商品購買の周期性を考慮した消費者購買行動モデルと購買予測

Consumer Purchase Behavior and Prediction Considering Product Purchase Cycle

経営システム工学専攻　齋藤　烈也

1　研究背景と目的

スーパーマーケット市場では，新型コロナウイルス感染症の拡大の中でも売上を維持しており，自宅で食事をするライフスタイルが定着してきていることを裏付けている．その一方で，人口減少や高齢化問題等の課題に対しての危機意識も高まってきている．そのほかにも，スーパーマーケットなどの小売業は顧客の離脱や，他店舗への転入がしやすく，店舗の代替性が高いこと，また人口減少の進行により，将来的に必ず市場規模が縮小することが考えられる．大手スーパーマーケットのようなネットスーパーを持たない企業において，活用できるデータは実店舗での購買関連情報に限られ，その情報をいかに活用するかが，企業が顧客管理を行うにあたって重要な課題となっている．このような問題に対する有効なマーケティング施策として，既存顧客との関係性を重視したCRM が重要であり，顧客の特徴を捉えたバスケット予測のようなパーソナライズされたソリューションが必要となってきている．

そこで本研究では，スーパーマーケットにおいて顧客が次回の取引，すなわち次回購買時に購入する商品の予測を行うことを目的とする．顧客の購入する商品を予測するには，顧客の行動の変化，繰り返し起こる購買パターン，およびそれらの周期的な変化に対応しなければならないので，企業にとっては顧客情報を把握するために必要な知見であると考えられる．

2　先行研究と本研究の位置付け

顧客の購買行動の推測の一つに，バスケット予測が挙げられる．バスケット予測とは，各ユーザが次回の来店時にどの商品を購買するかを予測するものであり，バスケット予測に関する研究は，効果的なレコメンダシステムの開発を目的に行われる．レコメンダは，一般レコメンダ，シーケンシャルレコメンダ，パターンベースレコメンダ，ハイブリッドレコメンダの 4 つに分類される．

一般レコメンダは，協調フィルタリングに基づいており，一般的な顧客の好みに基づいて顧客に対するレコメンデーションを作成する．また，一般レコメンダは，顧客の嗜好に基づいてレコメンデーションを行うが，商品間の繋がり等に関する逐次的な情報を考慮せず，時系列の要素を憂慮していない．対照的に，シーケンシャルレコメンダは，時系列推移に基づいており，連続的な情報と，最近の購買履歴を利用して，顧客に対するレコメンデーションを作成する．パターンベースレコメンダは，全ての顧客の購買履歴から抽出された頻出アイテムセットに基づいて予測を行うが，逐次的な情報は破棄される特徴がある．パターンベースのアプローチは，基本的なパターンを抽出するために Apriori アルゴリズム [1] を用いることが多いことで知られている．

ハイブリッドアプローチ [2] は，一般レコメンダとシーケンシャルレコメンダを組み合わせたものである．Rendle et al.[2] は行列因子分解とマルコフ連鎖を組み合わせて発展させた Factorized Personalized MC (FPMC) を用いて分析を行っている．また，モデルのパラメータ学習のために，Bayesian Personalized Ranking (BPR) を用いて，顧客が商品を購買する確率を計算している．

以上のモデルでは，商品の相互影響の要因を考慮できていない点，順序情報や顧客の再来性を考慮できていない等様々な課題がある．これらを踏まえ本研究では，同じバスケット内の商品の相互作用と，連続するバスケット内の商品間の相互作用だけでなく，顧客同士の類似性をモデル化する．

3　使用するデータ

本研究では，複数店舗のスーパーマーケットを展開する企業から提供いただいた ID 付き POS データを分析に使用する．

表 1: データ概要

データ概要	スーパーマーケットの ID 付 POS データ
期間	2014 年 4 月 1 日〜2014 年 12 月 31 日
対象商品	804 品
対象顧客	7,381 人
販売形式	実店舗のみ

表 2: バスケット概要

	バスケットサイズ	来店回数
平均値	7.6 品	34.6 回
中央値	6.8 品	22 回

4　分析手法

本研究において用いる TARS (Temporal Annotated Recurring Sequences) [3] は，FP-Growth アルゴリズム [4] を拡張したものであり，深さ優先型の tree 構造モデルである．このモデルの問題点としては，購入情報の少

ない顧客については予測精度が低くなることが挙げられる．これに対して本研究では，セグメンテーション手法である pLSA (probabilistic Latent Semantic Analysis) [5] を用いることで顧客間の相互関係をモデル化し，分析に用いる．

4.1 Temporal Annotated Recurring Sequences

TARS の抽出手順を Algorithm 1 に示す．

Algorithm 1 $extractTars(B)$

1: $\mathcal{S} \leftarrow extractBaseSequences(B)$;
2: $\{\delta_S^{max}\}, \{q_S^{min}\}, \{p_S^{min}\} \leftarrow parametersEstimation(B, \mathcal{S})$;
3: $\mathcal{S}^* \leftarrow sequenceFiltering(B, \mathcal{S}, \{\delta_S^{max}\}, \{q_S^{min}\}, \{p_S^{min}\})$;
4: $\mathbf{\Psi} \leftarrow buildTars\text{-}Tree(B, \mathcal{S}^*, \{\delta_S^{max}\}, \{q_S^{min}\}, \{p_S^{min}\})$;
5: $\mathbf{\Gamma} \leftarrow extractTarsFromTree(\mathbf{\Psi})$;
6: **return** $\mathbf{\Gamma}$;

最初のステップでは，購入履歴 B から，シーケンスを抽出する．次に，B に関する各シーケンス $S \in \mathcal{S}$ のパラメータ $\{\delta_S^{max}\}, \{q_S^{min}\}, \{p_S^{min}\}$ を推定する．そして，これらのパラメータに関して，シーケンス S をフィルタリングし，ベースとなるシーケンス S^* を抽出する (それ以外のシーケンスは検索効率を上げるために廃棄する)．最後に，シーケンス S^* と各パラメータによって $TARS\text{-}Tree$ を構築し，FP-Growth アルゴリズムに従って $\mathbf{\Psi}$ を作成，最後に TARS の集合 $\mathbf{\Gamma}$ を抽出する．

4.1.1 TARS Based Predictor

Algorithm 1 のような手順で抽出されたアクティブな TARS を元に，商品ごとにアイテムスコアを算出する．その手順を以下に示す．

Algorithm 2 $calculateItemScore(B, \hat{\mathbf{\Gamma}}, Q)$

1: $\Omega \leftarrow \emptyset$;
2: **foreach** $i \in I$ **do** $\Omega_i \leftarrow 0$;
3: **for** $\gamma = (S = \langle X, Y \rangle, \alpha, p, q) \in \hat{\mathbf{\Gamma}}$ **do**
4: **foreach** $i \in Y$ **do** $\Omega_i \leftarrow \Omega_i + (q - Q_\gamma)$;
5: **for** $i \in \{i | \exists \gamma = (S = \langle X, Y \rangle, \alpha, p, q) \in \hat{\mathbf{\Gamma}}, i \in Y\}$ **do**
6: $\Omega_i \leftarrow \Omega_i + sup(i)$
7: **return** Ω;

まず，各アイテム Ω_i のスコアをゼロにする．次に，アイテム $i \in Y$ を含むすべてのアクティブな TARS γ に対して，γ の平均的な発生回数 q と，γ のシーケンスが最近の購入履歴で発生した回数を示す Q_γ との差で，Ω_i を増加させる．最後に，Ω_i は，アクティブな TARS におけるアイテム i との商品の組み合わせによって，Ω_i を加点する．以上のようなアルゴリズムで算出されたアイテムスコアを基に予測する商品を決定する．予測するバスケットサイズについては，各顧客のバスケットサイズのヒストグラムを作成し，スタージェスの公式から適正な階級を推定することで，最大の階級値をとった値をバスケットサイズとして設定する．

4.2 提案モデル

本研究では，前節で記述した TBP (TARS Based Predictor) のアイテムスコア算出アルゴリズムに，pLSA を用いる．pLSA とは，確率的潜在意味解析法と呼ばれる次元圧縮の手法であり，クラスタリングの手法として使用される．pLSA では，潜在クラスの下で顧客とアイテムの共起が発生していると仮定し，その共起関係を条件付き確率で表現する．つまり，これまでの購買履歴を基に，潜在クラスを仮定することで，顧客の購買傾向およびアイテムの類似性をクラスタリングする．

TBP の問題点として，購買回数が少ない顧客については予測精度が低くなる点が挙げられる．これは，購買情報が少ないことによって TARS そのものが抽出できていないことに起因している．それに対して，提案モデルでは pLSA を用いることで潜在クラスを仮定し顧客の購買傾向及び商品の類似性をクラスタリングする．それにより，従来モデルより規則的なストラクチャーを構築する．

4.2.1 Probabilistic Latent Semantic Analysis

以下では，顧客の購買履歴を用いた pLSA による潜在クラス生成について述べる．

X 人の顧客と Y 個の商品を対象とし，それぞれ x_c $(c = 1, 2, ..., X)$ と y_i $(i = 1, 2, ..., Y)$ とする．また，潜在クラス数を Z とし，潜在クラス k を表す変数を z_k $(k = 1, 2, ..., Z)$ とする．ここで，x と y の共起確率 $P(\boldsymbol{x}, \boldsymbol{y})$ を式 (1) のようにモデル化する．$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ は，それぞれを含むベクトルとする．

$$P(x_c, y_i) = \sum_{k=1}^{Z} P(x_c|z_k)P(y_i|z_k)P(z_k) \tag{1}$$

また，x_c と y_i の同時出現頻度を N_{ci} とすると，その対数尤度は式 (2) となる．

$$\begin{aligned} l &= \sum_{c=1}^{X}\sum_{i=1}^{Y} N_{ci} \log P(x_c, y_i) \\ &= \sum_{c=1}^{X}\sum_{i=1}^{Y} N_{ci} \log \left\{ \sum_{k=1}^{Z} P(x_c|z_k)P(y_i|z_k)P(z_k) \right\} \end{aligned} \tag{2}$$

この潜在クラスモデルの対数尤度は EM アルゴリズムにより最大化できる．推定すべき条件付き確率は $P(\boldsymbol{x}, \boldsymbol{z})$，$P(\boldsymbol{y}, \boldsymbol{z})$，$P(\boldsymbol{z})$ であり，それぞれに初期値を乱数で与えると，式 (2) の変形から潜在変数の条件付き確率は，以下のように計算できる．

$$\begin{aligned} P(x_c|z_k, y_i) &= \frac{P(x_c, y_i, z_k)}{P(x_c, y_i)} \\ &= \frac{P(x_c|z_k)P(y_i|z_k)P(z_k)}{\sum_{k=1}^{Z} P(x_c|z_k)P(y_i|z_k)P(z_k)} \end{aligned} \tag{3}$$

また，ラグランジュの未定乗数法から反復計算ステップの式 (3) の条件付き確率を最大化する条件付き確率は，以下のように計算できる．

$$P(x_c|z_k) = \frac{\sum_{i=1}^{Y} N_{ci} P(z_k|x_c, y_i)}{\sum_{c=1}^{X} \sum_{i=1}^{Y} N_{ci} P(z_k|x_c, y_i)} \tag{4}$$

$$P(y_i|z_k) = \frac{\sum_{c=1}^{X} N_{ci} P(z_k|x_c, y_i)}{\sum_{c=1}^{X} \sum_{i=1}^{Y} N_{ci} P(z_k|x_c, y_i)} \tag{5}$$

$$P(z_k) = \frac{\sum_{c=1}^{X} \sum_{i=1}^{Y} N_{ci} P(z_k|x_c, y_i)}{\sum_{c=1}^{X} \sum_{i=1}^{Y} \sum_{k=1}^{Z} N_{ci} P(z_k|x_c, y_i)} \tag{6}$$

この反復を尤度が収束するまで実行することで，各条件付き確率を推定することができる．また，本研究においては潜在クラス数は赤池情報量規準とベイズ情報量基準の 2 つで評価し，12 クラスに決定した．

4.2.2 pLSA を用いたアイテムスコアリング

本節では，TBP のアイテムスコア算出アルゴリズムに pLSA を用いたスコアリング手法について記す．提案モデルにおいては，pLSA を用いて潜在クラスを仮定し，それらの条件付き確率の値を用いてアイテムスコアを算出する．以下にその概要を記載する．

$$\Omega^*_{c_i} = \Omega_{TBP_{c_i}} \times (\ \Omega_{pLSA_c} \times\ \Omega_{pLSA_i}) \tag{7}$$

ここで，$\Omega_{TBP_{c_i}}$ は TBP によって算出されたアイテムスコアであり，顧客，アイテムに関するウェイトを表す．また，$\Omega_{pLSA_c} = \{\Omega^{(1)}_{pLSA_c}, ..., \Omega^{(Z)}_{pLSA_c}\}, \Omega_{pLSA_i} = \{\Omega^{(1)}_{pLSA_i}, ..., \Omega^{(Z)}_{pLSA_i}\}$ であり，各 Ω_{pLSA} は以下のように定義される．

$$\Omega^{(z)}_{pLSA_c} = \frac{p(x_c|z)p(z)}{\sum_z p(x_c|z)p(z)} \tag{8}$$

$$\Omega^{(z)}_{pLSA_i} = \frac{p(y_i|z)p(z)}{\sum_z p(y_i|z)p(z)} \tag{9}$$

以上のスコアリング手法を用いることでアイテムスコアを再計算し，バスケット予測を行う．

5 結果と考察

本節では，第 5 章で扱った TBP による従来モデルと，従来モデルに pLSA による購買パターンの違いを考慮した提案モデルの精度比較及び pLSA を用いたことによる精度向上の要因について述べる．

表 3: 従来 TBP モデルの購買回数毎の精度一覧

	～9	10～19	20～29	30～39	40～	全体
recall	0.128	0.148	0.181	0.208	0.214	0.172
precision	0.274	0.287	0.272	0.274	0.259	0.273
fscore	0.161	0.178	0.202	0.221	0.221	0.194
顧客数	1779	1703	1105	828	1966	7381

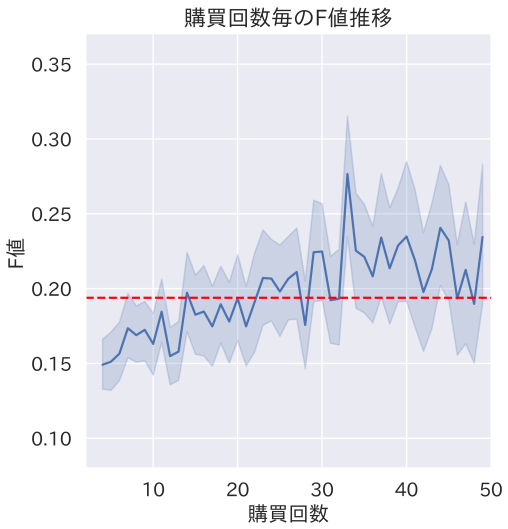

図 1: 従来モデルにおける購買回数毎の F 値の折れ線グラフ

表 3 の結果より，F 値については平均で 0.194 となっていることがわかる．従来モデルでも課題としてもあったように，購買回数の少ない顧客についてはうまく予測できていないことが，図 1，表 3 からも読み取れる．

次に，提案モデルの精度について表 4 に示す．

表 4: 提案モデルの精度一覧

	～9	10～19	20～29	30～39	40～	全体
recall	0.180	0.185	0.248	0.305	0.348	0.250
precision	0.263	0.268	0.230	0.213	0.171	0.229
fscore	0.194	0.194	0.212	0.226	0.203	0.203
顧客数	1779	1703	1105	828	1966	7381

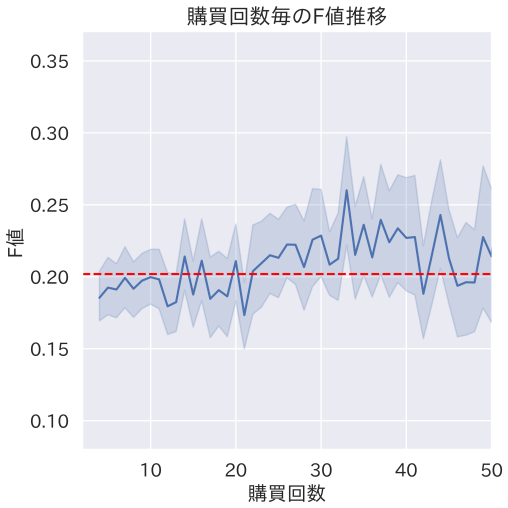

図 2: 購買回数毎の F 値の折れ線グラフ

表 4 の結果より，F 値については平均で 0.203 となっていることがわかる．また，購買回数が 10 回未満のセグメントの F 値は，0.194 であることがわかる．これは，他のセグメントと比較しても，そこまで差がない結果となっており，従来モデルの課題であった，購買回数の少

ない顧客においてもうまく予測できていることが考えられる．

アイテムスコアを算出する際に，pLSA によって潜在クラスへの所属確率を求め，アイテムスコアに反映する．アイテムスコア算出のアルゴリズム上，顧客の該当潜在クラスへの所属確率と購買すると予測する商品の該当潜在クラスへの所属確率が共に高い場合により高いスコアが算出される．すなわち，ある顧客が該当商品を購買することの尤もらしさをアイテムスコアに反映している．以上のことを踏まえ，以下に，同一顧客において従来モデルでは購買すると予測できず，提案モデルでは購買すると予測可能となった商品に関する各潜在クラスへの所属確率の一部をまとめる．

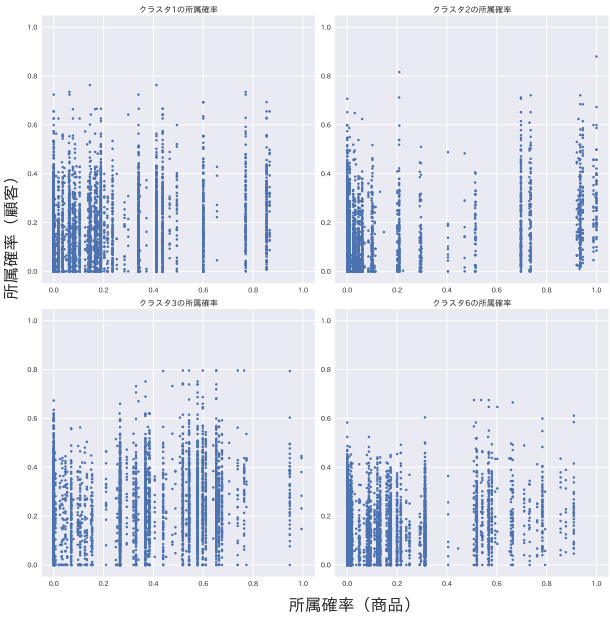

図 3: 各クラスタにおける所属確率

クラスタ 1，クラスタ 2，クラスタ 3，クラスタ 6 は商品の所属確率については極端な偏りはなく，顧客の所属確率が高い値を取るケースがあることがわかる．顧客と商品の該当潜在クラスが共に高い場合に，より高いアイテムスコアが算出されることからクラスタ 1，クラスタ 2，クラスタ 3，クラスタ 6 に高い所属確率の値を持つ商品は，アイテムスコアが高くなっていると考えられる．

ここで，pLSA によるアイテムスコア算出アルゴリズムに関する解釈を述べる．まず，提案モデルにおいて用いている pLSA を用いたアイテムスコアでは，顧客と商品の各クラスタへの所属確率が共に高い場合にアイテムスコアが高くなり，どちらか一方が低い場合にアイテムスコアは低くなることがわかっている．従来モデルにおいては，TARS そのものが抽出できなかった場合，それまでの購入情報をもとに累積購買点数をアイテムスコアに用いる構造となっている．その結果，購入情報の少ない顧客は TARS が抽出できず，予測精度も低い結果となっている．一方で，提案モデルにおいてはそのような顧客に対して潜在クラスを仮定することで顧客の相互作用を考慮し，予測をパーソナライズするのではなく，一般的な購買傾向をもとにアイテムスコアを算出していることが精度向上の要因であると考えられる．

6　まとめと今後の課題

本研究では，国内において実店舗のみを販売チャネルにもつスーパーマーケットの ID 付き POS データを用い，個人単位でのバスケット予測を行うことを目的に分析を行った．バスケット予測には，TARS を基にした購買予測モデルである TBP と，TBP によってパーソナライズされた観点だけでなく一般的な観点 (潜在性) を考慮することを志向した提案モデルの 2 種を用いた．提案モデルでは，pLSA を用いることで潜在クラスを仮定し，顧客の購買傾向及び商品の類似性をクラスタリングした．それにより，従来モデルよりも高い精度を得ることが可能となった．

今後の課題について，本モデルでは天気や来店時間帯などの外的要因については考慮していない．特に，予測のパーソナライズを志向した TARS を用いたモデルにおいて，それらの要因を考慮することは顧客理解を深める点でも必要な改善であると考えられる．

参考文献

[1] Agrawal, R., and Srikant, R. “Fast Algorithms for Mining Association Rules,” *Proc. 20th Int. Conf. Very Large Data Bases*, pp. 487–499, 1994

[2] Rendle, S., Freudenthaler, C., and Schmidt-Thieme, L., “Factorizing Personalized Markov Chains for Next-Basket Recommendation,” *Proc. 19th Int. Conf. on World Wide Web*, pp. 811–820, 2010

[3] Guidotti, R., Rossetti, G., Pappalardo, L., Giannotti, F., and Pedreschi, D., “Personalized Market Basket Prediction with Temporal Annotated Recurring Sequences,” *IEEE Transactions on Knowledge and Data Engineering*, Vol.31, No.11, pp. 2151–2163, 2018

[4] Han, J., Pei, J., and Yin, Y. “Mining Frequent Patterns without Candidate Generation,” *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Vol.29, No.2, pp. 1–12, 2000

[5] Hofmann, T. “Probabilistic Latent Semantic Indexing,” *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, pp. 289–296, 1999