日本オペレーションズ・リサーチ学会
2021年 秋季研究発表会

顧客の購買行動と購買傾向に着目した顧客来店予測

申請中　中央大学大学院　*齋藤烈也　SAITO Retsuya
05000907　東海大学　大竹恒平　OTAKE Kohei
01405390　中央大学　生田目崇　NAMATAME Takashi

1. 背景

成熟した市場においては，一般に，新規顧客の獲得にかかるコストが大きいことが知られており，既存顧客との関係性を重視した CRM (Customer Relationship Management) が重要視されている．CRM の観点を用いた有効なマーケティング施策の一つに，顧客のセグメンテーションがある．中でも，顧客行動に関する情報を用いたセグメンテーションは，EC などのインターネットを利用したオムニチャネルだけではなく，小売業を営む実店舗においても注目を集めている．

そこで本研究では，スーパーマーケットの ID 付き POS データを用い，顧客行動をセグメンテーションした場合に，購買予測の精度が向上する可能性について論じる．

2. 先行研究

セグメンテーションを用いた研究として，トランザクションデータと顧客の生活調査データから，商品カテゴリと顧客のライフスタイルに関する潜在クラス変数を作成し，ベイジアンネットワークを用いて関係をモデル化することにより，顧客のライフスタイル，時間帯等の多様な条件下でどのような潜在クラスの商品カテゴリが購買されるのかを明らかにした研究 [1] がある．このように，潜在クラスの概念を機械学習に取り入れた研究は存在しているが，説明変数としての有用性について考察している研究は少ない．

3. 使用データ

本研究では，スーパーマーケットの ID 付き POS データを用いる．詳細を表 1 に記載する．

表 1: 使用データの詳細

対象期間	2014/4/1 ~2016/3/31
対象人数	15,783
対象商品数	106,743
対象会計数	1,218,470

4. 分析手法

本研究は，将来一ヶ月間の購買回数についての予測を，購買状況を基にした顧客の潜在性に考慮して行う．予測については，決定木ベースの強化学習の一つである XGBoost (eXtreme Gradient Boosting) [2] を用いて行う．使用データから顧客ごとに 24ヶ月間のトランザクションデータを取得し，その期間を最初の 23ヶ月とそれ以降の一ヶ月間に分割する．顧客の将来の行動を予測するため，23ヶ月間を訓練データ，一ヶ月間をテストデータとしている．

4.1. 潜在変数の導入

本研究では，モデルの説明変数として，顧客と顧客行動に関する要因について潜在性を仮定した変数を作成する．そこで，1 つの潜在変数を仮定した，顧客，商品，潜在クラスの関係を pLSA (Probabilistic Latent Semantic Analysis) を用いてモデル化する．X 人と売上個数上位 1,000 商品 Y 個を対象とし，それぞれ x_i $(i = 1, 2, ..., X)$ と y_j $(j = 1, 2, ..., Y)$ 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 $P(x, y)$ を (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．

$$P(x_i, y_j) = \sum_{k=1}^{Z} P(x_i|z_k)P(y_j|z_k)P(z_k) \qquad (1)$$

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

$$\begin{aligned} l &= \sum_{i=1}^{X}\sum_{j=1}^{Y} N_{ij} \log P(x_i, y_j) \\ &= \sum_{i=1}^{X}\sum_{j=1}^{Y} N_{ij} \log \left\{ \sum_{k=1}^{Z} P(x_i|z_k)P(y_j|z_k)P(z_k) \right\} \end{aligned} \qquad (2)$$

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を計算することができる．

本研究においては，顧客は 15,783 人を対象とし，

顧客と商品との潜在クラス推定の他に，顧客と来店時間帯との関係，顧客と来店曜日との関係を同様にして推定する．また，潜在クラス数は AIC と BIC の 2 つで評価し，クラス数を決定している．表 2 に各潜在クラス数を示す．

表 2: 顧客と各要因間の潜在クラス数

商品	来店時間帯	来店曜日
8	9	5

4.2. モデル作成

本研究において用いる説明変数について，表 3 に示す．

表 3: モデルに使用した変数

変数名		説明
$Buy\_times_{t_2}$	目的変数	t_2 の期間に購買した回数
$Buy\_times_{t_1}$	基本	t_1 の期間に購買した回数
$Interval$		t_1 の期間中の平均来店間隔（対数）
$MinPrice$		t_1 の期間中に購買した合計金額の最小値
$MaxPrice$		t_1 の期間中に購買した合計金額の最大値
$DiffPrice$		t_1 の期間中に購買した合計金額の最大値と最小値の差
$DiffDays$		t_1 の期間中の最後の来店日と t_2 の期間までの残日数
$Itemavg$		t_1 の期間中の 1 取引での平均購買品目
$MostHour$	最頻値	t_1 の期間中に最も来店した時間
$MostWeek$		t_1 の期間中に最も来店した曜日
$BuyingItems$	潜在	pLSA によって分類された商品間の潜在クラスへの所属確率
$VisitTime$		pLSA によって分類された来店時間帯間の潜在クラスへの所属確率
$VisitWeek$		pLSA によって分類された来店曜日間の潜在クラスへの所属確率

比較に用いるパラメトリックな統計モデルの予測には GLM(Generalized Linear Model) による回帰分析モデルを用い，最尤推定による推定値を利用する．解析では，目的変数として，累積購買回数を用い，ポアソン回帰モデルにより予測を行なう．なお，元データの 80%を訓練データ，残りの 20%をテストデータに適用することで，RMSE を算出する．

5. 結果と考察

予測期間の購買回数について，XGBoost と GLM を用いて予測した結果の RMSE を表 4 に示す．また，精度の最も高かった基本変数と潜在変数を説明変数に加えたモデルの予測期間の購買回数の散布図を図 1 に示す．

表 4: 各モデルパターンにおける RMSE の結果

予測方法	基本＋最頻値	基本＋潜在	基本＋最頻値＋潜在
XGBoost	2.43	1.92	2.04
GLM	2.70	2.25	2.18

表 4 から，RMSE の値が最も改善したのは，XGBoost を用いた際の，基本変数と潜在変数のみを

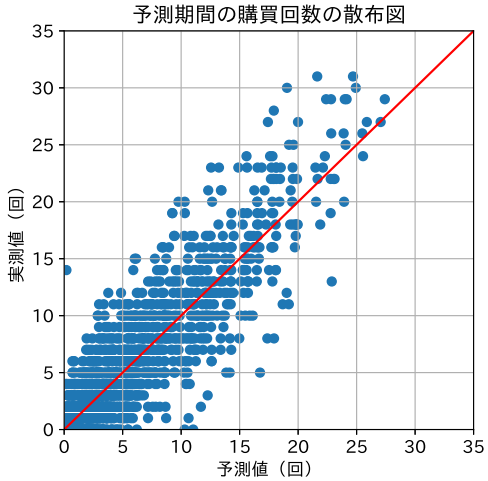

図 1: 予測期間の購買回数の散布図

説明変数に加えたモデルであることがわかる．また，XGBoost と GLM の予測結果の比較では，いずれの場合においても常に XGBoost の方が精度が改善していることも読み取れる．以上のことから，XGBoost の方が，GLM より精細に事象を表現できていると考えられる．

次に，本研究で用いた潜在変数の有用性について述べる．pLSA による潜在性の表現は，顧客と商品，顧客と時間などの関係を潜在クラスを仮定することでモデル化し，各顧客における潜在クラスの所属確率を低次元で表現することができる．また，表 4 からも，ダミー変数を使用している最頻値変数と基本変数で構成されたモデルよりも精度が改善されていることがわかる．このことから，カテゴリカルデータに対して，データのスパース性や精度の面において，従来から扱われているダミー変数よりも優れているという結果が示された．また，結果解釈の面では，購買傾向から算出される顧客ごとの潜在クラスへの所属確率を変数として用いているので，単に購買生起を変数化するよりも有用な顧客理解ができると考えられる．

参考文献

[1] 石垣司, 竹中毅, 本村陽一, “日常購買行動に関する大規模データの融合による顧客行動予測システム, ” 人工知能学会論文誌, Vol. 26, No.6D, pp. 670–681, 2011

[2] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System, ” *Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016