生花 EC サイトの購買履歴データに基づく商品特性分析モデル

経営情報学研究　5219F019-7　山極綾子
指導教員　後藤正幸

An Analysis Model of Flower Products' Features Based on Purchase Data

YAMAGIWA Ayako

1. 研究背景と目的

近年，EC サイト上に蓄積された顧客の購買履歴データを活用した，様々なマーケティング施策が求められている．例えば 1 つの観点として，顧客の嗜好に基づいた商品の類似性分析を行い，類似した商品を顧客に推薦することによる売り上げ増加を目的とするものがある．類似性の評価方法として，商品のジャンルなどの情報を用いるなどの単純な手法に加え，大量に蓄積されたデータを活用する新しい手法が多く研究されている [1]．後者の代表的な手法である Item2Vec[2] は，顧客の購買履歴データを用いて，同一顧客から似たタイミングで購入される商品を類似商品と見なし，埋め込み空間上に商品を表現するモデルである．この手法は，同一顧客が短期間に複数購買を行うような日用品や映画などに適用されている．しかし，本研究の対象事例である生花 EC サイトの様に，購買間隔が長い場合や購買の度に顧客の嗜好が変化する商品群では，従来手法では類似性の推定が難しい．

生花 EC サイト上での顧客の購買行動は，年に 1 回のみの “母の日” や “誕生日” など特定のイベントでの贈答を用途として購買が行われることが多く，購買間隔は長い．加えて，連続する購買であっても，贈答品としての購買の場合受領者が異なる可能性があり，その場合は購買行動が同一の嗜好に基づいていないと考えられる．ここで，既存顧客に対し，一度購買した用途に加え別用途でも購買を促すことが重要であり，実際にそのための商品の提示が行われている．しかし現状は担当者の経験に基づき，各用途の注目商品を一律に提示するに留まっている．

以上より本研究では，対象事例のような商品群に対して，顧客の嗜好を反映した商品の類似性分析手法を提案する．そこで，同じ商品 A を購入した顧客の嗜好は類似していると仮定し，それ以外の商品を購買した顧客との違いを，顧客属性や購入用途を説明変数とした二値分類器の係数で表現することで，商品の特性を推定することを考える．その際，二値分類器は正例と負例のデータを分類する係数を学習するため，負例の選択方法が重要となる．本研究では，既存顧客に別用途での購買が見込める商品を推薦するために，異なる用途で同じ嗜好を有する顧客から購入される可能性が高い商品ほど，類似性が高くなるように推定される必要がある．そのために，与えられた正例に対し，顧客の嗜好を推定するために適切な負例選択方法を採用することで，各用途ごとにその商品を購買した顧客属性の傾向を推定することを考える．また，同じ顧客属性であっても用途が変われば受領者が異なり，嗜好が変化する可能性がある．つまり，顧客属性と用途の間には関係性が存在している．そこで，二値分類器として，説明変数間の交互作用を表現することが可能なモデルである Factorization Machine（以下，FM）[3] を用いる．最後に，実際のデータに提案手法を適用しその有効性を示す．

2. 準備

2.1. 生花 EC サイトビジネスモデル

本研究では，生花 EC サイト A 社から提供された購買履歴データを対象とする．A 社が運営する生花 EC サイトを利用することにより，購入者は希望する商品を，希望する受領者に指定した日時に届けてもらうことができる．この生花 EC サイト上の商品は，顧客の購入用途となるイベントに基づき制作される．そのため，対象としたイベントごとに，商品にはカテゴリが付与され管理されている．購入者は多くの場合，購入用途と紐づくカテゴリの商品の中から選択し，購入することになる．これは EC サイト上で，A 社が購買用途に合わせ設定したカテゴリごとに商品が提示されるためである．例えば，“母の日” 用途で EC サイトを訪れた顧客は，“母の日” カテゴリの商品が掲載されている特集ページを訪れ，それらの商品群から購入商品を決める場合がほとんどである．

2.2. 関連研究

最も単純な商品の類似性分析の方法として，“カテゴリ” や “花材” などの商品属性に基づく手法が挙げられる．しかしその手法では顧客の嗜好を反映できず，購買を促すための商品類似性を評価するには不十分であると考えられる．一方，顧客の商品購買行動には嗜好が表れることから，顧客の購買履歴に基づき，商品を埋め込み空間上に分散表現する Item2Vec が提案されており，音楽や映画，ゲームなどの購買履歴に適用されている [4]．しかし，各顧客に対してある程度の長さの購買系列が必要になるため，購買間隔が長く同一顧客が多くの商品を購入することが無い商品群に対しては適用できない．

2.3. 対象データ概要

本研究で分析に用いる購買履歴データの期間は 2018 年 8 月–2019 年 7 月（注文日ベース），対象商品数は約 7,500 個，購買履歴数は約 70 万件をサンプリングした．補助情報として生花 EC サイト上での閲覧履歴データおよび商品情報マスターデータを用いている．ここで，各商品は特定用途のために設計されているため，購入用途はほとんどの商品で大きく偏っており，約 6 割の商品において 90%以上を単一の用途が占めていた．加えて購買傾向の特徴として，同一顧客の 1 年間における購買数が小さいことが挙げられる．実際に 69.26%の顧客が年に 1 回の購買に留まっていた．さらに説明変数間の交互作用について，例えば購入用途と年代傾向の関係性を見ると，“お供え” 用途での購買は 50 年代が 33.6%，“誕生日” 用途では 40 代が 32.3%を占め，最も多

い結果となった．つまり，用途と顧客属性の間には関係性があり，それらを表現することで商品特性を推定することができる．

3. 提案手法

3.1. 着想

本研究で対象とする生花 EC サイトの事例では，同一顧客の購買間隔が長く，かつ連続する購買であってもその嗜好は変化する場合が多いと考えられる．そのため，従来手法を用いた商品特性の評価，および類似性の推定が難しい．そこで，同じ商品 A を購入した顧客の嗜好は類似しており，かつその嗜好は顧客属性に現れると仮定し，商品購買有無を目的変数とした二値分類器を学習することを考える．このとき，分類器で推定される係数は，目的変数である対象商品購買有無と各説明変数の関係性を表していると言える．実際に，佐和 [5] は，「係数そのものが説明変数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」と述べている．

ここで商品 A の購買有無を分類器の目的変数とした場合，対象商品以外のすべての商品を購買したデータを負例として扱うことになり，データ数が極端に偏ってしまう．そのため，適切に負例データをサンプリングする必要がある．これは，分類器で学習される係数は，正例と負例を分類する識別平面を学習するためである．適切な負例の選択により，顧客の嗜好を表す係数を学習できると考えられる．さらに，分類器で得られた係数が類似している商品同士は，その商品を購入する顧客や，購買される状況が類似していると考えられる．

3.2. 利用する分類器

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，説明変数 $\boldsymbol{x}^A$ を購買用途などの購買に関する情報と顧客属性とする．本研究では FM を分類器として用いており，事前に定めた次元 k のもとで $d \times k$ 次元の交互作用行列 $\boldsymbol{V}^A$ を学習し，少ないパラメータ数で交互作用を表現することができる．以下の式 (1) 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 $\boldsymbol{w}^A = (w_0^A, w_1^A, ..., w_d^A)$ および，交互作用を表す行列 $\boldsymbol{V}^A \in \boldsymbol{\mathcal{R}}^{d\times k}$ が学習される．なお，$\boldsymbol{V}^A = (\boldsymbol{v}_1^A, \boldsymbol{v}_2^A, ..., \boldsymbol{v}_d^A)^\top$，$\boldsymbol{v}_i^A = (v_{i1}^A, v_{i2}^A, ..., v_{ik}^A)$ である．

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

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

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 (3) で表される．ここで，$\hat{y}^A$ は予測値を示している．

$$\ln\left(\exp(-y^A \hat{y}^A) + 1\right) + \lambda_w \|\boldsymbol{w}^A\| + \lambda_v \|\boldsymbol{V}^A\| \tag{3}$$

ただし，$\|\boldsymbol{\alpha}\|$ はベクトル $\boldsymbol{\alpha}$ の 2 次ノルムを表し，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．

3.3. 負例データ選択方法

対象商品 A を購買したデータを正例とし二値分類器を学習する場合，負例の数が極端に多くなることから，分類器を適切に学習させるためには，負例のサンプリングを行う必要がある．しかし，商品 A を購買していない全ての購買履歴データを対象とし負例のサンプリングを行った場合，不適切な負例が選択される可能性がある．例えば，商品 A のカテゴリが “母の日” であったとき，負例として “お盆用” や “開店祝い用” など，顧客が購入時比較対象にしない商品が選択されてしまうことが考えられる．その場合，商品 A とそれら負例データを分類する FM の係数に，顧客の嗜好を反映することができない．そのため，顧客が対象商品 A の比較対象とするであろう商品を分類器の負例として用いることが望ましい．

ここで，本研究の対象 EC サイトでは，顧客は購入用途のカテゴリに属する商品が掲載されているページを閲覧し，複数の商品を検討したうえで，最終的に自身の嗜好に合致する商品を購入することが多い．逆に，同じカテゴリ内で購買されなかった商品は，顧客の嗜好に合致しなかったものと考えられる．そのため，対象商品 A と同じカテゴリの商品を負例に選択することにより，顧客の嗜好を反映した商品の特性を推定することが可能となる．

3.4. アルゴリズム概要

分類器を用いて顧客属性や用途などの関連情報が商品購買に与える影響を推定し，それを商品の特性と見なすことで類似性を評価する手法のアルゴリズムを以下に示す．

1. 学習用データセットの作成
2. 分類器の学習
3. 係数を用いた類似性評価

まず，負例データとして同一カテゴリの商品を選択し，学習データセットを作成する．次に，FM を用いて係数の推定を行う．なお，本研究では FM の学習手法として交互最小二乗法を用いている．最後に，求めた係数を用いてコサイン類似度を算出する．ここで，FM の直接効果 $\boldsymbol{w}^A$ についてはその値を直接類似度の計算に用いる．一方，交互作用ベクトル $\boldsymbol{V}^A$ については，適切な処理を行い類似度を計算している．

4. 実データ分析

4.1. 分析対象データ詳細

分析対象の購買履歴データ詳細を以下に示す．

- 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 1,000 商品
- 購買履歴数：各商品最大 100 件（正例時）
- 目的変数：各商品の購買有無
- 説明変数：購買に関する情報と商品を購買した顧客属性

なお，事前に年間被購買数上位 200 位までの商品を対象とし，検証実験を行い適切なパラメータについて検討を行った．その結果，FM の交互作用パラメータは $k = 10$，正則化パラメータはそれぞれ $\lambda_w = 0.01$，$\lambda_v = 0.5$ とした．なお，パラメータ学習には交互最小二乗法を用いている．具体的な説明変数は購入用途，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 1 つにまとめる），性別，年代（10 歳ごと）およ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，最終的な説明変数の次元数 d は 102 となっている．

4.2. 類似性分析結果

分析事例として花材に “紫リンドウ” と “トルコ桔梗” を用いた，カテゴリが “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性の分析結果を示す．対象商品と高類似性商品の商品画像を図 1 に，高類似性商品のカテゴリを表 1 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 1: 分析対象商品とその高類似性商品群

表 1: 類似性上位商品

No.	カテゴリ	No.	カテゴリ
1	お中元	6	開店祝い
2	お任せ	7	お任せ
3	開店祝い	8	誕生日
4	誕生日	9	卒入学祝い
5	お盆	10	誕生日

表 1 より，異なるカテゴリの商品を高類似性商品として抽出できていることがわかる．さらに花材に着目すると，対象商品と同様に “トルコ桔梗” 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性は，顧客の嗜好が反映されていると言える．さらに，図 1 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性商品として評価されていることが分かった．つまり，同一カテゴリ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言える．

4.3. その他の負例選択方法との比較

比較手法として，購買履歴データおよび関連情報から得られる情報を用いてサンプリング対象の負例を決定し，負例選択方法による分析結果の違いを確認する．具体的には，すべての商品 (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) をそれぞれ対象とし，負例のサンプリングを行った．なお，基準となる N は，分析対象データの同一顧客購買商品と閲覧商品の傾向より決定した．提案手法と同様の条件で分析を行っており，その際のパラメータはそれぞれ検証実験の結果，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様に $\lambda_w = 0.01$ と $\lambda_v = 0.5$ とした．

4.3.1. 類似性分析結果

まず，負例データの選択方法として購買履歴データを用い，同一顧客による購買が少ない商品群を選択した際の類似性分析結果について図 2 に示す．

図 2: 購買履歴データに基づく負例選択時 分析結果

類似性上位 10 商品に同一カテゴリの商品は含まれておらず，異なるカテゴリの商品の類似性を高く評価することはできている．しかし図 2 より，高類似性と分析された商品は定性的に対象商品と類似しているとは考えにくい．異なる用途向けの多様な商品を顧客に提示することは重要であるが，あまりにも感覚とかけ離れた商品を提示することは適切であるとはいえず，この負例データの選択方法は適切であるとは言えない．また，対象商品と同様の花材を使った商品は 1 つのみである．この観点からも，最適な負例データ選択方法ではないと考えられる．

次に，EC サイト上での閲覧履歴に基づき，同一顧客からの閲覧が少ない商品を負例データとした際の類似性分析結果について述べる．類似性上位 10 商品の内，4 商品が対象商品と同じカテゴリに属していた．本研究ではある用途で購買した顧客に対し，別用途での購買を促すための類似性評価を目的としており，この負例データ選択方法は研究目的に照らして不適切であると言える．また，完全にランダムに選択した場合においても同様の傾向がみられた．

4.3.2. 同一カテゴリ商品の高類似性商品に占める割合

図 3 に，対象商品と同一カテゴリの商品がどの程度類似性が高いと推定されているかについて，負例データ選択方法による差異を示す．横軸は各商品の類似性ランク，縦軸はその類似性ランクの商品までに，対象商品と同一のカテゴリ “お盆” に属する商品が，全体のうち，いくつ含まれているかを示すものであり，その最大値は 1 である．

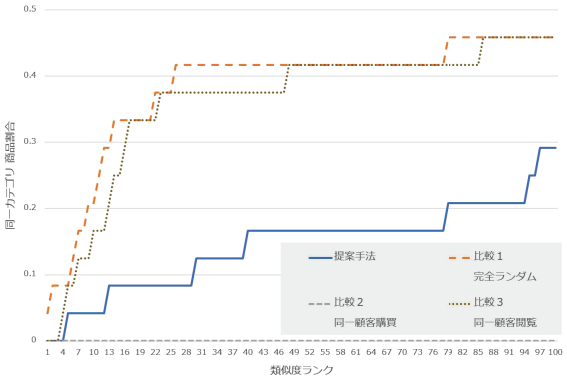

図 3: 同一カテゴリ商品の高類似性商品に占める割合

本研究では，同じ用途向け商品の類似性を低く評価することが必要であった．つまり，図 3 の縦軸の値は小さいほど，適切な負例選択方法であると考えられる．図 3 より，提案手法および比較手法 3 の同一顧客閲覧に基づく負例選択方法が，この観点からは適切であると言える．

4.3.3. 同一花材商品の高類似性商品に占める割合

次に高類似性商品の花材について，対象商品と同種類である“リンドウ”と“トルコ桔梗”のいずれかを用いている商品群のうち，高類似性商品群に含まれる割合を図 4 に示す．横軸は各商品の類似性ランク，縦軸は花材が同種類である商品割合を示しており，その最大値は 1 である．

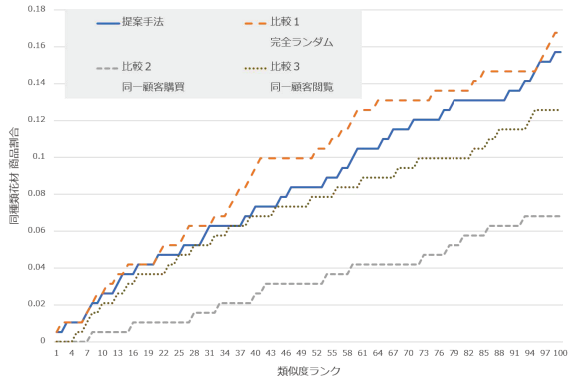

図 4: 同一花材商品の高類似性商品に占める割合

花材は顧客の購買有無に影響を与える要因の 1 つであり，同じ花材を使っている商品ほど購買されやすいであろうことを考えると，図 4 の縦軸の値は大きいほど研究目的に対して適切であると言える．従ってこの観点からは，提案手法および比較手法 1,2 が適切な負例選択方法であると判断される．

5. 考察

5.1. 提案手法の有効性

実験結果から，提案手法により分析した商品類似性は，定性的な観点や，花材の観点から対象商品と似ており，分類器の係数を用いた類似性分析により顧客の嗜好を評価できることを示した．また他の負例データ選択方法との比較から，同一カテゴリ商品を負例とすることが適切であることも示した．

特に図 4 より，同一花材を用いた商品群の類似性についても，比較手法 1，2 と同等の割合で高く評価できていることが分かった．これらと提案手法の違いは，同一カテゴリの商品が高類似性商品と分析されたか否かである．例えば，“母の日”といえば“カーネーション”といったように，カテゴリごとによく用いられる花材が存在する．そのため，同一カテゴリ商品の類似性を高く評価している比較手法 1，2 において，同一花材が高類似性商品に占める割合が高くなることは自然である．提案手法では，同一カテゴリの商品の類似性が低いにも関わらず，同一花材を用いた商品の類似性を高く評価できており，適切であったと言える．

5.2. 交互作用を用いることの有効性

本研究では説明変数間の交互作用を考慮するため，分類器に FM を用いた．交互作用による影響を明らかにするため，交互作用を考慮した場合としない場合について，各説明変数に対して商品毎に推定された回帰係数の分散を，降順で図 5 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，どちらも各説明変数の直接効果 $\boldsymbol{w}^A$ のみを示している．

ここで，分散の値が他の説明変数に比べて大きい説明変数

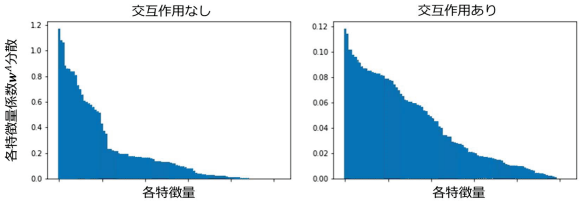

図 5: 説明変数別 係数分散

は，商品により異なる係数の値を持つということであり，すなわち商品特性に影響を与える説明変数である．図 5 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，一部の説明変数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，様々な説明変数によって商品特性をとらえることができたと考えられる．

5.3. 実ビジネスへの応用

従来のビジネスにおける課題として，ある用途で購買を行った顧客に対し，別の用途での購買を促す効果的なアプローチが求められている．現状では，それらは企業担当者の経験値に基づき，用途ごとに一律で商品提示を行うことがほとんどであるが，本提案手法で得られた類似性を用いることにより，顧客の嗜好を膨大な購買履歴データから抽出し，商品推薦などに役立てることが可能になる．このことは，1to1 マーケティングが求められる昨今の EC ビジネス業界において，コストを抑えながら適切なアプローチをするために有効である．

6. まとめと今後の課題

本研究では，同一顧客の購買間隔が長く，かつその嗜好が購買ごとに変化する商品群について，FM の主効果と交互作用の係数パラメータを用いて類似性を評価する手法を提案した．さらに，負例データの選択方法が重要であることを示した．本提案手法は，顧客ごとの購買点数に因らず適用することが可能であり，従来手法に比べ適用範囲を広げることができる．今後の課題として，対象商品の拡大と，実際のビジネスへの反映による有効性の検討が挙げられる．

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．

参考文献

[1] Ma, L. and Sun, B.: 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).

[2] Barkan, O. and Koenigstein, N.: Item2vec: Neural item embedding for collaborative filtering, *2016 IEEE 26th International Workshop on MLSP*, IEEE, pp. 1–6 (2016).

[3] Rendle, S.: Factorization machines, *2010 IEEE International Conference on Data Mining*, IEEE, pp. 995–1000 (2010).

[4] Barkan, O., Caciularu, A., Katz, O. and Koenigstein, N.: Attentive Item2vec: Neural attentive user representations, *2020 IEEE ICASSP*, IEEE, pp. 3377–3381 (2020).

[5] 佐和隆光：回帰分析，朝倉出版 (1979).