧客属性とする．なお，わかりやすさのため特徴量についても $\boldsymbol{x}^A$ と記載しているが，特徴量ベクトルの順序はどの商品を対象とした場合にも変化しないため，$\boldsymbol{x}$ はどの商品の購買を目的変数として扱っていても，常に一定の定義域をもっている．

以下の式 (3) より商品 A の特性を表現するパラメータとして，各変数の直接効果を表す係数 $\boldsymbol{w}^A = (w_0^A, w_1^A, ..., w_d^A)$ 及び，交互作用を表す行列 $\boldsymbol{V}^A \in \mathcal{R}^{d \times k}$ が学習される．なお，$\boldsymbol{V}^A = (\boldsymbol{v}_1^{A^\top}, \boldsymbol{v}_2^{A^\top}, ..., \boldsymbol{v}_d^{A^\top})^\top$，$\boldsymbol{v}_i^A = (v_{i1}^A, v_{i2}^A, ..., v_{ik}^A)^\top$ である．

$$\hat{y}^A(\boldsymbol{x}) = w_0^A + \sum_{i=1}^{d} w_i^A x_i^A + \sum_{i=1}^{d} \sum_{j=i+1}^{d} \langle \boldsymbol{v}_i^A, \boldsymbol{v}_j^A \rangle x_i^A x_j^A \tag{3}$$

$$\langle \boldsymbol{v}_i^A, \boldsymbol{v}_j^A \rangle = \sum_{l=1}^{k} v_{il}^A v_{jl}^A \tag{4}$$

なお，$\boldsymbol{w}^A$ と $\boldsymbol{V}^A$ を学習する際に最小化すべき損失関数は，以下の式 (5) で表される．ここで，$\hat{y}^A$ は予測値を示している．

$$\ln(\exp(-y^A \hat{y}^A) + 1) + \lambda_w \|\boldsymbol{w}^A\| + \lambda_v \|\boldsymbol{V}^A\| \tag{5}$$

ただし，$\|\boldsymbol{\alpha}\|$ はベクトル $\boldsymbol{\alpha}$，若しくは行列 $\boldsymbol{\alpha}$ のユークリッドノルムを表し，λ_w と λ_v はパラメータの過学習を防ぐための正則化パラメータである．本研究では，交互最小 2 乗法を用いてパラメータの推定を行っている．

3.3.3 係数を用いた類似度評価

学習された二値分類器から得られる偏回帰係数を用いて，商品の類似度評価を行う手法について説明する．本研究における学習データを作成する際，FM を適用するために購買データの各情報ごとに，質的変数をダミー変数を用いて変換していることに留意する必要がある．図 4 に，FM で学習される交互作用行列と，それから得られる特徴量間の交互作用を示す $d \times d$ 次元の行列 $\boldsymbol{V}^A \boldsymbol{V}^{A\top}$ について，その特徴を示す．

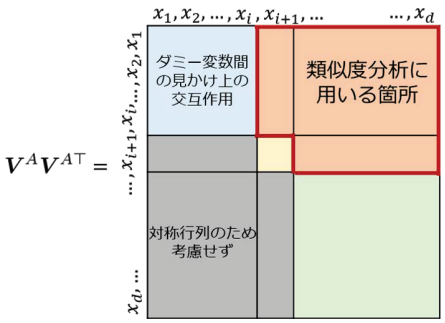

図 4　類似度分析に用いる箇所
Fig. 4　A part for similarity analysis.

まず，図 4 に示す $d \times d$ 次元の行列 $\boldsymbol{V}^A \boldsymbol{V}^{A\top}$ は，対角行列であるため右対角成分に着目する．ここで，特徴量 $\boldsymbol{x}$ はその値が 0, 1 となるよう，各質的変数ごとにダミー変数を用いて 1hot ベクトルとしている．そのため $d \times d$ 次元の行列中には，ダミー変数の導入により同じ質的変数間に生じた見かけ上の交互作用が存在しており，その値を類似度評価に用いることは適切ではない．したがって，本研究で類似度分析に用いる箇所は，直接効果 $\boldsymbol{w}$ と図 4 の右上の箇所となる．

次に，類似度評価指標の定式化を行う．交互作用行列 $\boldsymbol{V}^A$ の積 $\boldsymbol{V}^A \boldsymbol{V}^{A\top}$ の要素 $[\gamma_{ij}](i, j = 1, 2, ..., d)$ に対し，その集合を $\Gamma = \{\gamma_{ij}\}$ と定義する．ここで Γ の部分集合 $\tilde{\Gamma} \subset \Gamma$ を類似度分析に用いる要素集合とし，その要素を $\tilde{\gamma}_{ij} \in \tilde{\Gamma}$ とする．なお，$\tilde{\gamma}_{ij}$ の個数を $\tilde{d}_\gamma$ とする．類似度分析に用いる交互作用の要素を並べた $\tilde{d}_\gamma$ 次元ベクトル $\tilde{\boldsymbol{w}}^A = (\tilde{\gamma}_{ij})$ と，FM で学習された直接効果を表す $d+1$ 次元ベクトル $\boldsymbol{w}^A$ を用いて，商品 A と B 間の類似度 $S(A, B)$ を式 (6) のように定義する．ただし，$\boldsymbol{W}^A$ は $d + 1 + \tilde{d}_\gamma$ 次元の商品 A の商品ベクトルであり，$\boldsymbol{W}^A = (\boldsymbol{w}^A, \tilde{\boldsymbol{w}}^A)$ である．

$$S(A, B) = \frac{\langle \boldsymbol{W}^A \boldsymbol{W}^B \rangle}{\|\boldsymbol{W}^A\| \|\boldsymbol{W}^B\|} = \frac{\boldsymbol{W}^A \boldsymbol{W}^{B\top}}{\|\boldsymbol{W}^A\| \|\boldsymbol{W}^B\|} \tag{6}$$

ここで，商品 A について得られた商品ベクトル $\boldsymbol{W}^A$ は，直接効果と交互作用の係数推定値を並べたものである．直接効果は各特徴量と商品 A の関係性を示すものであり，交互作用は二つの特徴量と商品 A の関係性を示すものである．本提案手法で得られた商品ベクトル $\boldsymbol{W}^A$ は各商品 A が購入された購買履歴データを正例とした二値分類器により得られたものであり，大きい値であるほど購入する可能性を高くする要因であることを示している．つまり，これらの推定値を用いて各特徴量が各商品の購入され易さに与える影響を解釈することができる．

4. 実データ分析結果

4.1 適切なパラメータの選択

提案手法では二値分類器として FM を用いている．FM では，交互作用ベクトルの次元数 k と，正則化パラメータ λ_w と λ_v の値を事前に設定する必要がある．そこで，年間被購買数量上位 200 位の商品を対象とし，それらが商品 A を購買したデータであったか否かを予測する問題を作成することで，適切なパラメータの選択を行った．その際の実験条件を以下に示す．

- 期間：2018 年 8 月–2019 年 7 月（注文日ベース）
- 商品：年間被購買数上位 200 商品
- テストデータサンプリング数：各商品最大 100 件
- 学習データサンプリング数：各商品最大数
- 目的変数：各商品の購買有無
- 特徴量：購買に関する情報と商品を購買した顧客属性

なお，各商品についてその商品を購入した顧客の嗜好を学習するため，対象 200 商品それぞれについて，その商品を正例としたデータを作成している．また，商品 A の購買データを正例として用いる際，商品 A の購買データを全て正例として用いており，負例のデータ数はテストデータ，学習データ共に，正例と同じ数までアンダーサンプリングを行っている．事前実験の結果より，交互作用ベクトルの次元数 $k = 10$，正則化パラメータとして $\lambda_w = 0.01$，$\lambda_v = 0.5$ を用いることとした．

4.2 提案手法により得られた商品ベクトル概要

分析対象の購買履歴データ詳細を以下に示す．なお，期間，目的変数，特徴量については検証実験と同様である．

- 商品：年間被購買数上位 1,000 商品
- 対象購買履歴数：各商品最大 100 件（正例時）

具体的な特徴量は購入用途，受注時間帯，購買月，購買曜日，注文日と届け日までの差（一週間ごと，13 週以上は一つにまとめる），性別，年代（10 歳ごと）及び法人フラグであり，各属性ごとに 1hot ベクトルに変換を行っているため，最終的な特徴量の次元数 d は 102 となっている．

提案手法により得られた各商品の商品ベクトル $\boldsymbol{W}^A$ について，可視化手法である t-SNE [27] を用いて 2 次元に圧縮した結果を以下の図 5 に示す．なお，本分析

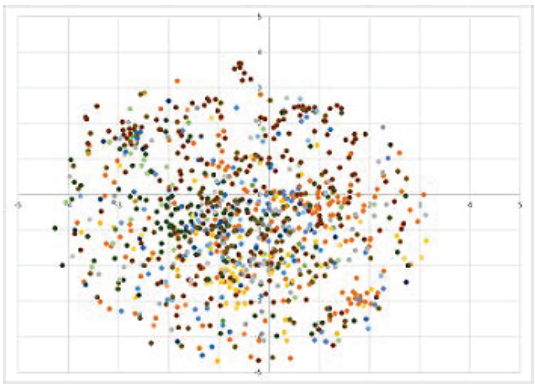

図 5　提案手法により得られた商品ベクトルの分布
Fig. 5　Distribution of the obtained product vector by proposed method.

では $d = 102$，$\tilde{d}_\gamma = 1020$ であるため，商品ベクトルの最終的な次元数は $d + 1 + \tilde{d}_\gamma = 1123$ となっている．

図 5 におけるプロットの各点は商品を示し，その色はカテゴリーを示している．この図から以下の 3 点を指摘できる．1 点目に，同じカテゴリーの商品が類似した商品ベクトルをもつことなく分布していることである．本研究の目的は，異なるカテゴリーに属しており，かつその商品を購入する顧客の嗜好が類似している商品の類似度を高く評価することであった．したがって，カテゴリーの分布は研究目的に合致しているといえる．更に 2 点目として外れ値が存在していないこと，3 点目として，幾つかの商品がクラスタを生成していることがわかる．例えば図中右下の商品の塊について，実際に得られた商品ベクトルに対し k-means [28] 法を用いてクラスタリングを行った結果，同じクラスタに属していた．

4.3　提案した負例サンプリング手法の有効性の検証

本提案手法では，ある商品の購買有無を識別する二値分類器を学習するためのデータセット生成時負例データのサンプリングを行っており，選択された負例データによって得られる偏回帰係数は変化すると考えられる．実際に，負例の選択方法によって分析結果が変化することを検証するため，提案アルゴリズムにおいて負例選択方法を変化させた場合の類似度分析結果について示す．更に，提案した手法が顧客嗜好を反映した商品ベクトルの学習に有効であることを，クラスタリング後の各クラスタの偏りから評価する．

4.3.1　比較対象となる負例選択手法概要

本節では，以下の商品群を負例選択対象とした比較手法と提案手法の分析結果を比較検証する．

- 全ての商品 (以下，比較手法 1)
- 商品 A を購買した顧客による購買数下位 N% の商品 (以下，比較手法 2)
- EC サイト上での商品 A との同一顧客閲覧数下位 N% の商品 (以下，比較手法 3)

以上の商品群を対象として負例のサンプリングを行う手法を比較手法 1 3 とする．

ここで，比較手法 1 は，商品 A 以外の商品を購買した全ての購買データを負例の対象とする方法である．比較手法 2 では，購買履歴データに含まれる顧客 ID を用いて，同一顧客の購買商品を抽出している．1 年間の購買商品が 2–99 個の顧客の購買履歴データを対象とし，データの傾向を考慮し，同一顧客による購買が少ない下位 $N = 60$% の商品を負例の対象データとして用いた．比較手法 3 は，EC サイト上の閲覧履歴に基づく負例選択を行う手法である．具体的には，同一 IP アドレスからのアクセスがあったデータを同一顧客からのアクセスであるとみなし，負例抽出の基準としている．商品 A を閲覧した顧客が閲覧した商品を調べ，データの傾向を考慮し，その回数が下位 $N = 20$% の商品を負例の対象として抽出した．それぞれ，負例として選択される商品群が異なっていることが確認された．

また，提案手法と同様の条件で検証実験を行い，FM のパラメータについて適切な値の推定を行った．その結果，交互作用ベクトル次元数 k について，それぞれ比較手法 1 では $k = 10$，比較手法 2 では $k = 3$，比較手法 3 では $k = 2$ とし，正則化パラメータは提案手法と同等の $\lambda_w = 0.01$ と $\lambda_v = 0.5$ を用いることとした．

4.3.2　負例選択手法による分析結果差異

提案手法とこれら三つの比較手法から得られた商品ベクトルを用いて，式 (6) で計算される類似度のランキングを計算した結果を図 6 に示す．図 6 は，各列と行が商品を示す商品数 × 商品数の行列において，i 行目の商品に対して j 番目の商品が類似度ランク何位であるかを示したものである．本分析では 1,000 商品を対象としていたため，実際には 1 位から 1,000 位までのランクが振り分けられているが，見やすさのため，各商品上位 50 位までを表現している．なお，色が濃い箇所が類似度ランクが上位の商品を示している．

図 6 より，それぞれの手法によって異なる類似度ランクが得られることがわかる．すなわち，負例の選択方法によって分析結果が変化するといえる．

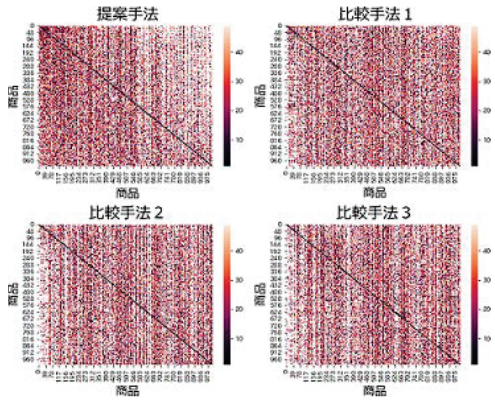

図 6　負例サンプリング手法ごとの類似度ランク
Fig. 6　Similarity rank for each negative data sampling method.

表 1　各手法で得られた商品ベクトルによるクラスタ間の顧客の属性比分散
Table 1　Variance of customer features of each cluster based on methods.

	提案手法	比較手法 1	比較手法 2	比較手法 3
性別	0.0075±0.003	0.0062±0.001	0.0052±0.001	0.0058±0.001
年代	0.0147±0.000	0.0145±0.000	0.0144±0.000	0.0144±0.000

次に，それぞれの手法によって得られた商品ベクトルを用いて商品のクラスタリングを行い，各クラスタに属する商品の購買履歴を分析し，提案した負例サンプリング手法の有効性を検証する．ここでクラスタリングには $k = 10$ とした k-means 法 [28] を用いた．各クラスタに属する商品を購入した顧客の属性比について，それぞれの負例選択方法ごとに分散を 10 回計算した結果の平均と標準偏差を表 1 に示す．

得られた商品ベクトルが顧客の嗜好を表しているとき，そのベクトルに基づくクラスタリングの結果と同じクラスタに属すると推定された商品群は，それを購入する顧客の嗜好が偏ると推測される．加えて，顧客の属性が異なる場合，その嗜好は変化することが多いと考えられる．したがって，商品ベクトルが顧客の嗜好を反映しているとき，その商品ベクトルに基づくクラスタ間において，購入する顧客の属性が変化するといえる．表 1 より，特に性別において提案手法の分散が最も大きくなっていることがわかる．つまり，クラスタによって，それぞれに属する商品を購入する顧客の性別比が異なっており，クラスタには顧客の嗜好がより反映されていると考えられる．

4.4　個々の商品を対象とした場合の詳細分析

個々の商品を対象とした場合の詳細分析について，複数の商品の例を示し，提案手法の有効性を示す．

図 7　分析対象商品とその高類似度商品群 (“お盆” カテゴリー)
Fig. 7　The objective product and products with high similarity (Obon).

4.4.1　“お盆” カテゴリー商品を対象とした分析

分析事例として花材に “紫リンドウ” と “トルコキキョウ” を用いた，カテゴリーが “お盆” で，商品名が “お供え用のアレンジメント” の商品を対象商品とし，提案手法による類似度の分析結果を示す．なお，対象商品との類似度は式 (6) を用いて算出した．対象商品と高類似度商品の商品画像を図 7 に示す．なお，商品画像を囲む四角は，その商品に対象商品と同様の花材が用いられていることを意味している．

図 7 に示す高類似度 10 商品のうち，対象商品と同じカテゴリー “お盆” に属する商品は 5 位の商品のみであり，その他の商品は全て “お誕生日” や “開店祝い” など，異なるカテゴリーの商品であった．更に花材に着目すると，対象商品と同様に “トルコキキョウ” を利用する商品が 5/10 含まれていた．花の種類には顧客の嗜好が現れていると考えられるため，本提案手法で算出した類似度には，顧客の嗜好が反映されているといえる．更に，図 7 に示す実際の商品写真を見ると，全体の形状や雰囲気など，定性的な観点からも分析対象商品に似ている商品が高類似度商品として評価されていることが分かった．実際に，企業担当者の視点からも，定性的な観点について同様の結論を得ることができた．すなわち，同一カテゴリーの商品を分類器の負例として学習することで，異なるカテゴリーの商品を高類似性商品として評価することができており，研究目的に合致した指標を得られたといえる．

4.4.2　“母の日” カテゴリー商品を対象とした分析

カテゴリーが “母の日” の商品の分析事例として，花材に “カーネーション” と “バラ” を用いた，商品名 “ティーブーケ” の商品について，類似度分析結果を示す．なお，対象商品との類似度は式 (6) を用いて算出した．対象商品と高類似度商品の商品画像を図 8 に示す．なお，商品画像を囲む四角は，その商品に対象商品と同様の花材が用いられていることを意味している．

また，これら類似度上位 10 商品には対象商品と同じカテゴリー “母の日” に属する商品は存在しておらず，異なるカテゴリーの商品を高類似度商品として抽出できていることがわかる．更に花材に着目すると，対象商品と同様に “カーネーション” 若しくは “バラ” を利用する商品が 8/10 含まれていた．また，図 8 に示す実際の商品写真を見ると，全体の形状や雰囲気など，定性的な観点からも分析対象商品に似ている商品が高類似度商品として評価されていることが分かった．したがって，本提案手法で算出した類似度には，顧客の嗜好が反映されていると考えられる．

4.4.3 “父の日” カテゴリー商品を対象とした分析

カテゴリーが “父の日” の商品の分析事例として，花材に “オレンジバラ” を用いた，商品名 “オレンジバラのアレンジメント” の商品について，提案手法による類似度分析結果を図 9 に示す．

図 9 より，その商品の形状や色味が対象商品と似ていると考えられる．また商品情報上では対象商品の花材として “バラ” のみが記載されているが，画像を見る限り，“カーネーション” や “トルコキキョウ” も用いられている．それらの花材が用いられているかという観点からいえば，提案手法により評価された高類似度商品は同じ花材を用いており，したがって顧客の嗜好を反映した類似度評価が行えていると考えられる．

図 8　分析対象商品とその高類似度商品群 (“母の日” カテゴリー)

Fig. 8　The objective product and products with high similarity (Mother’s day).

図 9　分析対象商品とその高類似度商品群 (“父の日” カテゴリー)

Fig. 9　The objective product and products with high similarity (Father’s day).

5. 考　　察

5.1 従来手法の定量的評価

本研究で対象としているデータは一人当たりの購入数が少なく，購入のたびに顧客の嗜好が変化するため従来手法の適用が困難であると考えられる．そのことを定量的に示すために，Item2Vec を用いて対象データへの適用，及び人工データを用いた提案手法との比較実験を行った．

5.1.1 実データへの適用

本研究で用いたものと同様のデータを対象として，Item2Vec により推定した商品の特徴ベクトルの分布を図 10 に示す．なお，分析対象のデータのうち約 7 割の顧客は年に 1 点のみの購入に留まっており，Item2Vec を用いて分析を行った対象は複数購入を行った 3 割の顧客のデータのみである．

図 10 の各プロットは商品を表し，その色は商品のカテゴリーを示している．図 10 より，Item2Vec により得られた特徴ベクトルはカテゴリーごとに類似した値をもっていることが多いことがわかる．ここで Item2Vec は同一顧客に購入された商品を埋め込み空間上で近しい空間に埋め込む手法であり，近い場所に埋め込まれた商品以外との関係性について述べることはできず，近くに存在しないが次点で近い場所に埋め込まれた別カテゴリーの商品を推薦することは適切ではない．以上より，本研究の対象データに従来手法である Item2Vec を適用し，異なるカテゴリーに属する商品の顧客への推薦を行うことは難しいといえる．

5.1.2 人工データを用いた比較実験

一人当たりの購入数が少ないデータにおいて，提案

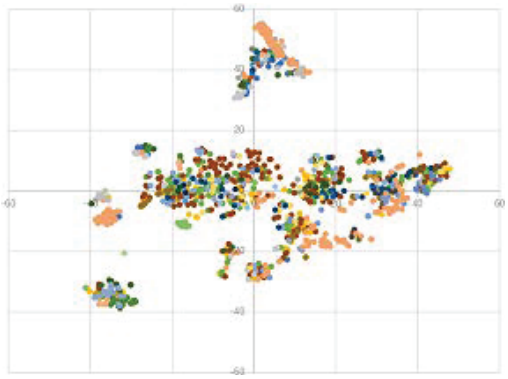

図 10　Item2Vec により得られた商品ベクトルの分布

Fig. 10　Distribution of the obtained product vector by Item2Vec.

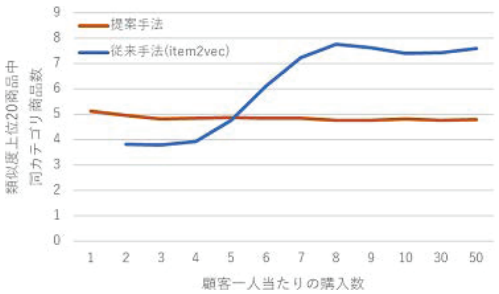

図 11　顧客一人当たりの購入数による類似商品推定精度の推移

Fig. 11　Accuracy of similar product estimation by the number of purchases per A customer.

手法及び従来手法である Item2Vec が類似した商品を正しく識別できるかを評価するために，人工データを用いて実験を行った．人工データにおける各商品は購買確率を決める特徴ベクトルの値によって 5 クラスタに分割が可能であり，各クラスタには 20 商品があるものとする．提案手法と従来手法で得られる特徴ベクトルを用いたクラスタリング結果が，真のクラスタをどの程度推定できるかについて，顧客一人当たりの購入数推移による変化を調べる．なお顧客数は 5,000 とし，評価指標として，得られた特徴ベクトルに基づく類似度上位 20 商品のうち同じクラスタに属する商品が含まれている個数を用いた．実験結果を以下の図 11 に示す．なお，顧客一人当たりの購入数が 1 の場合，従来手法では商品の特徴ベクトルの推定ができないため，評価指標についても記載していない．

図 11 より，顧客一人当たりの購入数が少ない場合において，提案手法の方が良い精度を示していることがわかる．以上より，本研究対象のように一人当たりの購入数が少ないデータにおいては，提案手法を用いることで精度良くモデル化が可能であるといえる．なお，生成した人工データの詳細については付録に示す．

5.2　交互作用を用いることの有効性評価

本研究では目的変数に対する特徴量間の交互作用を考慮するため，分類器に FM を用いた．**4.4** で分析対象として用いた商品について，二値分類器に重回帰モデルを用いた場合の類似度分析結果を図 12 に示す．

ここで，類似度上位商品のうち 5 商品が対象商品と同じ “お供え” カテゴリーに属する商品である．これは，商品購買に大きな影響を与える購入用途は多くの場合，特定の日付において発生するものであり，特徴量の一つである購買月と強い関連性があるためである．

図 12　重回帰モデル利用時類似度上位商品

Fig. 12　The objective product and products with high similarity based on logistic regression.

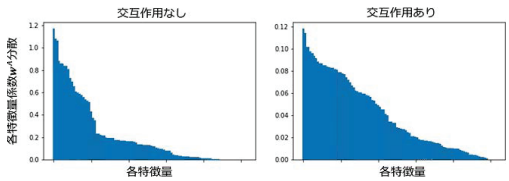

図 13　特徴量別係数分散

Fig. 13　Variance of coefficient of each explanatory variable.

そのため，少数の特徴量のみが商品購買有無の予測に影響を与える状況となり，結果として対象商品と同じカテゴリーに属する商品の類似度が高く評価されたといえる．

実際に，交互作用の考慮による特徴量の偏回帰係数への影響を明らかにするため，交互作用を考慮した場合と考慮しない場合について，各特徴量に対して商品ごとに推定された偏回帰係数の分散を降順に並べたグラフを図 13 に示す．左図が交互作用を考慮しない場合，右図は提案手法の結果を表す．また，1 対 1 の比較を行うため，どちらも各特徴量の直接効果 $\boldsymbol{w}^A$ のみを示している．

ここで，分散の値が他に比べて大きい特徴量は，商品により異なる係数の値をもつということであり，すなわち商品特性に影響を与える特徴量である．図 13 より，交互作用を考慮しないモデルで推定された係数の分散は，一部の特徴量についてのみ分散が大きくなっている．一方，交互作用を考慮することにより，比較的多くの特徴量について係数の分散が大きくなっていることがわかる．すなわち，交互作用をモデルに組み込むことにより，様々な特徴量によって商品特性を捉えることができたといえる．

特定の変数のみ係数の分散が大きいということは，それらが目的変数である商品購買有無に与える影響が大きく，関係性が強いことを意味している．すなわち，その特徴量が商品の類似性分析に与える影響が大きいといえる．具体例として，カテゴリーが “お祝い”

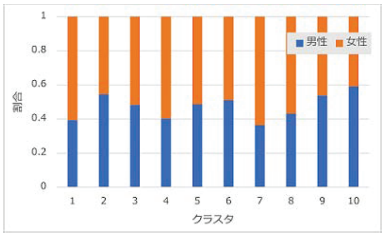

図 14　クラスタ別商品購入顧客属性の割合 (性別)
Fig. 14　Customers feature ratio of each cluster (gender).

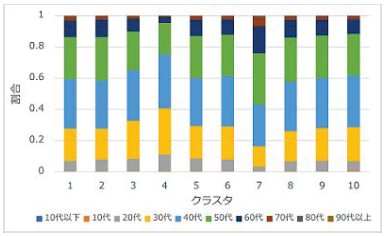

図 15　クラスタ別商品購入顧客属性の割合 (年代)
Fig. 15　Customers feature ratio of each cluster (age).

の，送別会を主な用途として購買される商品 *A* について考える．多くの送別会は 3 月に実施されるため，購入月に関する特徴量のうち，“3 月” の係数が大きくなることが予想される．ここで，交互作用を用いなかった場合，購入月とその他の特徴量の関係性は考慮されず，直接効果を示す係数のみが大きくなる．一方，交互作用を用いることで，購入月と性別，年代など，さまざまな特徴量間の関係性をモデル化することが可能となる．すなわち，交互作用を考慮することにより，影響が大きい特徴量について，その影響を他の特徴量との交互作用の形で分割することが可能となる．その結果，多くの特徴量との関係性から商品特性を捉えることができ，分類精度の向上が見られるなど，適切な係数を学習できたと考えられる．

5.3　得られた商品ベクトルの活用

提案手法により得られた商品ベクトルを，どのように活用できるかについて考察する．まず，商品ベクトルの値そのものの活用が考えられる．Embedding などの手法とは異なり，回帰モデルで推定される偏回帰係数は，それ自体が目的変数と特徴量の関係性を示すものである．本研究において商品の購買有無を目的変数，購入した顧客の属性情報などを特徴量としているため，そのため商品ベクトルの各要素は「どの属性の顧客に商品が買われる傾向があるか」ということを示している．更に FM を二値分類器に用いていることにより，特徴量間の交互作用を評価することも可能となっており，例えば女性かつ 20 代の顧客に購入されやすい商品は何か，という分析を行うこともできる．

次に，商品ベクトルをクラスタリングした結果を活用することを考える．図 14 と図 15 に，提案手法で得られた商品ベクトルに k-means 法 [28] を適用し，10 個のクラスタに分けた際，それぞれのクラスタに所属する商品を購入した顧客の属性を示す．図 14 は性別の比率を表し，図 15 は年代の比率を表している．

図 14 と図 15 より，各クラスタに属する商品がどのような属性の顧客に好まれているかを分析することができる．例えばクラスタ 7 に属する商品は女性かつ，比較的年代が上の顧客に好まれていることがわかる．この結果を活用することにより，個々の商品と商品ベクトルに着目するだけではなく，全体的な傾向を把握することができるといえる．

6.　む　す　び

本研究では，1 人当たりの購入数が少なく従来の分析手法の適用が難しい商品群を対象とした商品分析モデルを提案した．具体的には各商品の購買データに着目し，商品購買有無を識別する二値分類器の偏回帰係数を用いて商品ベクトルを表現することで，商品間の類似度を分析している．更に，二値分類器を学習する際に用いる学習データについて，顧客の嗜好を分類基準に反映できるような負例データのサンプリングが必要となる．実際に，実際に上記のような特徴をもつ購買履歴データとして生花 EC サイト上のデータに提案手法を適用し，提案手法における負例の対象データの選択基準が適切であることを示した．最後に，購買履歴データの分析結果を考察することによりその有効性を示した．

今後の課題として，被購買数が少ない商品への提案手法の適用が挙げられる．本論文では，全商品の内，年間被購買数が上位 1,000 商品のみを対象として分析を行っており，対象外とした商品の方が多く存在している．しかし，分析対象外の商品の中には年間の購買数量が極端に少ないものがあり，本提案手法ではそれら商品の購買有無を識別する識別器の学習に十分な学習データセットを用意することができない恐れがある．つまり，学習データセットが十分用意できない際にも適用可能なモデルを構築することで，分析可能商品の範囲を広げることができるといえる．

謝辞　本研究を行うにあたり用いた貴重なデータは花キューピット株式会社様よりご提供いただきました．深く感謝致します．なお本研究は，日本学術振興会（JSPS）科学研究費 No.21H04600 の助成を受けたものです．

文　献

[1] 経済産業省，“令和元年度内外一体の経済成長戦略構築にかかる国際経済調査事業（電子証取式に関する市場調査），”2020.
[2] J.K. Gerrikagoitia, I. Castander, F. Rebón, and A. Alzua-Sorzabal, “New trends of intelligent e-marketing based on web mining for e-shops,” Procedia-Social and Behavioral Sciences, vol.175, no.1, pp.75–83, 2015.
[3] L. Ma and B. Sun, “Machine learning and ai in marketing–connecting computing power to human insights,” International Journal of Research in Marketing, vol.37, no.3, pp.481–504, 2020.
[4] O. Barkan and N. Koenigstein, “Item2vec: neural item embedding for collaborative filtering,” IEEE 26th International Workshop on Machine Learning for Signal Processing, IEEE, pp.1–6, 2016.
[5] Z. Li, H. Zhao, Q. Liu, Z. Huang, T. Mei, and E. Chen, “Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors,” Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.1734–1743, 2018.
[6] Y. Gui and Z. Xu, “Training recurrent neural network on distributed representation space for session-based recommendation,” 2018 Int. Joint Conference on Neural Networks IEEE, pp.1–6, 2018.
[7] T. Tran, K. Lee, Y. Liao, and D. Lee, “Regularizing matrix factorization with user and item embeddings for recommendation,” Proc. 27th ACM Int. Conf. Information and Knowledge Management, pp.687–696, 2018.
[8] R. Rahutomo, A.S. Perbangsa, H. Soeparno, and B. Pardamean, “Embedding model design for producing book recommendation,” 2019 Int. Conf. Information Management and Technology, vol.1, IEEE, pp.537–541, 2019.
[9] O. Barkan, A. Caciularu, O. Katz, and N. Koenigstein, “Attentive item2vec: Neural attentive user representations,” IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, pp.3377–3381, 2020.
[10] O. Barkan, A. Caciularu, I. Rejwan, O. Katz, J. Weill, I. Malkiel, and N. Koenigstein, “Cold item recommendations via hierarchical item2vec,” 2020 IEEE Int. Conf. Data Mining IEEE, pp.912–917, 2020.
[11] W. Pei, J. Yang, Z. Sun, J. Zhang, A. Bozzon, and D.M. Tax, “Interacting attention-gated recurrent networks for recommendation,” Proc. 2017 ACM Conf. Information and Knowledge Management, pp.1459–1468, 2017.
[12] 佐和隆光，回帰分析，朝倉出版，1979.
[13] 藤井流華，岡本一志，“線形回帰による推薦の透明性を有したモデルベース協調フィルタリング，”人工知能学会論文誌，vol.35, no.1, pp.D–J61_1, 2020.
[14] S. Rendle, “Factorization machines,” 2010 IEEE Int. Conf. Data Mining, IEEE, pp.995–1000, 2010.
[15] S. Burt and L. Sparks, “E-commerce and the retail process: a review,” Journal of Retailing and Consumer Services, vol.10, no.5, pp.275–286, 2003.
[16] 後藤正幸，小林　学，入門パターン認識と機械学習，コロナ社，2014.
[17] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” Computer, vol.42, no.8, pp.30–37, 2009.
[18] 江本　守，大澤幸生，“トピックモデルを用いた顧客行動分析と商品 dna 作成手法の提案，”信学技報，AI2016-33, Feb. 2017.
[19] J. Jin, Q. Geng, H. Mou, and C. Chen, “Author–subject–topic model for reviewer recommendation,” J. Information Science, vol.45, no.4, pp.554–570, 2019.
[20] E.O. Park, B.K. Chae, J. Kwon, and W.-H. Kim, “The effects of green restaurant attributes on customer satisfaction using the structural topic model on online customer reviews,” Sustainability, vol.12, no.7, p.2843, 2020.
[21] Y. Lu, R. Dong, and B. Smyth, “Coevolutionary recommendation model: Mutual learning between ratings and reviews,” Proc. 2018 World Wide Web Conference, pp.773–782, 2018.
[22] C. Xu, “A novel recommendation method based on social network using matrix factorization technique,” Information Processing & Management, vol.54, no.3, pp.463–474, 2018.
[23] R. He, W.-C. Kang, and J. McAuley, “Translation-based recommendation,” Proc. 11th ACM Conference on Recommender Systems, pp.161–169, 2017.
[24] J. Tang and K. Wang, “Personalized top-n sequential recommendation via convolutional sequence embedding,” Proc. 11th ACM Int. Conf. Web Search and Data Mining, pp.565–573, 2018.
[25] Z. Sun, J. Yang, J. Zhang, A. Bozzon, L.-K. Huang, and C. Xu, “Recurrent knowledge graph embedding for effective recommendation,” Proc. 12th ACM Conf. Recommender Systems, pp.297–305, 2018.
[26] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, and F. Herrera, Learning from Imbalanced Data sets, vol.11, Springer, 2018.
[27] L. Van derMaaten and G. Hinton, “Visualizing data using t-sne,” J. Machine Learning Research, vol.9, no.11, pp.2570–2605, 2008.
[28] K. Krishna and M.N. Murty, “Genetic k-means algorithm,” IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics), vol.29, no.3, pp.433–439, 1999.

付　録

1.　人工データ生成過程詳細

5.1.2 に述べた人工データ実験で用いたデータの生成過程の詳細を以下に示す．まず各商品の特徴ベクトルについて，本研究では特徴量間の交互作用が存在するデータを対象としている．そこで，人工データの特徴ベクトルにおいても直接効果を示す特徴ベクトル $\boldsymbol{\alpha}$ と，交互作用を示す特徴ベクトル $\boldsymbol{\beta}$ を各商品について定義する．i 番目の商品 p_i について，d 番目の特徴量

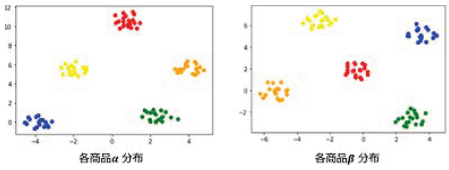

図 A·1　各商品の特徴ベクトル $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ 分布
Fig. A·1　The distribution of each Products' feature vector.

の直接効果を表す偏回帰係数 $\boldsymbol{\alpha}$ は平均 $\mu^{\alpha}_{i,d}$，分散 $\sigma^{\alpha}_{i,d}$ の正規分布に，d 番目と d' 番目の特徴量の交互作用の偏回帰係数 $\boldsymbol{\beta}$ は平均 $\mu^{\beta}_{i,d,d'}$，分散 $\sigma^{\beta}_{i,d,d'}$ の正規分布に従うものとする．ここで，商品は複数のクラスタに分割され，カテゴリーごとに $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ の平均値と分散が異なるものとする．本実験では各クラスタに所属する商品数を 20，クラスタ数を 5 とし，実際に生成した各商品の $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ について，t-SNE を用いて二次元に可視化したものを以下の図 A·1 に示す．各プロットは商品を表し，色は各商品のクラスタを表している．図 A·1 より，各クラスタに属する商品ごとに類似した特徴ベクトルが生成されていることがわかる．

次に，特徴ベクトル $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ を用いて購買履歴データを生成する．各購買における特徴量 $\boldsymbol{x} = \{x_0, x_1, ..., x_D\}$ より，各商品の購買確率 $P_i(\boldsymbol{x})$ を以下の式 (A·1) で求める．なお，特徴量 $\boldsymbol{x}$ については，$1, 2, ..., D$ 上の自然数のいずれかをランダムに取る，各値の確率が $1/D$ の乱数を生成し，その箇所の特徴量が 1，それ以外の特徴量は全て 0 を取るような 1-hot ベクトルを構成する操作を繰り返した．ここで，x_0 はバイアス項であり全てのデータにおいて $x_0 = 1$ である．なお，$\boldsymbol{x}$ には顧客固有のものと購買のたびに変化するものがあるとし，特徴量数の比は本研究の対象データに従い 1 : 10 とした．

$$P_i(\boldsymbol{x}) = \frac{1}{1 + e^{-(f_i(\boldsymbol{x}))}} \tag{A·1}$$

$$f_i(\boldsymbol{x}) = \sum_{d=0}^{D} \alpha_{i,d} x_d + \sum_{d=1}^{D} \sum_{d'=d+1}^{D} \beta_{i,d,d'} x_d x_{d'} \tag{A·2}$$

$P_i(\boldsymbol{x})$ に従い購買商品を確率的に選択し，購買履歴データを作成した．

（2021 年 6 月 25 日受付，11 月 15 日再受付，
12 月 29 日早期公開）

山極　綾子

2021 早稲田大学大学院創造理工学研究科経営デザイン専攻修士課程了．現在，同大学大学院創造理工学研究科経営デザイン専攻博士課程在学中．機械学習を用いたデータ分析に関する研究に興味をもつ．

雲居　玄道

2008 早稲田大学理工学部経営システム工学科卒．2008 同大学理工学術院総合研究所嘱託研究員．2015 浄土真宗本願寺派総合研究所研究助手．2017 早大大学院創造理工学研究科博士後期課程入学．2019 早稲田大学創造理工学部経営システム工学科助手，現在に至る．情報数理応用・テキストマイニングの研究に従事．経営情報学会，日本気象学会各会員．

後藤　正幸　（正員）

1994 武蔵工業大学大学院修士課程了．2000 早稲田大学大学院理工学博士課程了．博士（工学）．1997 同大学理工学部助手．2000 東京大学大学院工学研究科助手．2002 武蔵工業大学環境情報学部助教授．2008 早稲田大学創造理工学部経営システム工学科准教授．2011 同大学教授．情報数理応用とデータサイエンスの研究に従事．著書に，『入門パターン認識と機械学習』コロナ社 (2014)，『ビジネス統計～統計基礎とエクセル分析』オデッセイコミュニケーションズ (2015) 等．IEEE，INFROMS，情報処理学会，人工知能学会，日本経営工学会，経営情報学会等各会員．