論文・事例研究

セール品に注目した顧客の購買行動の解析
—2値データのクラスタリングを考慮したロジスティック回帰分析—

山下　遥，鈴木　秀男

1. はじめに

近年，インターネットショッピングは，その市場規模を増やし続けており，2012 年には 9.5 兆円の規模となっている [1]．その中でも特に，衣料・アクセサリー小売業は国内外において多くの顧客を有し [1]，その動向が大きく注目されている．今回の経営科学系研究部会連合協議会が主催した平成 25 年度データ解析コンペティションは，こうしたファッション EC サイトにおける顧客の購買データを解析するものであった．インターネットショッピングのビジネスにとって，「どのように新規顧客を獲得するのか」，さらには「どのように新規顧客の再購買を促すのか」という問題は，重要な課題であり，バーゲン（割引）が幅広く取り入れられている [2]．今回，データを提供していただいたファッション EC サイトにおいても，2011 年 9 月から 2013 年 3 月までの間に会員登録した，49,814 人の新規顧客のうち，15,593 人，すなわち，3 割強の顧客が初期購買においてすべての商品をバーゲン価格で購買している．バーゲン販売は，新規顧客獲得に有効な方法となっていることがわかる．

しかしながら，バーゲン品のみを購買した顧客は必ずしもリピーター（i.e., 初期購買の後に購買を行う顧客）となるわけではなく，その後の行動により，企業にとっての利益が大きく異なってくるものと思われる．そこで，バーゲン品のみを初期購買した顧客に関し，その後の購買パターンに従い，3 種類の顧客を定義していくことにする．

1 つ目の行動パターンは，バーゲン品のみを一度だけ購買し，その後は購買行動を行わないというパターンである．本研究では，この行動パターンをもつ顧客を「初期退出者」として定義することにする．さらに，初期購買の後に購買行動を行う顧客（リピーター）であっても，その後の購買行動パターンは，バーゲン品のみを買い続けるパターンと，バーゲン品以外の商品も購買するパターンに分けられる．本研究では，バーゲン品のみを買い続ける顧客を「バーゲンハンター」，バーゲン品以外の商品も購買する顧客を「バーゲン優良顧客」と定義する．実際にバーゲン品のみを初期購買した顧客をこの 3 つのパターンへと分類すると，それらの購買金額は表 1 のようにまとめられ，顧客の種類によって明らかな違いがあることがわかる．

表 1　顧客の種類，人数，および平均購買金額

顧客の種類	人数	平均購買金額
初期退出者	5,899 人	6296.8 円
バーゲンハンター	5,233 人	16227.8 円
バーゲン優良顧客	4,461 人	50006.6 円

本研究では，バーゲン品のみを初期購買した顧客が，リピーターになるか初期退出者になるかにはどのような要因があるのか，さらに，リピーターになった顧客が，バーゲン優良顧客になるか，バーゲンハンターになるかにはどのような要因が存在しているのかを，顧客の初期の購買行動データおよび顧客の属性データを用いて分析していくことにする．これにより，顧客が初期購買をした時点で，どの種類の顧客になるかを予想することができ，ビジネスへの活用が期待される．

リピーターになるか初期退出者になるかの分析や，バーゲン優良顧客になるかバーゲンハンターになるかの分析のように，2 値で表されるような反応の要因を説明変数を用いて分析する方法としてロジスティック回帰分析が幅広く適用されている．ここで，本研究で

やました　はるか，すずき　ひでお
慶應義塾大学
〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
受付 14.7.25　採択 14.11.10

扱う変数のうち，「初期購買品目」は，6 変数のカテゴリカルデータとなっており，「どの品目を購買しているか」を 2 値のベクトルで表すことができる．これに対して，2 値のベクトルを説明変数に入れてロジスティック回帰式を求めることができるが，変数の交互作用をどこまで考慮するべきか（例えば，2 因子交互作用まで，3 因子交互作用までなど）という問題が存在し，交互作用を取り入れるほどパラメータの数が大きく増加してしまい，予測モデルとして適用しようとした場合，汎化能力が低下してしまうことが懸念される．

本研究では，Yamashita and Suzuki [3] の 2 値型 principal points をロジスティック回帰分析に応用して，2 値型の代表点を求め，2 値型代表点をもとにデータをクラスタリングしながらデータに対する尤度が最大になるようなロジスティック回帰式を求める方法を提案する．この方法は，パラメータの数を少なく抑えながら，データに対する当てはまりがよくなるような 2 値型代表点およびクラスタと，それらを用いたロジスティック回帰式を同定することを可能とする．さらに本研究では，バーゲン品のみを初期購買した顧客の初期購買データおよび属性データから，バーゲン品のみを初期購買した顧客が初期退出者になるのかリピーターになるのか，また，リピーターになった顧客がバーゲンハンターになるのかバーゲン優良顧客になるのかの予測モデルを，提案モデルを用いて構築し，得られた予測モデルから初期購買がバーゲン品のみの顧客をリピーターにするためのアプローチ，およびバーゲン優良顧客とするためのアプローチについて考察する．

2. 本研究の基礎となる 2 値型代表点とロジスティック回帰分析

2.1 ロジスティック回帰分析

本研究におけるリピーターになるか初期退出者になるかの判別，バーゲン優良顧客になるかバーゲンハンターになるかの判別のように反応が 2 値となるデータは数多く存在する．この反応に対していくつかの要因を説明変数を用いて反応との関係を明らかにしようとする場合にロジスティック回帰分析が幅広く用いられている．本節では，3 節にて 2 値データのクラスタリングを考慮したロジスティック回帰分析を提案するために，本研究の基礎となるロジスティック回帰分析について概説する．

まず N 個の $l+m+1$ 変量説明変数ベクトル $\boldsymbol{c}_i \in \mathbb{R}^{l+m+1}$ $(i=1,...,N)$ を，i 番目の l 変量 2 値型データベクトル $\boldsymbol{a}_i \in \{0,1\}^l$ と，それ以外の m 変量データベクトル $\boldsymbol{b}_i \in \mathbb{R}^m$，および，切片に対応する要素 1 を $\boldsymbol{c}_i^T = (\boldsymbol{a}_i^T, \boldsymbol{b}_i^T, 1)$ のように並べたベクトルとして定義する．ただし，$\{0,1\}^l$ は，$\{0,1\}^l = \{0,1\}\times\{0,1\}\times\cdots\times\{0,1\}$ で表される直積空間とする．ここで，2 値の反応を表す確率変数を $Z=\{0,1\}$ とおき，$Z=1$ となる確率を p，$Z=0$ となる確率を $1-p$ とおく．このとき，反応を起こす説明変数ベクトルと p との関係を，パラメータベクトル $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_a{}^T, \boldsymbol{\beta}_b{}^T, \beta_0)$（$\boldsymbol{\beta}_a$: 2 値型の説明変数ベクトルに対応するパラメータベクトル，$\boldsymbol{\beta}_b$: それ以外の説明変数ベクトルに対応するパラメータベクトル，β_0: 切片を表すパラメータ）を用いてロジスティック関数に当てはめると，

$$p_i = \frac{\exp(\mathbf{c}_i{}^T\boldsymbol{\beta})}{1+\exp(\mathbf{c}_i{}^T\boldsymbol{\beta})} \tag{1}$$

で表すことができる．

一方，2 値の反応を表す確率変数 Z は，2 項分布に従うものと考えられるため，N 個の反応 $z_i=\{0,1\}$ $(i=1,\ldots,N)$ が得られた場合，その対数尤度関数 $\log\ L(p_1,\ldots,p_N \mid z_1,\ldots,z_N)$ は，

$$\begin{aligned}&\log\ L(p_1,\ldots,p_N \mid z_1,\ldots,z_N)\\ &=\sum_{i=1}^{N} z_i\ \log\frac{p_i}{1-p_i}+\sum_{i=1}^{N}\log(1-p_i)\end{aligned} \tag{2}$$

となる．(1) 式を (2) 式に代入すると，以下の式が導かれる．

$$\begin{aligned}&\log\ L(\boldsymbol{\beta})\\ &=\sum_{i=1}^{N} z_i\ \mathbf{c}_i{}^T\boldsymbol{\beta}-\sum_{i=1}^{N}\log\left[1+\exp(\mathbf{c}_i{}^T\boldsymbol{\beta})\right]\end{aligned} \tag{3}$$

この対数尤度を最大化するようなパラメータベクトル $\boldsymbol{\beta}$ を求めることで，データへの尤度が最も大きくなるロジスティック関数を同定することができる．

2.2 2 値型代表点

本研究で扱う説明変数の中には，「どの商品群を買ったのか」の 2 値のデータが含まれており，さらに，買った商品群のパターンは，いくつかのグループに分けられる（クラスタリングが可能である）ものと思われる．2 値の確率分布の解析方法として，主に統計の分野においてその理論的な性質に関する研究 [4〜6] や，応用研究 [7, 8] が展開されている principal points [4] を基礎とした 2 値型 principal points [3] が提案されている．これは，確率変数 $\boldsymbol{X}$ が多変量 2 値ベクトルの場合，すなわち，$\boldsymbol{X}$ が 2 値の l 次元空間上 $\{0,1\}^l$ のみに確率をもつ確率分布 F に従う場合において，取り

うる値を多変量2値分布上の k 個の点 $\boldsymbol{\xi}_j \in \{0,1\}^l$ に限定した2値型の principal points として位置づけられ，以下のように定式化される．

まず，k 個の l 次元 2 値ベクトルを $\boldsymbol{y}_j \in \{0,1\}^l$ $(j=1,...,k)$ とし，確率変数 $\boldsymbol{X}$ の実現値ベクトル $\boldsymbol{x}_i \in \{0,1\}^l$ と $\boldsymbol{y}_j$ との距離の二乗 d^2 は以下のように表されるものとする．

$$d^2(\boldsymbol{x}_i \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) = \min_{1 \le j \le k} (\boldsymbol{x}_i - \boldsymbol{y}_j)^T (\boldsymbol{x}_i - \boldsymbol{y}_j)$$

このとき，多変量 2 値分布 F に従う確率変数 $\boldsymbol{X}$ の k-principal points は，下式を最小化する k 個のベクトル $\boldsymbol{\xi}_j \in \{0,1\}^l$ として与えられる．

$$\begin{aligned} & E_F[d^2(\boldsymbol{X} \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)] \\ &= \min_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \in \{0,1\}^l} E_F[d^2(\boldsymbol{X} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k)] \\ &= \min_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \in \{0,1\}^l} \sum_{i=1}^{2^p} P[\boldsymbol{X}=\boldsymbol{x}_i] d^2(\boldsymbol{x}_i \mid \boldsymbol{y}_1, ..., \boldsymbol{y}_k) \quad (4) \end{aligned}$$

本研究では，確率変数 $\boldsymbol{X}$ は，N 個のサンプルに基づく経験分布 $\hat{F}$ に従うものとする．さらに，確率変数 $\boldsymbol{X}$ の実現値ベクトル（サンプル）を $\boldsymbol{x}_i$ $(i = 1, ..., N)$ とすることで，(4) 式は，

$$\begin{aligned} & E_{\hat{F}}[d^2(\boldsymbol{X} \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)] \\ &= \min_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \in \{0,1\}^l} \sum_{i=1}^{N} d^2(\boldsymbol{x}_i \mid \boldsymbol{y}_1, ..., \boldsymbol{y}_k)/N \quad (5) \end{aligned}$$

と表される．本研究では，このサンプルに基づいた経験分布における principal points を 2 値型代表点と呼び，この考え方をロジスティック回帰分析に応用していくことにする．

2.3 Firth 法を用いたロジスティック回帰モデル

2 値型代表点をロジスティック回帰分析に応用する（詳しくは 3 章に記述していくことにする）際に，「すべての反応が同じクラスタ」が存在する準完全分離 [9] の状態となることが考えられる．このような場合，尤度を最大化する際にパラメータの推定値が発散してしまい，最尤推定量を求めることができない [9]. この問題への対策として，Firth 法 [10] を用いて最尤推定量のバイアスを除くことでパラメータを推定する方法が存在する [9]. これは，(3) 式で表す対数尤度関数に対して修正対数尤度関数

$$\log\ FL(\boldsymbol{\beta}) = \log\ L(\boldsymbol{\beta}) + \frac{1}{2} \log\ |\, I(\boldsymbol{\beta})\,| \quad (6)$$

を用いることで，最尤推定量を求めることを可能とする．ただし，$|\, I(\boldsymbol{\beta})\,|$ は，Fisher の情報行列の行列式とする．

3. 2 値データのクラスタリングを考慮したロジスティック回帰モデル

3.1 2 値データのクラスタリングを考慮したロジスティック回帰モデルの定式化

本研究の目的は，顧客の初期購買の情報から，リピーターになるか，初期退出者になるか，また，バーゲン優良顧客になるか，バーゲンハンターになるのか，といった 2 値の応答と，説明変数との関係を明らかにすること，およびその予測式を同定することである．また，説明変数の中には「どの商品群を買ったのか」の 2 値データが含まれている．このような場合，2 値データをダミー変数として 2.1 節のようにロジスティック回帰分析によって回帰式を同定することができる．しかしながら，このような分析を行う際には，交互作用をどこまで考慮すべきか（例えば，2 因子間交互作用まで，3 因子間交互作用まで）といった問題が存在する．これに対して，買った商品群のパターンをいくつかのグループに分け，それを新しいダミー変数を作成することで，ダミー変数同士の関係を考慮するといったアプローチが有効であるものと思われる．

これに対してまず，前節において述べた 2 値データを 2 値型代表点を用いて k 個のクラスタへと分割し，それを「どのクラスタに属するのか」を表す変数へとダミー変数化して偏回帰係数を推定するモデルを構築することができる．しかしながら，2 値型代表点を用いた k 個のクラスタへの分割は，「データとの距離の二乗和が最小になるような点を用いたクラスタリング」であり，それが必ずしもロジスティック回帰式とデータとの当てはまりを改善させるとは限らない．そこで，多変量 2 値データにおけるクラスタリングの基準を，「求められる Firth 法を用いたロジスティック回帰式とデータとの当てはまりの最大化（実績値と予測値の残差二乗和の最小化）」へと変更するアプローチを考えていくことにする．

まず，2.1 節のように表される説明変数ベクトル $\boldsymbol{c}_i^T = (\boldsymbol{a}_i^T, \boldsymbol{b}_i^T, 1) \in \mathbb{R}^{l+m+1}$ を，$\boldsymbol{c}_i'^T = (\boldsymbol{d}_i^T, \boldsymbol{b}_i^T, 1) \in \mathbb{R}^{k+m}$ へと変換する．ただし，$\boldsymbol{c}_i'$ は，$\boldsymbol{a}_i$ と k 個の代表点の候補 $\boldsymbol{y}_j$ によって (10) 式のように決定されたクラスタを表す 2 値のダミー変数のうちの 1 つを除外し（i.e., ランク落ちの問題のため，効果を 0 に固定し）ベクトルで表した $\boldsymbol{d}_i = (d_{ig})$ $(g = 1, \ldots k-1)$，連続値ベクトル $\boldsymbol{b}_i \in \mathbb{R}^m$，さらに Firth 法を用いたロジスティック回帰モデルにおける切片に対応する要素 1 を加えたベクトル

とする．また，Firth 法を用いたロジスティック回帰モデルの切片の値を $\hat{\beta}_0$ とし，ロジスティック回帰モデルのパラメータベクトルを $\hat{\boldsymbol{\beta}}_i'^T = (\hat{\boldsymbol{\beta}}_d{}^T, \hat{\boldsymbol{\beta}}_b{}^T, \hat{\beta}_0) \in \mathbb{R}^{k+m}$，とする．このとき，2 値データのクラスタリングを考慮したロジスティック回帰モデルを，(3) 式，(5) 式，(6) 式を用いて以下のような最適化問題として定式化する．

$$\max_{\hat{\boldsymbol{y}}_j,\ 1\le j\le k}\left\{\max_{\hat{\boldsymbol{\beta}}'}\left(\log\ L(\hat{\boldsymbol{\beta}}')+\frac{1}{2}\log\ \left|\ I(\hat{\boldsymbol{\beta}}')\ \right|\right)\right\}, \quad (7)$$

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\boldsymbol{c}}_i'^T\hat{\boldsymbol{\beta}}', \quad (8)$$

$$\boldsymbol{c}_i'^T = \{\boldsymbol{d}_i{}^T, \boldsymbol{b}_i{}^T, 1\}, \qquad \hat{\boldsymbol{\beta}}'^T = \{\hat{\boldsymbol{\beta}}_d{}^T, \hat{\boldsymbol{\beta}}_b{}^T\hat{\beta}_0\}, \quad (9)$$

$$d_{ig} = \begin{cases} 1 & (\text{if } \min_{1\le j\le k}(\boldsymbol{a}_i-\boldsymbol{y}_j)^T(\boldsymbol{a}_i-\boldsymbol{y}_j) \\ & \quad = (\boldsymbol{a}_i-\boldsymbol{y}_g)^T(\boldsymbol{a}_i-\boldsymbol{y}_g)) \\ 0 & (\text{if 上記以外}) \end{cases} \quad (10)$$

ただし，(10) 式が一意に定まらない場合は，データベクトルに対して最も近い複数の代表点の中からランダムに代表点を決定する．(7) 式で表される最適化問題は，(10) 式によって決定する「それぞれのデータがどのクラスタに属するのか」という変数を用いた Firth 法を用いたロジスティック回帰モデルの対数尤度が最大になるように k 個の代表点 $\hat{\boldsymbol{y}}_j$ $(j = 1, \ldots, k)$ を求める問題として位置づけることができる．また，(7) 式の中括弧の中の最大化問題は，(6) 式で表される Firth 法を用いたロジスティック回帰モデルの尤度の最大化に相当する．

この最大化問題 (7) 式は，取りうる $\boldsymbol{y}_j$ の値の組み合わせを以下のように全探索することにより，その最適解を求めることができる．

STEP1 2 値ベクトル $\boldsymbol{y}_j$ $(j = 1, \ldots, k)$ を設定する．

STEP2 (8) 式のパラメータを推定する．

STEP3 STEP1–STEP2 をすべての代表点のパターンにおいて繰り返し，最適な代表点，およびそのときのパラメータを決定する．

これにより，回帰モデルとデータとの当てはまりの度合いを最大化するような 2 値型代表点を求め，クラスタリングしながらロジスティック回帰式を同定することができる．

3.2 AIC によるモデル選択

本研究の提案モデルは，代表点の数 k をいくつに設定するかによって (7) 式の値が変化するモデルとなっている．これに対して k の値に予め制約がある場合（e.g., 求めたいクラスタの数が決まっている場合）は，その k の数を用いればよいが，そうでない場合は，k の数だけ存在する分析結果の中から，最適なモデルを選択しなければならない．そこで，以下のような AIC（赤池情報量基準）を用いたモデル選択方法を提案する．

まず，k 個の代表点を用いて (7) 式を最適化したときのモデルを MODEL(k)，このときの修正対数尤度を $\log FL(k)$ で表すことにする．このときの AIC は

$$-2\times\log\ FL(k)+2\times(k+m) \quad (11)$$

で表すことができる．k の値が特に決められていない場合には，この値が最も小さくなるような MODEL(k) を選択していくことにする．

4. セール品に着目した顧客の購買行動モデルの解析への提案モデルの適用

4.1 解析データおよび解析内容

本節では，提供された顧客の購買データのうち，初期購買においてすべてバーゲン品を購買した 15,593 人を対象として，リピーターになる（$z_i = 1$ とする）か初期退出者になる（$z_i = 0$ とする）か，また，バーゲン優良顧客になる（$z_i = 1$ とする）かバーゲンハンターになる（$z_i = 0$ とする）か，の 2 種類の予測モデルを構築していく．そこで，顧客の属性データおよび顧客が初期購買を行う時点で得られる以下のような 11 個の説明変数（男女（男性：1，女性：0）・年齢 ・会員登録から初期購買までの日数・初期購買で使った金額 ・初期購買の購買数・どのような商品大分類の組み合わせを買ったか（トップス，パンツ，ワンピース，シューズ，バッグ，ジャケットを買ったかどうかの 6 変量 2 値変数））を用いて解析していく．

これにより，顧客が初期購買をした時点で，どの種類の顧客になるかを予想することができるようになる．ただし，予測モデルの妥当性について考察するために，2 種類のデータセットそれぞれにおいて顧客の初期購買日が均一になるように学習データおよびテストデータへと 2 分割し，(7) 式を用いて 2 値データの部分をクラスタリングしながら Firth 法を用いたロジスティック回帰式を同定する提案モデルと，2 値変数をそのまま変数として用いて (6) 式で表される Firth 法を用いたロジスティック回帰モデル（従来モデル）によってそれぞれ解析していくことにする．また，提案法での解析における k の値は，(11) 式の値を用いて決定する．

表 2　AIC の比較

k の値	AIC
3	9662.748
4	8859.782
5	**8655.037**
6	8656.037

4.2　解析結果の評価

本研究の提案モデルは，回帰モデルとデータとの当てはまりの度合いを最大化するような 2 値型代表点を求め，クラスタリングしながらロジスティック回帰式を求めるための方法として位置づけられる．そこで，得られたモデルについて，(i) パラメータ数の差異，(ii) 精度の差異，(iii) 回帰係数の解釈の差異，を従来法によって得られた結果と比較していくことにする．

ただし，(ii) 精度の差異については，学習データとテストデータを，提案法および従来法によりそれぞれ解析し，修正対数尤度の値と判別の精度の値，予測値と実績値のクロス表，適合率，再現率，および F 値を求め，比較していくことにする．

4.3　解析結果

4.3.1　リピーターと初期退出者の判別

リピーターと初期退出者の判別の結果は以下のとおりである．この解析において，k の値を 3 から 6 まで 1 つずつ増やしていったところ，表 2 のように $k = 5$ のときの AIC の値が最も小さくなったため，その解析結果を採用した．

このときの提案法で求めた代表点は（該当する品目以外），（パンツ・シューズ・ジャケット），（トップス・パンツ・ワンピース・シューズ・バッグ・ジャケット），（トップス・パンツ・ワンピース・シューズ・バッグ），そして（バッグ・ジャケット）であり，これら初期購買の商品分類のパターンをもとに顧客を 5 つのグループへと分割した．

まず，(i) の自由度について，従来法と提案法とで比較する．従来法では 11，提案法では 10 となり，提案法のほうが自由度が少ないモデルとなっている．次に，(ii) の精度について，それぞれの解析における (6) 式または (7) 式の値（i.e., 修正対数尤度），判別率，クロス表，適合率，再現率，および F 値を表 3–5 に示す．ただし，精度およびクロス表は，ROC 曲線で最適となる境界値を用いて計算している（提案法，従来法ともに，0.250 を境界値とした）．

表 3–5 より，学習データおよびテストデータのすべての評価指標において提案法が優れていることがわかる．よって提案法により，少ないパラメータ数でデータにフィットし，汎化能力が高いロジスティック回帰分析による予測式を求めることができていると考えられる．

表 3　修正対数尤度の値と判別精度の比較

	学習		テスト	
方法	修正対数尤度	精度	修正対数尤度	精度
提案	−4317.519	0.683	−4351.432	0.700
従来	−4426.487	0.675	−4604.862	0.667

表 4　解析結果のクロス表（行：実際，列：予測）

	学習		テスト	
提案法	リピーター	初期退出者	リピーター	初期退出者
リピーター	3,811	1,011	3,817	1,055
初期退出者	1,457	1,518	1,287	1,637
従来法	リピーター	初期退出者	リピーター	初期退出者
初期退出者	3,780	1,042	3,753	1,119
リピーター	1,493	1,482	1,472	1,452

表 5　適合率，再現率，F 値の比較

	学習			テスト		
方法	適合率	再現率	F 値	適合率	再現率	F 値
提案	0.723	0.790	0.755	0.748	0.784	0.765
従来	0.717	0.784	0.749	0.718	0.770	0.743

次に，(iii) 回帰係数の解釈の差異について，提案法および従来法によって求めた偏回帰係数，標準誤差，χ^2 の p 値を表 6 および表 7 に示す．ただし，提案法において，代表点をダミー変数化する際には，この大分類に該当するものを買っていない人の効果（i.e., 偏回帰係数）を 0 に設定している．

ここで提案法と従来法によって得られたモデルの違いを解釈していくことにする．まず，上記の結果から，提案法においても，従来法についても，男性であること，年齢が高いこと，サイト登録から購買までの時間の短さが初期退出者のなりやすさに影響を与えていることは共通して読み取ることができるが，合計金額についてはプラスとマイナスが逆になっている．

次に得られた代表点について考察していくことにする．提案方法では，それぞれの代表点に近いデータのクラスタリング結果を説明変数としている．この解析結果を見ると，それぞれのパターンに近い購買をしている場合の効果がわかる．この解析では，該当する大分類以外の効果を 0 に固定しており，購買した大分類の種類が少ない顧客の多くは，このクラスタに属している．すなわち，多くのパターンを購買している顧客のほうがリピーターになりやすいという解釈を得るこ

表 6　提案法による解析結果

学習	β	標準誤差	χ^2 の p 値
切片	−0.0486	0.001	0.625
男性	−0.169	0.052	0.000
年齢	−0.003	0.274	0.216
購買金額	0.000	0.000	0.000
サイト登録からの日数	0.006	0.000	0.000
購買数	−0.016	0.178	0.355
パンツ・シューズ・ジャケット	4.321	1.716	0.000
トップス・パンツ・ワンピース・シューズ・バッグ・ジャケット	6.442	1.416	0.000
トップス・パンツ・ワンピース・シューズ・バッグ	6.508	1.416	0.000
バッグ・ジャケット	1.027	1.440	0.000

表 7　従来法による解析結果

学習	β	標準誤差	χ^2 の p 値
切片	−0.643	0.109	0.000
男性	−0.230	0.054	0.000
年齢	−0.006	0.003	0.028
購買金額	−0.000	0.000	0.000
サイト登録からの日数	0.006	0.000	0.000
購買数	−0.016	0.018	0.377
トップス	1.375	0.062	0.000
パンツ	0.113	0.084	0.178
ワンピース	1.184	0.095	0.000
シューズ	0.907	0.096	0.000
バッグ	0.312	0.081	0.000
ジャケット	0.330	0.092	0.000

とができる．一方，従来法では，それぞれの商品大分類の効果ごとの購買する効果をモデルに組み込んでいる．ただし，このモデルは交互作用を考慮していないため，購買した商品大分類の組み合わせについての情報を得ることはできない．当然，交互作用をモデルに組み込むことも可能であるが，さらに自由度が大きくなり，汎化能力が悪化することが考えられる．

よって，(i)，(ii)，(iii) の評価指標により，提案法によって得られたモデルのほうが少ない自由度でより精度のよい結果を得られるとともに，購買した商品大分類の組みわせについての情報を考慮したモデルとなっていることがわかり，結果の妥当性が示されている．

4.3.2　バーゲン優良顧客とバーゲンハンターとの判別

この解析においても，k の値を 3 から 6 へ変化させてそれぞれの AIC を求めたところ，表 8 のように $k = 4$ に設定したときに提案法による AIC の値が最小になったため，$k = 4$ のときの解析結果を採用した．また，このときの提案法で求めた代表点は（トップス・パンツ・ワンピース・シューズ・バッグ・ジャケット），（トップス・パンツ・ワンピース，ジャケット），（トップス・パンツ），そして（ワンピース，バッグ）であり，初期購買の商品大分類の組み合わせをもとにリピーターの

表 8　AIC の比較

k の値	AIC
3	6296.072
4	**6277.744**
5	6278.991
6	6277.961

表 9　対数尤度の値と判別精度の比較

	学習		テスト	
方法	修正対数尤度	精度	修正対数尤度	精度
提案	−3129.872	0.654	−3160.876	0.641
従来	−3133.100	0.651	−3234.628	0.633

表 10　解析結果のクロス表（行：実際，列：予測）

	学習		テスト	
提案法	優良顧客	ハンター	優良顧客	ハンター
優良顧客	1,108	1,109	1,084	1,160
ハンター	570	2,059	577	2,027
従来法	優良顧客	ハンター	優良顧客	ハンター
優良顧客	1,089	1,128	1,068	1,176
ハンター	562	2,067	605	1,999

表 11　適合率，再現率，F 値の比較

	学習			テスト		
方法	適合率	再現率	F 値	適合率	再現率	F 値
提案法	0.660	0.500	0.569	0.653	0.483	0.555
従来法	0.660	0.491	0.563	0.638	0.476	0.545

顧客を 4 つのグループへと分割した．

このときの (i) 自由度について比較すると，提案法が 9，従来法が 11 となっており，提案法により求めたモデルのほうが自由度が少ないことがわかる．さらに (ii) 精度の比較のために，(6) 式の値または (7) 式の値（修正対数尤度），判別率，クロス表，適合率，再現率，および F 値を表 9–11 に示す．ただし，精度およびクロス表は，ROC 曲線で最適となる境界値を用いて計算している（提案法，従来法ともに，0.250 を境界値とした）．

この指標においても，大きな差があるわけではないが，提案法がすべての指標において従来法よりも優れていることがわかる．すなわち，バーゲンハンターになるのか，バーゲン優良顧客になるのかのよりよい予測モデルを同定することができていると考えられる．

次に，(iii) 得られたモデルの解釈について記述していくことにする．そこで，提案法および従来法によっ

表 12　提案法による解析結果

学習	偏回帰係数	標準誤差	χ^2 の p 値
切片	0.591	0.133	0.000
男性	0.260	0.063	0.000
年齢	−0.014	0.003	0.000
購買金額	0.000	0.000	0.010
サイト登録からの日数	0.005	0.000	0.000
購買数	−0.270	0.024	0.000
トップス･パンツ･ワンピース･シューズ･バッグ･ジャケット	−0.724	0.554	0.191
トップス･パンツ･ワンピース･シューズ･ジャケット	−0.878	0.225	0.000
トップス・パンツ	−0.197	0.064	0.002

表 13　従来法による解析結果

学習	β	標準誤差	χ^2 の p 値
切片	0.538	0.131	0.000
男性	0.268	0.065	0.000
年齢	−0.014	0.003	0.000
購買金額	0.000	0.000	0.001
サイト登録からの日数	0.005	0.000	0.000
購買数	−0.242	0.027	0.000
トップス	−0.208	0.071	0.003
パンツ	−0.343	0.093	0.000
ワンピース	−0.041	0.087	0.638
シューズ	−0.130	0.096	0.176
バッグ	0.050	0.091	0.584
ジャケット	−0.135	0.112	0.226

て求めた偏回帰係数，標準誤差，χ^2 の p 値を表 12 および表 13 に示す．ただし，提案法においては，（ワンピース・バッグ）の組み合わせのグループに属する顧客の偏回帰係数を 0 に固定している．これらの結果から，提案法および従来法の両方において男性であること，年齢が低いこと，購買金額が高いこと，サイトに登録してからの日数が長いこと，そして購買数が少ないことがバーゲン優良顧客になるための条件として効いていることが示唆される．

また，提案法から得られた代表点より，含まれる大分類の種類が少ないデータが（ワンピース・バッグ）または（トップス・パンツ）のクラスタ，大分類の種類が多いデータがその他の代表点に含まれることがわかる．次に分割されたクラスタそれぞれの効果に着目すると，（トップス・パンツ）の組み合わせの代表点に含まれるデータのクラスタの負の効果が小さいことから，購買品の種類が多いことがバーゲンハンターとになる要因として効いていると解釈することができる．

ここで，従来法によって得られたモデルの p 値に着目すると，ワンピース，バッグ，そしてジャケットの効果の p 値が大きくなっていることがわかる．これに対して，提案法のように代表点によるクラスタリングというアプローチをとることにより，p 値の大きい変数も考慮に入れたモデルを構築することができている．

以上の結果から，バーゲン優良顧客になるかバーゲンハンターになるかのデータに関しても，提案手法の適用により，少ない自由度でより精度のよいモデルを同定することができることがわかる．

4.4　考察

本節では，モデルの妥当性について考察し，解析結果からの示唆を性別の違いに着目しながら論じていくことにする．

まず，本研究では，データによりフィットしたロジスティック回帰モデルを構築するために，2 値型 principal points を応用し，パラメータ数を減らしながら最適なロジスティック回帰式を同定するための方法を提案している．実際のデータに対して提案法および従来法を適用することにより，(i) 自由度，(ii) 精度，そして (iii) モデルの解釈の 3 つの指標により，提案法と従来法を比較し，提案法を適用することの妥当性を示すことができた．

次に，2 つの解析結果を「性別の違い」という視点から考察していく．まず，リピーターになるか初期退出者になるかの判別では，男性はリピーターになりにくいという結果になったが，バーゲン優良顧客になるかバーゲンハンターになるかの判別では，男性はバーゲン優良顧客になりやすいという結果になった．この結果から，男性は，2 回目の購買につなげることができれば，バーゲン品以外の商品も購買するバーゲン優良顧客になりやすいと解釈することができる．そこで，バーゲン品を初期購買した男性の顧客に対してはバーゲン品を押し出した宣伝が有効なのではないかと思われる．

一方，女性は，男性に比べてリピーターになりやすいが，バーゲン優良顧客になりにくいという解析結果が示された．よって女性の新規顧客に対しては，値段をセールスポイントとして宣伝するのではなく，品質やファッション性などをアピールしながら，プロパー価格での購買を促すべきであるものと考えられる．

5.　おわりに

本研究では，2 値の応答をもつデータに対して，少ないパラメータでより正確な予測モデルを構築するために，2 値型 principal points のアプローチを応用したロジスティック回帰モデルを提案した．さらに，提案したモデルを，バーゲン利用者の行動パターンのデータに関する 2 つのデータセットへと適用し，予測モデルを構築した．また，得られた予測モデルから初期購

買がバーゲン品のみの顧客がリピーターになるためのアプローチおよびバーゲン優良顧客になるためのアプローチについて考察した.

謝辞　データを提供していただいたデータ解析コンペティション事務局の皆様，および貴重なコメントを頂いた 2 名の査読者の方に感謝いたします.

参考文献

[1] 経済産業省，「平成 24 年度我が国情報経済社会における基礎基盤（電子商取り引きに関する市場調査）」，http://www.meti.go.jp/press/2013/09/20130927007/20130927007-4.pdf（2014 年 7 月 16 日閲覧）
[2] T. Doganoglu and J. Wright, "Exclusive dealing with network effects," *International Journal of Industrial Organization*, **28**, 145–154, 2010.
[3] H. Yamashita and H. Suzuki, "Heuristic approximation methods for principal points for binary distributions," *Journal of Japan Industrial Management Association*, **65**, 131–141, 2014.
[4] B. Flury, "Principal points," *Biometrika*, **77**, 33–41, 1990.
[5] T. Tarpey, L. Li and B. Flury, "Principal points and self-consistent points of elliptical distributions," *The Annals of Statistics*, **23**, 103–112, 1995.
[6] 清水信夫，水田正弘，佐藤義治，"Principal Points の対称性に関する定理について，" *Journal of the Japanese Society of Computational Statistics*, **12**, 45–53, 2000.
[7] 村木千恵，大瀧慈，水田正弘，"主要点解析法による極東夏期天気図の分類，" *Japanese Journal of Applied Statistics*, **27**, 17–31, 1998.
[8] S. Matsuura, "Optimal partitioning of probability distributions under general convex loss functions in selective assembly," *Communications in Statistics – Theory and Methods*, **40**, 1545–1560, 2011.
[9] 大倉征幸，鎌倉稔成，"小標本かつ応答変数発現確率が高い場合のロジスティック回帰モデルにおける回帰パラメータの検定法，" *Japanese Journal of Applied Statistics*, **40**, 41–51, 2011.
[10] D. Firth, "Bias reduction of maximum likelihood estimates," *Biometrika*, **80**, 27–38, 1993.

