論文・事例研究

グラフ研磨を利用した顧客クラスタリングによる多様性を考慮した特徴抽出

中原　孝信，丸橋　弘明，羽室　行信，宇野　毅明

1. はじめに

データマイニングの分野では相関ルールに基づく手法が提案されており，その一つに顕在パターンと呼ばれる方法がある [1]．顕在パターンは，目的変数として正例・負例を定義し，正負の集合にそれぞれ特徴的なルールを抽出する方法である．目的変数は，購入・非購入のような事象を単純に正例・負例と定義することが多く，それらを分類するためのルール発見が行われてきた．本研究では，ある商品の購入を正例，その非購入を負例として単純に定義するのではなく，顧客の類似性を考慮したセグメンテーションを行い，類似する顧客集合の中で正例・負例を定義してその違いを明らかにする．

たとえばチャーン分析を実施する際に，ブランドを乗り換えた顧客集合と，その否定である継続集合を比較して顕著なルールを抽出するよりも，乗り換えた顧客と「乗り換えそうで継続している顧客」を比較し，それらの違いを得ることのほうが経営的な観点からは有用なルールが抽出できると考えられる．つまり，継続顧客の中にもロイヤルティの高い顧客や低い顧客など多様な顧客が存在しており，それらの顧客を同一の集合として分析することは，判別力の高いモデルが構築できたとしても，それが経営的な施策に結びつく有用なルールかどうかは疑問が生じる．これまでデータマイニングでは，得られるルールは当たり前のルールが多いと言われてきたが，それはこのような比較の方法にも問題があるのではないだろうか．

なかはら　たかのぶ
専修大学
まるはし　ひろあき，はむろ　ゆきのぶ
関西学院大学
hiro.maruhashi@gmail.com
うの　たけあき
国立情報学研究所
受付 18.7.25　採択 18.11.2

そこで本研究では，ターゲット商品の購入・非購入を目的変数として単純にモデルを構築するのではなく，ターゲット商品を購入した正例のサンプルのみを対象にクラスタリングし，負例の各サンプルは，得られたクラスタの中から距離条件を満たすクラスタに組み入れていく．ここに本研究で提案する手法の新規性がある．すなわち，分類モデルを構築する前処理として，サンプルをクラスタリングし，同一のセグメンテーションの正例・負例を対象に分類モデルを構築するのである．

セグメンテーションごとにモデルは構築されることになるが，個々のモデルが対象とするサンプル（顧客）の性質を明確化できるため，モデルの意味解釈が容易になるのである．クラスタリングの手法としては，グラフ研磨 [2] と呼ばれるグラフクリーニングをベースとした方法を用いる．そして，得られた各クラスタの中で，顧客の購買パターンとデモグラフィック属性を説明変数としたロジスティック回帰モデルを構築する．購買パターンとしては顕在パターンを利用し，そのほかに顧客の生まれ年，初回来店の年，担当者の情報，来店に関する情報などを説明変数に用いる．

以降 2 節で関連研究について述べ，3 節では提案手法を説明し，4 節ではヘアサロン利用履歴のデータに対して分類モデルを構築しその結果を示す．そして 5 節でまとめと今後の課題について述べる．

2. 関連研究

マーケティングを対象にした正例・負例の集合を用いた特徴抽出の研究として，Web のアクセスログを対象にした研究 [3] や顧客が店舗を選択する際に重視する購買要因を捉えようとした研究がある [4]．これらはいずれも単一のターゲット変数を対象としており，前者はソフトウェアの申し込みの有無，そして，ある特定サイトの直帰・滞在をそれぞれ目的変数に利用している．また，後者は健康志向の有無を目的変数にしている．

Lavrač et al. [5] は，特定ブランドの認知の有無を対象に Subgroup Discovery による方法を利用したルール抽出を行っている．Subgroup Discovery とは顕在パターン，コントラストパターンとともに Supervised Descriptive Rule Discovery と呼ばれるデータマイニング手法に位置づけられる方法である．また彼らは，別のケースとして特定ブランドの飲用経験の有無に，マーケティングエキスパートの知見を加えた目的変数を作成している．具体的には特定ブランドの飲用者をさらに他ブランドの炭酸飲料飲用者と未飲用者に分けてサンプルを細分化した．ターゲットとなる特定ブランドの飲用経験に別の要因を考慮することでサンプルを細分化している点は，本研究で扱う枠組みと共通しているが，本研究ではエキスパートの知見を取り入れる方法ではなく，クラスタリングを利用し共通する特徴をまとめてクラスタを生成する点が異なる．

また分類モデルでは，モデル木を用いた分類モデルの構築手法が提案されている．MD5 Model Tree [6] は，標準偏差減少 (Standard Deviation Reduction) を利用し結果の均一性が高くなるような分割を行う．そしてルールに基づき分割されたインスタンスを利用して，線形回帰モデルを構築する．われわれの提案手法は，経営的な視点を考慮したセグメンテーションを行うことでデータの分割を行い，その後にモデルを構築することで意味の解釈性を高めようとしている．一方で，MD5 Model Tree はモデルの中でアルゴリズムによるデータの分割とモデル構築を同時に行っているため，説明力のある分類モデルが構築できる可能性はあるが，意味の解釈性については考慮していない．

3. クラスタリングと顕在パターンを利用した分析フレームの提案

本研究で提案する分析フレームを図 1 に示す．まず①では，ターゲットとなる変数を一つ決定する．本研究ではケラスターゼをターゲット変数に選択した．ケラスターゼ (Kerastase) はフランスのロレアルが展開しているヘアサロン向けのヘアケア製品で，付加価値商品として扱われており，施術料金も比較的高額な商品である．ケラスターゼの購買経験のある顧客群を正例，それ以外の顧客を負例と定義した．

次に②では，正例集合に含まれる顧客の購買履歴データを利用して顧客をクラスタリングする．クラスタリングの方法としては，グラフ研磨 [2] と呼ばれるグラフのクリーニング手法を適用し購買アイテムが類似する顧客のクラスタを構成する（詳細は後述）．

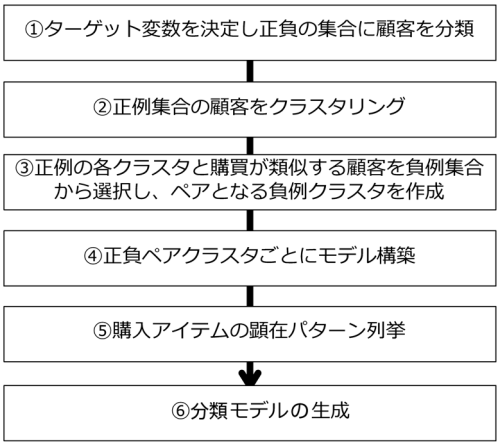

図 1　分析フレーム

③では，各正例クラスタ内で共通するアイテム集合を選択し，それと類似するアイテム集合をもつ顧客を負例集合から選択して負例クラスタを生成する．この方法で生成される負例クラスタは，ターゲット変数であるケラスターゼの購入経験はないが，それ以外の購買アイテムで正例クラスタの顧客群と購買行動が類似しているという点がポイントである．つまり顧客の購買行動の多様性を考慮しながらケラスターゼの購入有無に基づく目的変数を生成することができる．

④，⑤，⑥で購買行動が類似している正負の各クラスタをペアにして分類モデルを構築する．説明変数としては，購入アイテムの違いを表すルールを顕在パターンにより列挙して説明変数を作る．ケラスターゼの購入の有無を除いては，正例・負例ともに購買行動の類似する顧客群を対象にしており，ここから得られる購買パターンを解釈することで，その顧客群に適したケラスターゼの購入に関するルールの抽出が期待できる．④，⑤，⑥の処理をクラスタ数と同じだけ繰り返すことで，クラスタごとに分類モデルを構築する．

以下では②からの手順をより詳細に記述する．

3.1 正例顧客のクラスタリング

正例の顧客群を対象に購買アイテムの類似性に基づき顧客をクラスタリングする．クラスタリングの方法は，各顧客を節点とし，購買行動が互いに類似した節点間に枝を張った一般グラフを構築し（「類似度グラフ」と呼ぶ），そこから密な部分グラフをクラスタとして抽出する．

まず類似度グラフを作成するために，顧客間の類似度を Jaccard 係数で計算する．そして，その値がユーザの設定した最小 Jaccard 係数以上の場合に枝が張られる．ある顧客が購入したことのあるすべてのアイテム

の集合をトランザクションと呼ぶ．そして，任意の二人の顧客のトランザクションを X, Y とすると，Jaccard 係数は，以下のように定義される．

$$\mathrm{Jaccard}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

この値をもとに，最小類似度 δ_0 を閾値として与え，変数ペアを全列挙し，それらの変数間に枝を張ることで類似度グラフを作成する．類似度グラフでは，互いに似た購買行動をもつ顧客群には枝が多く接続されるため枝密度が濃くなり，一方で似ていない顧客群は枝密度が薄くなる．この性質を利用して，ある程度密度の濃い部分グラフをクラスタとして抽出することで，購買行動の似た顧客群を抽出することができる．本研究ではわれわれがこれまでに提案してきたグラフのクラスタリング方法であるグラフ研磨 [2] を適用し，一般グラフのクラスタリングを行う[1]．

グラフ研磨は，グラフをクラスタリングする前に，枝を張り直すことでグラフを再構成し，できる限り構造を明確化する方法である．研磨の方法は，すべての節点ペアについて，その類似度がユーザの指定した閾値以上であれば接続し，そうでなければ接続しないというルールにしたがって，新たなグラフを再構成する．類似度としては式 (1) の Jaccard 係数を用いる．ここではグラフ上での二つの節点を u, v とし，節点 u に直接接続のある節点集合と節点 v に直接接続のある節点集合を対象に $\mathrm{Jaccard}(u,v)$ を計算する．つまり，u と v が互いに共通の節点と隣接しているほど Jaccard 係数が高くなる．そしてユーザが与えた最小類似度 δ_1 以上の類似度をもつ変数ペアに枝を張ることでグラフを再構成する．このようにグラフを再構成すると，共通する隣接節点の多い節点間に枝が張られ，少ない節点間の枝は切断される．

そして新たに構成されたグラフを入力として同様の研磨手法を繰り返し適用し，グラフの構成に変化がなくなるか，もしくはユーザの指定した最大繰り返し回数に達すれば終了する．最終的に得られたグラフが研磨グラフである．この研磨グラフからグラフの連結成分を計算することでクラスタを抽出する．このクラスタを正例クラスタ $p_i (i = 1, 2, \ldots, m$ で，m は連結成分数）と呼ぶ．抽出された正例クラスタは，正例顧客群の中で購買行動が類似する顧客のクラスタである．

[1] さまざまな一般グラフのクラスタリング方法が提案されてきたが，どの手法も一長一短があり筆者らが提案する方法が比較的安定していた [2].

3.2 負例顧客から類似顧客の選択

負例集合から購買行動の類似する顧客を選ぶために，任意の正例クラスタ p_i に属する顧客の多くが購入しているアイテム集合（特徴商品リストと呼ぶ）l_i を計算する．特徴商品リストは，対象となる正例クラスタのアイテム出現割合を計算し，その値がユーザによって与えられた閾値以上の場合にアイテムが選択される．正例クラスタ p_i のトランザクション集合を D_p，また D_p におけるアイテム a の出現集合を $Occ_p(a)$ とすると出現割合は $|Occ_p(a)|/|D_p|$ となる．この値が閾値 δ_2 以上の値をもつ場合に特徴商品リストに加えられる．顧客によっては特徴商品リストに含まれる商品を全く購入していない場合があり，その場合はその顧客を正例クラスタから除外している．

次に特徴商品リスト l_i と負例集合の顧客のトランザクションを比較し，l_i と類似するトランザクションをもつ顧客を負例クラスタ n_i として選択する．類似度の計算はこれまでと同様に Jaccard 係数を利用する．つまり，特徴商品リスト l_i と負例集合の各トランザクションとの Jaccard 係数を計算し，ユーザが設定した閾値 δ_3 以上の値をもつ顧客集合を負例クラスタ n_i として生成する．その際に，同一の顧客が複数の特徴商品リスト l_i の閾値を満たす場合があり，その場合は複数の異なる負例クラスタに同一顧客が割り当てられることになる．

3.3 顕在パターンの列挙

正例クラスタ p_i と負例クラスタ n_i に対して顕在パターンを列挙する．顕在パターンとはユーザの設定した最小支持度および最小増加率以上の相関ルールである．

支持度と増加率の定義は次のとおりである．正例クラスタ p_i，負例クラスタ n_i のトランザクション集合をそれぞれ D_p, D_n で示す．また D_p においてパターン e が出現する部分集合を $Occ_p(e)$ で表す．D_p におけるパターン e の支持度 $support_{D_p}(e)$ は，式 (2) に示されるように，パターン e の出現確率として定義される．

$$support_{D_p}(e) = \frac{|Occ_p(e)|}{|D_p|} \tag{2}$$

また，パターン e の D_n に対する D_p の増加率 $GR_{D_n \to D_p}(e)$ は，式 (3) のとおり定義される．これは，負例でのパターンの出現確率に対する正例での出現確率の比であり，1.0 より大きければ，パターン e は正例に特徴的な出現パターンと言える．同じ方法で負例に特徴的な出現パターンも列挙する．

$$GR_{D_n \to D_p}(e) = \frac{support_{D_p}(e)}{support_{D_n}(e)} \tag{3}$$

以上により得られた顕在パターンとそれ以外のデモグラフィックデータを説明変数として分類モデルを構築する．顕在パターンとしての変数は，列挙された顕在パターンを構成するアイテムのすべてが顧客のトランザクションに含まれている場合に 1 をとる 2 値の変数と定義した．分類モデルにはロジスティック回帰モデルを用いる．分類モデルにおける目的変数を $y \in \{0, 1\}$（0：負例，1：正例），顕在パターンを含む p 個の説明変数ベクトルを $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ とすると，ロジスティック回帰モデルは式 (4) で表される．

$$\Pr(y = 1|\boldsymbol{x}) = f\left(\boldsymbol{\beta}^\top \boldsymbol{x} + \beta_0\right) \tag{4}$$

$f(\cdot)$ はロジスティック関数であり，$f(a) = 1/(1 + \exp(-a))$ で定義される．$\boldsymbol{\beta} \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$ は，それぞれ回帰係数ベクトルと定数項であり，これらは訓練サンプルから推定する．$\boldsymbol{\beta}$ の推定には Lasso と呼ばれる罰則付きの対数尤度を最大化する L_1 罰則付き最尤推定法 [7] を交差検証法によって実施し，モデルを構築した．

4. 計算実験

本研究で利用するデータは，首都圏のあるヘアサロンチェーンの 2015 年 7 月 1 日〜2017 年 6 月 30 日の店舗 POS データである．本研究ではケラスターゼのプロモーション計画を立案することを前提に，ある 1 店舗における女性客 2,665 人（うち，ケラスターゼ購入客 509 人，非購入客 2,156 人）のデータに対して提案手法を適用し分析を行った．ケラスターゼはケミカルトリートメントの施術の中で高級ランクのものであり，単価が高く効率的なプロモーション効果が期待できる商品である．

まず，ケラスターゼ購入客 509 人を対象に 3.1 節の内容に従い顧客類似度グラフを作成した．ここでアイテムは各商品名に対応し，各商品に対する期間内の購入の有無により，トランザクションを構成した．

類似度は Jaccard を用い，最小類似度 $\delta_0 = 0.4$ とした．δ_0 はあまりにも小さいと似ていない顧客に枝を張ることになり，一方でこの値が高すぎると，顧客間に枝が張られにくくなり疎なグラフになる．また次のステップのグラフ研磨の適用を考慮すると，比較的頑健な構造を得るために，類似度グラフは，ある程度多くの枝を残しつつクラフ研磨の枝の追加と削除によって構造を安定化させるほうがよい．そこで，今回は δ_0 を 0.1 刻みで 0.3〜0.8 まで変更しながら同様の分析を実施して，枝密度の濃淡を観察しながら恣意的に $\delta_0 = 0.4$ を選択した．

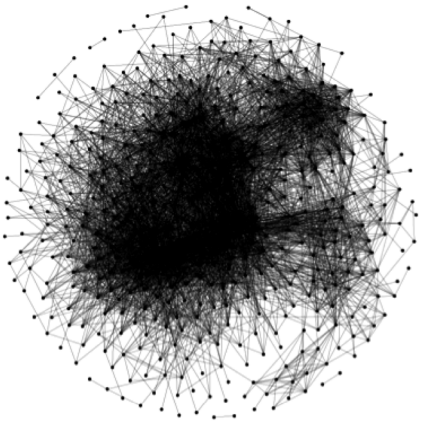

図 2　類似度グラフ
gephi (Fruchterman Reingold) にて描画

表 1　δ_1 を変化させたときの研磨後のグラフ

δ_1	枝の削除	共通枝数	枝の追加	連結成分数
0.2	40,376	12,374	714	15
0.3	49,880	2,870	9	13
0.4	51,721	1,029	0	10
0.5	52,422	328	0	9
0.6	52,507	243	0	4
0.7	52,729	21	0	1

$\delta_0 = 0.4$ の類似度グラフに対するグラフ研磨の結果

この顧客類似度グラフを可視化ツールで描画したものを図 2 に示す．図の節点が顧客 1 名を表し，枝は顧客間の類似度が閾値以上のものを表しており，枝の数は合計で 52,750 本であり，全体的に密度が濃い部分が多くあることが確認できる．

次に，この顧客類似度グラフに対してグラフ研磨を実行することで顧客クラスタを得る．表 1 は，$\delta_0 = 0.4$ の類似度グラフに対して δ_1 を 0.2 から 0.7 まで変えたときに得られた研磨後のグラフの特徴量を示している．δ_1 の値が大きくなるに従い，共通する隣接節点数に関する条件が厳しくなるため枝は削除されやすくなり，また追加される枝の数は減る傾向にある．実際に表の値を見ると，δ_1 の値が大きくなると削除される枝の数は増えている．また，追加された枝の数は減っており，$\delta_1 = 0.4$ 以降ではなくなっている．共通枝数は，もとの類似度グラフと共通する枝の数であり，枝の追加と削除をすることによって共通する枝数は減っておりグラフの構造が徐々に変わっていっていることが確認できる．連結成分数は，研磨後のグラフに対して求

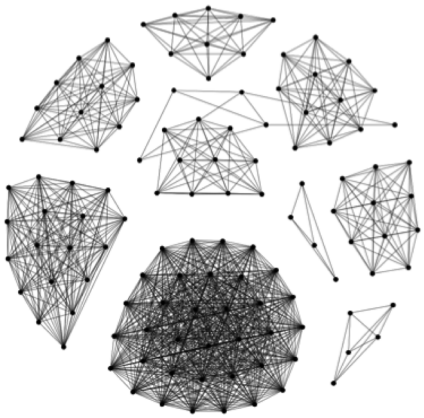

図 3　研磨後グラフ
gephi (Fruchterman Reingold) にて描画

めたものであり，δ_1 の値が増えるにしたがい，枝が削除されて単一の節点が多くなることで連結成分が小さくなり，連結成分数が減っていくため，最終的に一つのクラスタになっている．

この表の値からある程度もとのグラフ構造を残しつつ，分析可能なクラスタ数の観点から本研究では $\delta_1 = 0.4$ のグラフ研磨を採用した．δ_1 が 0.2 や 0.3 の場合は，グラフ研磨の特徴である枝の追加が行われているが，連結成分数が増えるため，一つのクラスタあたりの顧客人数が減ることになる．表 2 は $\delta = 0.4$ の場合に得られた各クラスタ ($i = 1 \sim 10$) に属する人数などを示しており，購入客数・非購入客数を確認すると $\delta_1 = 0.4$ の場合でも小規模なクラスタがあるため，さらにクラスタ数が増えるようなパラメータを採用することはモデル構築の観点からは難しい．研磨後のグラフを図 3 に示す．図 2 と比べると，研磨後のグラフは密な部分がそれぞれ明確なクラスタになっており，ノイズが除去されて密な構造が浮かび上がっていることが確認できる．

最終的に得られた顧客クラスタから 3.2 節の内容に基づき出現割合の閾値を $\delta_2 = 0.4$ で特徴商品リストを抽出した．そして，負例集合であるケラスターゼ非購入客から類似度の閾値 $\delta_3 = 0.4$ として特徴商品リストに共通するアイテムをトランザクションに含む顧客を負例クラスタとして生成した．表 2 は，各クラスタ p_i と n_i に共通する特徴商品リストを示している．それぞれの特徴商品リストの解釈は 4.1 節で示す．

最後に，これら 10 個の顧客クラスタに対して，それぞれクラスタ別に購入商品の顕在パターンを列挙し，顧客属性を説明変数に追加してロジスティック回帰を行った．その際の目的変数は，各クラスタの正例 p_i と

表 2　クラスタ別の特徴商品リストとモデル精度

i	購入客数	非購入客数	特徴商品リスト	精度
1	35	92	エグゼクティブ・カット	0.795
2	11	254	ディレクター・カット，カラー，PB・トリートメント	0.953
3	11	27	追加シャンプー・ブロー	0.692
4	5	10	追加シャンプー・ブロー，グロス	0.600
5	18	106	ディレクター・カット	0.895
6	5	145	スタイリスト・カット	不可
7	9	24	追加シャンプー・ブロー	0.818
8	11	59	追加シャンプー・ブロー，ディレクター・カット，カラー，シャンプー・ブロー，PB・トリートメント，前髪カット，部分カラー	0.864
9	4	76	シャンプー・ブロー	0.920
10	12	62	追加シャンプー・ブロー，ディレクター・カット，PB・トリートメント，シャンプー・ブロー，カラー，部分カラー，前髪カット，リラクゼーションシャンプー, PB2・コース 1	0.826

表 3　ロジスティック回帰の説明変数

購入商品の顕在パターン
顧客属性（初回来店主担当者 ID）
顧客属性（主担当者 ID）
顧客属性（会計指名区分）
顧客属性（DM 送信可否）
顧客属性（累積来店回数）
顧客属性（初回来店年）
顧客属性（誕生年代）
顧客属性（現金払）
顧客属性（半年以内の来店有無）
顧客属性（来店確率 = 来店回数/期間日数）
顧客属性（平均単価）
顧客属性（担当スタイリスト数）

負例 n_i であり，顕在パターンのパラメータは最小支持度が 0.1，最小増加率を 1.1 で列挙した．また，ロジスティック回帰は Lasso を用い，表 3 に示す変数を説明変数として利用した．顧客の嗜好性を少しでも深く浮き彫りにするため，商品購入履歴のほかに提供データからわかる限りの顧客属性を説明変数に追加している．また，モデルの精度は表 2 に示している．これはホールド・アウト法で訓練データを 6 割，残りをテストデータとしてモデルの精度を検証し，正しく予測できた割合を示している．$i = 6$ は正・負例の人数の偏りが大きく，ホールド・アウト法ではモデルの構築ができなかったため不可と記述した．

4.1　分類モデルの意味解釈

抽出された顧客クラスタ（表 2）について，$i = 1, 5, 6$ のクラスタはカットを担当する主担当スタイリストの

表 4　顧客クラスタ $i=1$ の分析結果

(a) 正例：ケラスターゼ購入客

回帰係数	説明変数（特徴量）	正例客数	負例客数
2.7398	主担当者 ID53	1	0
1.8636	主担当者 ID5	1	0
1.384	主担当者 ID34	1	0
1.8281	エグゼクティブ・カット，カール・パーマ	9	0
0.0003	平均単価	8,023	12,713

「平均単価」の正例客数・負例客数は正例客・負例客における「平均単価」の平均値を表す．

(b) 負例：ケラスターゼ非購入客

回帰係数	説明変数（特徴量）	正例客数	負例客数
0.4596	主担当者 ID150	2	30

ランクの違いによってクラスタが形成されていると解釈できる．また，$i=8,10$ のクラスタは前髪カットと部分カラーという最近の若者に定番の商品構成をもっており，両者の違いはシャンプーなどの付加価値商品の差である．$i=3,4,7,9$ のクラスタはシャンプー（およびその後のヘアセット）を主体としたクラスタであり，対象店舗の立地特有の客層が想定される．以下では $i=1,8$ のクラスタにおける分析結果の解釈例を示す．

$i=1$ のクラスタ，つまりエグゼクティブのスタイリストを嗜好するクラスタの分析結果を示したものが表 4 である．主担当者 ID53, 5, 34 を説明変数にもつ正例の顧客数は 1 件，負例の顧客数は 0 件であった．このような変数が選択されている理由は，サンプル数が少ないため変数選択の正則化ペナルティより最小二乗誤差を減少させる効果のほうが高くなっているからである．これらの説明変数は解釈の対象から除外する．

そこで，正例（購入客）の顕在パターンに現れている「カール・パーマ」の商品と負例（非購入客）の変数である主担当 ID150 に注目すると以下のような解釈ができる．一般的にケラスターゼなどのケミカルトリートメントはカラーやパーマなどの施術で傷んだ髪質を化学物質でコーティングしなおすもので，カラーやパーマとセットで施術するとその効果が期待できるものである．しかしながら，パーマのできばえはスタイリストの技術力に大きく依存するため，技術力の低いスタイリストがパーマとセットでケラスターゼを施術した場合，顧客はパーマのできばえの悪さに引きずられてケラスターゼの効果を過少に評価する傾向がある．そのため，その顧客がケラスターゼを再度購入しない，もしくはそのスタリストが同じ顧客にケラスターゼを再度提案しない，ということが起こりやすい．このクラスタの分析結果はまさにその状態を表していると考えられる．すなわち，ケラスターゼを購入している顧客の担当スタイリストのパーマの技術力がかなり高く，一方で主担当 ID150 のスタイリストのパーマの技術力がその域に達していない，というものである．

本提案手法を用いることでこのような具体的な仮説を考えることが可能となる．この仮説は現場および当事者のスタイリスト間で検証するしかないが，この仮説が正しければ，主担当 ID150 のスタイリストに対し

表 5　顧客クラスタ $i=8$ の分析結果

(a) 正例：ケラスターゼ購入客

回帰係数	説明変数（特徴量）	正例客数	負例客数
5.2919	前髪カット，ホイルワーク (1-5)・カラー，部分カラー，カラー，PB・トリートメント，シャンプー・ブロー	3	0
3.1717	ディレクター・カット，前髪カット，カラー，PB・トリートメント，追加シャンプー・ブロー	8	6
2.0698	ディレクター・カット，前髪カット，カラー，PB・トリートメント，追加シャンプー・ブロー，シャンプー・ブロー	6	2
0.5615	ディレクター・カット，前髪カット，カラー，PB・トリートメント	9	7
1.4901	会計主担当者 ID33	11	17
0.3064	初回来店担当者 ID2	8	12

(b) 負例：ケラスターゼ非購入客

回帰係数	説明変数（特徴量）	正例客数	負例客数
1.8166	部分パーマ，部分カラー，PB・トリートメント，追加シャンプー・ブロー，シャンプー・ブロー	0	6
0.9985	部分パーマ，前髪カット，カラー，PB・トリートメント，追加シャンプー・ブロー	0	7
0.5261	部分パーマ，前髪カット，PB・トリートメント，追加シャンプー・ブロー	0	10
1.1736	初回来店担当者 ID5	2	43
0.5593	担当スタイリスト数	1	1.29
0.0226	DM 送信可否拒否	1	37

「担当スタイリスト数」の正例客数・負例客数は正例客・負例客における「担当スタイリスト数」の平均値を表す．

表 6　ケラスターゼの購入経験の有無のみの顕在パターンの結果

係数	顕在パターン	提案手法のクラスタとの共通性
0.42	エグゼクティブ・カット	$i=1$ の特徴商品リスト
0.86	エグゼクティブ・カット，カール・パーマ	$i=1$ 内の正例
0.03	カラー，PB・トリートメント，カール・パーマ	$i=2$ 内の負例
0.29	ディレクター・カット，カラー，PB・トリートメント	$i=2$ の特徴商品リスト
0.41	ディレクター・カット，PB・トリートメント	$i=2$ の特徴商品リスト
1.44	追加シャンプー・ブロー	$i=3,7$ の特徴商品リスト
0.54	ディレクター・カット	$i=5$ の特徴商品リスト
0.24	スタイリスト・カット	$i=6$ の特徴商品リスト
−0.88	追加シャンプー・ブロー，PB・トリートメント	$i=8$ の特徴商品リスト
0.16	追加シャンプー・ブロー，カラー，PB・トリートメント	$i=8$ の特徴商品リスト
−0.32	エグゼクティブ・カット，PB・トリートメント	$i=10$ 内の負例
−0.09	リラクゼーションシャンプー，カラー，PB・トリートメント	$i=10$ の特徴商品リスト
0.89	ルネ・コース 1	$i=10$ の特徴商品リスト
1.36	前髪カット，PB・トリートメント	$i=10$ の特徴商品リスト
−0.22	エグゼクティブ 080・カット	なし
−0.17	PB・トリートメント，カール・パーマ	なし
0.07	DS アドジュネス	なし
0.07	DS アドジュネス，PB・トリートメント	なし
0.12	カラー，カール・パーマ	なし

クラスタリングを行わずケラスターゼの購入非購入を目的に顕在パターンを列挙した結果と提案手法のクラスタがもつ特徴商品リストの共通性を比較している．顕在パターンの多くが特徴商品リストと共通していることから，従来の方法ではクラスタを代表する購買パターンレベルのものしか抽出されていないことがわかる．

て技術力向上の訓練を行い（それが触媒となり），対象顧客に対するケラスターゼの提案とその受注可能性を上げることができる．

次に $i=8$ のクラスタの分析結果を表 5 に示す．表 5 の結果を見ると，負例クラスタの特徴（表 5(b)）としては，「部分パーマ」「部分カラー」など髪の毛の一部にアクセントを出す施術が行われており，これらは若者を中心としたクラスタであることが考えられる．

また，特に購入客の購入商品には現れていない「部分パーマ」（表 5(b) 下線部）に注目すると，ケラスターゼの購入を促すためには，部分パーマの扱いがポイントとなる．部分パーマは非購入客の特徴的な施術であり，ケラスターゼ導入の阻害要因になっていることが考えられる．これは，スタイリストが，部分パーマとセットでケラスターゼを施術することでパーマの技術力とケラスターゼの効果を顧客に過小評価されるリスクを意識したり，このクラスタが若者を中心としていることから顧客のコスト負担を意識したりすることで，顧客にケラスターゼを適切に勧めていないのではないかという仮説が考えられる．そこで，「今回の施術では部分パーマをせずにケラスターゼを試してみないか」と誘導する，という施策が考えられる．一時的に目先の「部分パーマ」の注文を失うが，ケラスターゼは一度その商品のファンになると再度購入する可能性が一般に高いことから，以降の商品提案の幅が拡がり，結果として今まで以上の売上貢献が見込める可能性が高いと考えられる．

4.2　クラスタリング効果の検証

最後に，提案手法の核心であるクラスタリングの効果について検証しておく．比較対象として，クラスタリングは行わずにケラスターゼの購入，非購入を目的変数としたデータから顕在パターンの列挙と Lasso によるモデルの構築を行った．表 6 はその結果を示している．ただしここでは特徴商品リストと比較するために説明変数には顕在パターンだけを利用している．

表 6 の「提案手法のクラスタとの共通性」は，提案手法で求めたクラスタのもつ特徴商品リストや顕在パターンと，クラスタリングなしの顕在パターンとの共通性を示したものである．たとえば「$i=1$ の特徴商品リスト」と記載されている場合は，この顕在パターンは表 2 のクラスタの $i=1$ の特徴商品リストと一致していることを示している．また，「$i=1$ 内の正例」や「$i=1$ 内の負例」は，クラスタの顕在パターンと一致していることを表しており，「$i=1$ 内の正例」の場合はクラスタ $i=1$ の正例に顕著な顕在パターンと一致している．また，なしはクラスタの特徴商品リストや顕在パターンには一致しないものを表している．

この結果から，クラスタリングなしに得られた顕在

パターンは提案手法により求められたクラスタの特徴商品リストとほぼ同程度のパターンしか列挙できておらず，クラスタを代表するような購買パターンレベルのものしか抽出できていないことがわかる．本研究では，これらの購買行動を顧客嗜好の無視できない要素つまり多様性として考慮したうえでクラスタを構築しており，その中でどのような購買特徴の違いが正例と負例で現れているのかという点を明らかにしている．

5. おわりに

本研究では，目的変数を定義する際にクラスタリングを行うことで顧客の多様性を考慮した分類モデルが構築できることを示した．特にサンプル全体をクラスタリングするのではなく，正例のサンプルのみを対象にクラスタリングし，負例の各サンプルは，得られたクラスタの中から距離条件を満たすクラスタに組み入れていく方法を提案した．この方法によって正例集合の特徴を反映したノイズの少ないクラスタを構築することができ，意味解釈が可能なより明確な特徴を捉えることができた．

またグラフのクラスタリングにおいて類似度グラフにグラフ研磨手法を適用することで，顧客の購買行動が類似している密なグラフ構造を浮かび上がらせることができた．しかし，グラフ研磨のパラメータについては，実質科学的な観点から試行錯誤による方法だけではなく，たとえば分類モデルの精度が向上するようなパラメータを設定するなど最適化の観点からのアプローチなども必要であり，この点は今後の課題としたい．

謝辞　本研究は，JST CREST（グラント番号：JPMJCR1401）および JSPS 科研費：15K17146 の研究助成を受けている．

参考文献

[1] G. Dong, X. Zhang, L. Wong and J. Li, “CAEP: Classification by aggregating emerging patterns,” *Discovery Science*, S. Arikawa, K. Furukawa (eds.), Springer, pp. 30–42, 1999.

[2] T. Uno, H. Maegawa, T. Nakahara, Y. Hamuro, R. Yoshinaka and M. Tatsuta, “Micro-clustering by data polishing,” In *Proeedings of 2017 IEEE International Conference on Big Data*, pp. 1012–1018, 2017, DOI: 10.1109/BigData.2017.8258024.

[3] 羽室行信，中西正雄，山本昭二，“統合化顕在パターン判別モデルによる Web アクセスログデータの分析，”オペレーションズ・リサーチ：経営の科学，**53**(2), pp. 75–84, 2008.

[4] 中原孝信，羽室行信，宇野毅明，“グラフ研磨手法を用いた顧客の店舗選択モデルの構築，”オペレーションズ・リサーチ：経営の科学，**60**(2), pp. 89–95, 2015.

[5] N. Lavrač, B. Cestnik, D. Gamberger and P. A. Flach, “Decision support through subgroup discovery: Three case studies and the lessons learned,” *Machine Learning Special Issue on Data Mining Lessons Learned*, **57**, pp. 115–143, 2004.

[6] J. R. Quinlan, “Learning with continuous classes,” In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, pp. 343–348, 1992.

[7] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, **58**, pp. 267–288, 1996.