6A-04

非負値行列因子分解による顧客購買パターン抽出と顧客生涯価値予測

蓮本　恭輔†　　雲居　玄道†　　後藤　正幸‡

†早稲田大学大学院創造理工学研究科　　‡早稲田大学創造理工学部

1. 研究背景・目的

企業経営において顧客生涯価値(Customer Lifetime Value: CLV)を把握することは，マーケティング施策決定のみならず，基本戦略の策定において重要な意味を持つ．これは CLV の大きさが将来の企業の収益を左右するためである．しかし CLV は顧客の将来の行動の結果であり，それを把握するには何らかの予測モデルが必要となる．これに対し，実務レベルでは購買履歴から優良顧客を判別する手段として RFM 分析によるセグメンテーションが頻繁に用いられているが，CLV を予測するための指標としてそのままでは不十分であることも指摘されている[1]．実際には，単純な購買回数や金額よりも，どの店舗で，どのような商品を，どのような頻度で購入するといった顧客の購買パターンが CLV に影響しており，それらを考慮した予測モデルが必要である．そこで本研究では非負値行列因子分解(NMF: Non-negative Matrix Factorization)[2]を用いて顧客の購買パターンを抽出し，CLV 予測に適用する方法を提案する．加えて，実データに対して提案モデルを適用し，得られる結果に対して考察を与える．

2. RFM 分析と顧客生涯価値(CLV)

RFM 分析は，最新購買日(Recency)，総購買回数(Frequency)，総購買金額(Monetary)の 3 指標をもとに顧客をランク付けし，優良顧客セグメントを判別する手法である．一方で CLV はその顧客が対象の事業者に生涯にわたってもたらす収益を指し，未来の購入頻度や顧客の生存確率によって決まるため，何らかのモデル化や予測が必要となる．RFM 分析は過去の購買データに対し適用でき，かつ簡便な手法であることから，RFM 指標を用いて CLV の予測やモデル化する試みが様々な形で行われている[1]．本研究では RFM 指標と NMF で抽出された購買パターン情報を回帰分析に適用し CLV の予測を行う．

3. 準備

3.1. 対象事例

本研究では協力事業者より提供された新規顧客の 1 年分の購買情報及びその翌年 1 年間の CLV を対象とした．この事業者のサービスは様々な店舗で利用されており，顧客がこの事業者を通して商品やサービスの購入を行うと，事業者には手数料収入が入る．顧客が継続的に商品を購入し，売上に貢献すれば手数料収入も増えるため，収益は基本的に売上に比例している．一方で店舗毎に手数料率が異なるため，どの店舗で購入されたかによって事業者の収益は異なり，また顧客の購入頻度や購入単価も店舗毎に異なる．事業の性質上，取引に伴う商品レベルの情報はない一方で複数の店舗にまたがる取引情報は存在する．また事前分析により，購入数や購入金額に加えて複数店舗での購入が CLV に影響があることも判明している．そこで購入先や時系列の購入傾向から購買パターンを抽出する．

3.2. Non-negative Matrix Factorization (NMF)

非負値行列因子分解(NMF: Non-negative Matrix Factorization)は，非負の行列データを低ランクの非負行列の積に因子分解することで，それらのデータに潜在的なパターンを抽出する手法である．比較的単純なアルゴリズムでデータの特徴を抽出する手法として，画像解析やテキスト分類など様々な問題に適用されている．本研究では提供された購買情報より，購入先と時系列の購入傾向を NMF でパターン抽出する．ここでは次の 2 つを入力行列として定義した．1. ***顧客×購入月***，2. ***顧客×購入店舗カテゴリ***（以下，購入月行列，店舗行列とする）．また行列の要素は購買回数，購買金額でそれぞれ 2 パターン作成する．ただし店舗行列においては，購入先の店舗数が膨大であり，そのまま統計モデルに取り込むことはできない．そこで店舗が取り扱っている商品（ファッション，食品など）によって店舗カテゴリを構成し，顧客がどの店舗カテゴリの店舗から購入しているのかを行列表現して NMF で行列分解する．以上のもとで，入力非負行列を$\boldsymbol{X} \in \mathbb{R}_{+}^{I \times J}$とし，それが顧客（$\boldsymbol{U} \in \mathbb{R}_{+}^{I \times K}$）と，購入月または購入店舗カテゴリ（$\boldsymbol{V} \in \mathbb{R}_{+}^{K \times J}$）に特徴数 Kのもと分解されるとすると行列は式(1)で近似される．

$$\boldsymbol{X} \cong \boldsymbol{UV} \tag{1}$$

4. 提案手法

本研究では購買パターンを考慮した CLV の回帰予測モデルを提案する．提案手法は NMF による購買パターン抽出とランダムフォレスト回帰(RF 回帰)[3]による CLV 予測の 2 フェーズに分かれる．最初に購入月行列と店舗行列を NMF で行列分解し，時系列と購入先の購買パターンを抽出する．NMF の出力結果として顧客×クラス，クラス×購入月/購入店舗カテゴリの行列が作成されるが顧客×クラスを説明変数として RFM 指標と共に使用し，CLV の予測を行う．

5. 実験データ

実験データは 16,009 人分の新規顧客データと初

Extraction of customers' purchasing patterns and prediction of customer lifetime value by non-negative matrix factorization

† Kyosuke Hasumoto. Gendo Kumoi ・ WASEDA University, Graduate School of Creative Science and Engineering

‡ Masayuki Goto ・ WASEDA University, School of Creative Science and Engineering

回購入後 1 年間の購買情報，さらに翌年 1 年間の CLV である．購買情報は顧客毎に月単位，購入店舗カテゴリで集計し，それぞれデータ件数は 69,768 件，27,659 件となる．これらの入力行列を購入回数，金額で 2 種類作成した．購入月行列は初回購入月を 1 として 12 列である．また店舗行列の購入店舗カテゴリ数は 46 であった．同様に顧客毎の RFM 指標を作成したが事前分析より購入店舗数の CLV への影響を考慮し，RFM3 指標に追加して 4 指標とした（以下 RFMs 指標とする）．回帰分析は 10 fold の交差検証を行い，すべてのデータを学習・テストに使用した．

6. 実行結果と考察

6.1. NMF 実行結果と考察

クラス数 K は 1〜30 の範囲で事前分析を行い，NMF 実行時の残差と実行結果の解釈容易性から定性的な評価により，購入数，購入金額にかかわらず，購入月行列は K=5，店舗行列は K=10 に設定した．購入月行列を分解した結果を時系列でグラフ化したものを図 1 に，クラス別の CLV を示したものを表 1 に示す．CLV 値については平均を 1 とし正規化している．

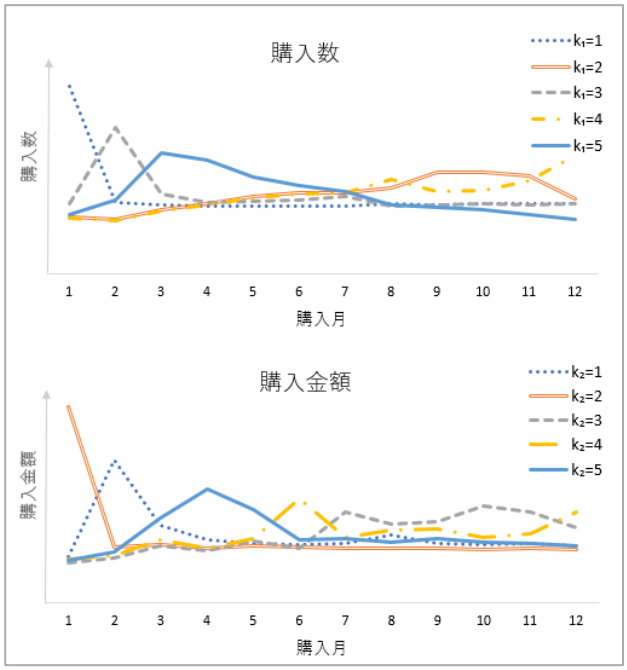

図 1 購入月行列の分解結果

表 1 購入月行列の分解結果：クラス別 CLV

Class(k)		1	2	3	4	5
Average CLV	購入数	0.84	1.04	0.92	1.10	0.95
	購入金額	0.90	0.94	1.06	1.01	0.93

購入数，購入金額ともに利用初期に購入が多く後半にかけて減っていく場合は CLV が低く，逆にコンスタントに購入があり顧客の成長が見られるようなクラスでは CLV が高くなっていることがわかる．また購入月 12 における数値が高いクラスは CLV も高い傾向にあり，Recency がその後の CLV にポジティブな影響を与えていることを示している．表 2 は同様に店舗行列の分解結果を示しているが，クラス間で CLV により大きな差が出ることが確認された．CLV の低いクラスはデジタルコンテンツなど比較的単価の低い店舗からの利用が多く，逆に CLV の高いクラスは嗜好品など高単価商品・サービスを扱う店舗が多いことから単価の違いによる影響が見えた．

表 2 店舗行列の分解結果：クラス別 CLV

Class(k)		1	2	3	4	5
Average CLV	購入数	0.86	0.92	1.54	1.00	0.73
	購入金額	0.58	0.89	0.88	1.35	0.87

Class(k)		6	7	8	9	10
Average CLV	購入数	1.10	1.13	0.96	0.73	0.90
	購入金額	1.19	1.06	0.97	1.39	0.93

6.2. ランダムフォレスト回帰の結果と考察

NMF の行列分解より各行列に対してクラス数 K 分の要素を新たな特徴量と捉えることができる．具体的には，顧客毎に購入月行列から 5 次元，店舗行列から 10 次元，さらに購買数，購買金額のそれぞれ 2 要素で合計 30 次元の特徴ベクトルが得られる．これらの特徴量と RFMs 指標 4 変数を説明変数として RF 回帰で CLV 予測を行った．比較として RFMs 指標のみを使用した場合，RFMs 指標と NMF の入力行列（購入月行列 12 列，店舗行列 46 列が購入数と金額で 116 変数）を使用した場合の実行結果も合わせて表 3 に示す．()内の数字は説明変数数を表す．なおランダムフォレストは決定木数を 500，基準をジニ係数とした．決定係数は RFMs 指標と NMF 出力結果の組み合わせが最も良い数値となった．購買パターンを考慮した予測モデルにすることで単純な購買データからの予測より精度が改善したと考えられる．

表 3 ランダムフォレスト回帰実行結果

説明変数	決定係数
RFMs指標(4)	0.2513
RFMs指標(4)+NMF入力行列(116)	0.2402
RFMs指標(4)+NMF出力結果(30)	0.2657

7. まとめと今後の課題

本研究では顧客の潜在的な購買パターンを NMF に抽出し，その結果を回帰分析で利用して予測精度を向上する手法を提案した．購買パターンの抽出では顧客のどのような購買行動が CLV に影響があるのかを明らかにし，マーケティング施策を立案する上で有益な示唆を得ることができた．また予測精度においては一定の改善を示すことができた．しかし絶対的な決定係数の数値が低く，さらなるモデルの改良や最適なクラス数，パラメータの決定方法など研究を進めていく予定である．

参考文献

[1] 阿部誠: RFM データを用いた顧客生涯価値の算出: 既存顧客の維持介入と新規顧客の獲得. マーケティングジャーナル, Vol.34, No.1, pp.73-90 (2014).
[2] Lee, D.D. and Seung, H.S.: Algorithms for non-negative matrix factorization, Proc. NIPS'01, pp.556-562 (2001).
[3] Breiman, L.: Random Forests, Machine Learning, Vol.45, No.1, pp.5-32 (2001).