論文・事例研究

百貨店における隠れた親近性の発掘

オウ　ロ，吉原　亜弥，矢島　安敏

1. はじめに

本研究は百貨店におけるクレジットカード利用履歴データを分析することで，顧客と商品の間に潜む隠れた親近性を発掘するための手法を提案し，その有効性の検証を行う．一般にクレジットカードの利用は，ポイント制の現金カードとは異なり，顧客によっては高額商品などの一部の購買に限定されてしまう．したがって，このデータには，百貨店での顧客の購買行動のごく一部しか記録されていないと考えられる．このような状況では，購買記録のない商品の購入数（金額）を一律に「0」であるとみなし分析を進めることは，明らかに適当でない．そこで，購入履歴のある商品の組み合わせを分析することによって，未購買商品に対して（もちろん購入履歴のある商品に対しても），ある種の購入の可能性や商品への選好の程度を表す指標，いわば隠れた「親近性」の算出を試みる．

本研究では，強力なパターン認識法として近年注目されているサポート・ベクタ・マシン（Support Vector Machine, SVM)[8]を応用した，1クラスSVM（1-SVM)[5]と呼ばれる手法を用い，未購買商品に対する親近性の算出を試みる．1-SVMは，属性の空間に分布しているデータ（点）をできるだけ多く包含しつつ，かつできるだけ小さな領域を算出する．また，特殊な凸二次計画問題の最適化アルゴリズムを用いることで，データ数が数万を超えるような大規模データに対しても適応可能な手法である．さらに，カーネル関数を用いた非線形変換を組み合わせると，複雑な形状の領域で分布の様子を表現することが可能になる．従来，1-SVMは，求めた領域の外部にある点に注目し，例えば，外れ値の検出といった分野に用いられている手法である．本研究では，トレードオフパラメータと呼ばれる値を変化させながら，最適化問題を複数回解き，大きさの異なる領域を算出し，これを基に親近性の算出を行う．

百貨店のデータに限らず，CDや書籍などの購買履歴データに対しては，未購買商品を顧客に推薦するための手法が研究されている．特に協調フィルタリング（collaborative filtering）と呼ばれる技術として，例えば，GroupLens project[4]やRingo[6]などとして研究が行われている．本研究では提案手法の有効性を確認するため，cross-validationによりこれらの方法との比較・検証を行った．その結果，提案法は従来から用いられているこれらの手法を上回る性能が発揮できることを確認した．

2. データの概要

分析に用いたデータには，百貨店の持つタイプの異なる三つの店舗のデータが含まれていた．本研究では，売り上げ最大の店舗一つに注目し，2001年7月から2003年6月までの2年間分のデータを用いた．各レコードは，一回のカード利用ごとに，顧客ID，日付，価格および商品に関する情報から構成されている．各商品には，約600の「アイテム」番号と約600の「売場」番号が付与されている．ここで「アイテム」と呼んでいるのは，例えば，婦人靴，婦人スカート，ハンドバッグといったレベルでの商品の分類である．「売場」とは，グッチ，ルイ・ヴィトンといったテナント名あるいは，アクセサリ，鞄といった店内の売場の名称である．すべてのアイテムがすべての売場で販売されてはいないので，「アイテム」と「売場」の組み合わせで商品を分類した場合には約4,400種類となる．以降，この分類のことを単に商品と呼び，分析を進める上での商品に対する最小の分類とする．本研究では，提案する手法をダイレクトメール戦略等に適用することを考慮に入れ，ある程度に細かく分類された商品ご

おう　ろ
東京海上日動火災保険㈱ 財務企画部
〒100-0004 千代田区大手町1-5-1
よしはら　あや，やじま　やすとし
東京工業大学 経営工学専攻
〒152-8552 目黒区大岡山2-12-1
受付04.7.28　採択04.9.5

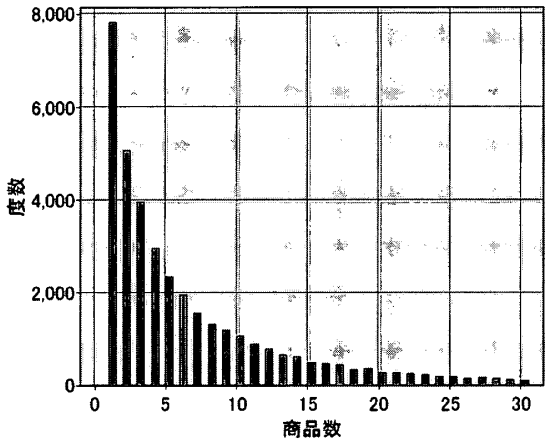

図1　商品数による顧客の分布

とであっても親近性の算出が行える手法を提案するものである．

本研究で用いたデータでは，顧客IDの総数が約3.8万人であるのに対し，レコード総数が約56.6万となっており，顧客IDの総数に比べてレコード数があまり多くない．顧客ごとに購入した商品の種類を集計し，ヒストグラムとしたものが図1である（30商品以上購入した度数（人数）は省略した）．この図から明らかなように，大半の顧客は購入した商品の種類がわずか数個にとどまっている．実際，顧客一人が購買した商品の種類は平均約7.6であり，また，12種類以下の顧客で全体の90%以上を占めており，商品が4,000種類以上あるのと比べ極めて少ない．これは，データがクレジットカードの利用という非常に限定された購買行動のみの記録であるためと考えられる．節3では，顧客一人一人の商品の購買パターンを用いて，データ上では購入数「0」となっている商品に対しても，顧客の嗜好を反映した，いわば「隠れた親近性」の算出を試みる．

3.　1クラスSVMを用いた親近性の算出

3.1　1クラスSVM

本節では，提案する親近性の算出方法について述べる．商品の種類をNとして，顧客iの購買の様子を商品ごとの購入回数を要素とするN次ベクトルで$x_i \in \mathbb{R}^N$と表し，以降では顧客iの購買ベクトルと呼ぶことにする．

今，ある商品yに注目し，顧客と商品yとの親近性を考える．まず，商品yを購入したことのある顧客のみの購買ベクトルx_iからなる集合をX_yとする．なお簡単のため，yを購入した顧客はL人として，

$$X_y=\{x_1, x_2, \ldots, x_L\}$$

と仮定する．1-SVMでは，集合X_yに属するN次空間の点をできるだけ小さな球で包含し，X_yに属するベクトルの特徴を抽出しようと考える．そこで，未知変数として球の半径の二乗をR，また球の中心をベクトル$a \in \mathbb{R}^N$とすれば，半径最小の球は次の凸計画問題：

$$\left|\begin{array}{ll} \text{最小化} & R \\ \text{制約} & \|x_i-a\|^2 \le R,\ i=1,2,\ldots,L \end{array}\right. \tag{3.1}$$

を解いて求めることができる．

1-SVMでは，包含されない点が存在することも許容し，その代わりに，包含されない点に対してはペナルティを与え半径とともに最小化を考える．すなわち，ペナルティを表す非負の変数$\xi_i (i=1,2,\ldots,L)$を導入し，次の最適化問題：

$$\left|\begin{array}{ll} \text{最小化} & R+\dfrac{1}{\nu L}\displaystyle\sum_{i=1}^{L}\xi_i \\ \text{制約} & \|x_i-a\|^2 \le R+\xi_i,\ i=1,2,\ldots,L, \\ & \xi_i \ge 0,\ i=1,2,\ldots,L, \end{array}\right. \tag{3.2}$$

を考える．ただし，νはペナルティ項ξ_iへの重みをコントロールするパラメータで，$0<\nu<1$の範囲で定められるものとする．

いま，適当なパラメータνのもとで問題(3.2)を最適化し，求められた球の中心を$a^* \in \mathbb{R}^N$，半径の二乗をR^*とする．ことき，ある顧客の購買ベクトル$x \in \mathbb{R}^N$が球の内部に存在する，すなわち

$$\|x-a^*\|^2 \le R^* \tag{3.3}$$

を満たせば商品yを購入した顧客の購買ベクトルと似ていると判断する．なお，問題(3.2)については，この後述べる理由により，その双対問題を解けば最適解が算出される．問題(3.2)の双対問題は，$\alpha_i (i=1,2,\ldots,L)$を双対変数として

$$\left|\begin{array}{ll} \text{最小化} & \displaystyle\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i\alpha_j\langle x_i, x_i\rangle - \sum_{i=1}^{L}\alpha_i\langle x_i, x_j\rangle \\ \text{制約} & \displaystyle\sum_{i=1}^{L}\alpha_i=1,\ 0\le\alpha_i\le\frac{1}{\nu L},\ i=1,2,\ldots,L, \end{array}\right. \tag{3.4}$$

となる．ただし，$\langle x_i, x_j\rangle$はベクトルx_iとx_jの内積を表す．双対問題は，等式制約が1本と変数の上下限制約のみの非常に単純な二次計画問題となるため，この構造を用いた高速な最適化アルゴリズム[1]が提案されている．また，主問題(3.2)および双対問題(3.4)の最適解をそれぞれ，(a^*, R^*, ξ^*)およびa^*と書けば，R^*は双対問題(3.4)の等式制約に対応した双対変数であり，また，

$$a^*=\sum_{i=1}^{L}\alpha_i^* x_i$$

の関係[7]から，主問題の解を得ることができる．

3.2 カーネル関数を用いた定式化

前節の定式化では，単純な球形の領域しか扱うことができない．一般に購買の組み合わせには相関があることから，領域の形状として楕円体あるいはより複雑なものを考慮すべきであろう．集合 X_y の分布の特徴をよりよく反映させるため，SVMでは，カーネルトリックと呼ばれる方法が通常用いられる．これは，適当な写像 $\phi: \mathbb{R}^N \mapsto \mathcal{F}$ を使い非線形に変換したデータ

$$\{\phi(x_i)\in\mathcal{F} | x_i\in X_y\}$$

に対して，問題 (3.2) あるいは問題 (3.4) と同様な最適化問題を考えるというものである．このとき，双対問題を考えれば，

$$\left|\begin{array}{ll}\text{最小化} & \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i\alpha_j\mathcal{K}(x_i,x_i)-\sum_{i}^{L}\alpha_i\mathcal{K}(x_i,x_j)\\ \text{制約} & \sum_{i=1}^{L}\alpha_i=1,\ 0\le\alpha_i\le\frac{1}{\nu L},\ i=1,2,\ldots,L\end{array}\right. \tag{3.5}$$

と書き表される．ただし，$\mathcal{K}(x_i, x_j)$ は空間 $\mathcal{F}$ の元 $\phi(x_i)$ と $\phi(x_j)$ の内積を表す関数，すなわち

$$\mathcal{K}(x_i, x_j)=\langle\phi(x_i), \phi(x_j)\rangle$$

とする．双対問題 (3.5) の係数を定めているのは，空間 $\mathcal{F}$ における内積の値のみであることに注目すれば，非線形写像 ϕ を陽に与える必要はなく，元のデータ x_i と x_j から $\mathcal{F}$ での内積の値 $\langle\phi(x_i), \phi(x_j)\rangle$ が求められればよい．このような場合に，最も頻繁に用いられるものの一つにRBFカーネル関数があり，

$$\mathcal{K}(x_i, x_j)=\langle\phi(x_i), \phi(x_j)\rangle=\exp(-\|x_i-x_j\|^2/2\sigma) \tag{3.6}$$

と定義されている．ここで，σ はあらかじめ定められたパラメータで，σ を変えることで非線形変換 ϕ が変化することになる．さらに，本研究では，購買ベクトル x_i を

$$x_i \leftarrow \frac{x_i}{\|x_i\|}$$

とノルムを1に基準化した上で，式(3.6)を用いることにする．したがって，$\|x_i\|^2=\|x_j\|^2=1$ であるから，式(3.6)は，

$$\mathcal{K}(x_i, x_j)=\exp((x_i^T x_j-1)/\sigma) \tag{3.7}$$

と書き直される．

ある ν に対して，問題 (3.5) の最適解を α_i^* また，等式制約に対応した最適な双対変数を R^* とすれば，非線形変換した場合の領域も，

$$f(x)=1+\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i^*\alpha_j^*\mathcal{K}(x_i, x_j)-2\sum_{i=1}^{L}\alpha_i^*\mathcal{K}(x_i, x)$$

とすれば，

$$S(\nu)\equiv\{x | f(x)\le R^*\}$$

となり，ϕ を使うことなく内積の値のみで記述することが可能である．特に，式(3.7)より，x_i と x_j に共通な購買商品がない場合，すなわち $x_i^T x_j=0$ と直交する場合には，カーネル関数は $\mathcal{K}(x_i, x_j)=\exp(-1/\sigma)$ と定数となる．したがって，購買ベクトル $\hat{x}$ が X_y に属するすべてのベクトルと直交する場合には，

$$f(\hat{x})=1+\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i^*\alpha_j^*\mathcal{K}(x_i, x_j)-2\exp(-1/\sigma)$$

も $\hat{x}$ によらず定数となる．また，ベクトル x_i は非負であり，$0\le x_i^T x_j\le 1$ となることより，

$$\exp((x_i^T x_j-1)/\sigma)\ge\exp(-1/\sigma)$$

であり，ゆえに $f(\hat{x})$ は $f(x)$ の上界となる．すなわち，X_y のいずれとも共通な商品の購入のない購買ベクトル $\hat{x}$ が，最も X_y と似ていないこととなる．

3.3 親近性の指標の算出

最後に上で求めた領域 $S(\nu)$ より，親近性の指標を算出する．いま，ある ν に対して問題 (3.5) の最適解を α_i^* とする．ペナルティ項への重み ν に関連して次の定理[5]が知られている．

定理3.1 νL は球の外部となる点の個数の上限であり，かつ α_i^* が正となるものの個数の下限である．

すなわち，小さな ν に対応した領域 $S(\nu)$ は多くの X_y の点を含むものとなり，したがって，共通性の基準を広範囲に捕らえたものと考えられる．一方，ν を大きくするほど $S(\nu)$ はより半径の小さな領域となり，特徴的なベクトルのみを捕らえたものとなると考えられる．

そこで，適当な間隔で，

$$0<\nu^1<\nu^2<\cdots<\nu^r<1$$

と r 個の ν をあらかじめ定め，それぞれに対応する領域

$$S(\nu^1), S(\nu^2), \ldots, S(\nu^r)$$

を計算する．その上で，購買ベクトルが x である顧客の商品 y に対する親近性を，

$$R(x)=\sum_{k=1}^{r}I(\nu^k, x),$$

ただし

$$I(\nu, x)=\begin{cases}1 & \text{if } x\in S(\nu)\\ 0 & \text{o. w.}\end{cases}$$

と定めることにする．多くの領域 $S(\nu^k)$ に含まれる購買ベクトル x ほど $R(x)$ は大きな値となることから，

ある商品 y への親近性と考えることにする．

3.4 データを用いた検証

前節で提案した親近性指標の妥当性を，百貨店のデータを用いて検証する．まず，2年間で30人以上の顧客に対して販売実績のあった商品約1,700に限定し，かつ，これらの商品を3種類以上購入した顧客約2.4万人を対象にした．なお，後述する比較手法の計算を実行する際，非常に多くの主記憶を必要としてしまうため，さらにランダムに25%の顧客をサンプリングした小さなデータセットを作成した．その結果，商品数1,668種類，顧客数6,007人のデータに対して実験を行った．1-SVMによる提案手法の場合は，サンプリング前のデータ規模であっても十分に計算可能であるが，ここでは比較のため同一のデータを用いた．

実験は購入顧客数の多い50商品それぞれで，顧客に対する親近性を算出した．表1には，この50の商品名—すなわちアイテム名と売場名—と実際の購入顧客数を示した．大半の商品は購入顧客数が100人程度であり，これは総顧客数の3%以下となっている．

実験では次のように4-fold cross-validationを行った．まず，全顧客をランダムに四つのグループに等分する．その内一つのグループに属する顧客の購入履歴から，商品 y に関するデータを削除，すなわち購入数を0としたテストデータを作成する．残りの三つのグループに属する顧客の履歴をトレーニングデータとして用い，テストデータの顧客に対して商品 y への親近性を算出する．得られた親近性にしたがって顧客を選んだ場合，実際に商品 y を購入した顧客をどれだけ正確に抽出できるかを，リフト図や再現率の値で評価する．同じ実験を購入数を0とする顧客のグループを変えながら4回行い，平均的なパフォーマンスを考える．1-SVMのカーネル関数としては，式(3.6)のRBFカーネルを用い，パラメータは $2\sigma=1,668\times10$（商品数×10），ν は(0,1)区間を50等分($r=50$)と設定した．

提案手法との比較のため，協調フィルタリングで標準的に用いられている相関係数法[3]，および近年Hofmann等の提案した確率的潜在クラスモデル(probabilistic Latent Semantic Analysis, pLSA)[2]を用いて同様の実験を行った．なお，pLSAでは潜在クラス数を $K=50, 100, 200, 300, 400$ と5通り設定して実験を行った．前述したように，顧客数を25%に減らした場合でも，潜在クラス数 $K=400$ で計算を行

表1 実験を行った商品名とその購入顧客数

アイテム名	売場名	購入顧客数
婦人靴	キャラクタシューズ	813
婦人靴	婦人靴	794
ハンドバッグ	ハンドバッグ	615
婦人衣料 ご奉仕品	ミセスカジュアル	464
ハンドバッグ	ルイ・ヴィトン	370
ファンデーション	ファンデーション実績	355
婦人衣料 ご奉仕品	ドレス&コート	334
婦人衣料 ご奉仕品	クローバーミセス	295
ショーツ	ショーツ実績	290
婦人パンツ	ジーンズ	277
婦人財布・小物ケース	ルイ・ヴィトン	268
婦人衣料 ご奉仕品	セータ(自主)	239
ハンドバッグ	タウンハンドバッグ	237
婦人衣料 ご奉仕品	ヤング その他	227
ネクタイ	ネクタイ	215
⋮	⋮	⋮
男児衣料	カジュアルウェア	151
婦人ニット	セータ(NB)	145
婦人ブラウス	ミセスカジュアル	144
女児衣料	ガールズ	143
メンズゴルフウェア	ゴルフウェア	143
メイクアップ	C・ディオール	140
パジャマ・ネグリジェ	ナイトウェア実績	139
紳士スーツ	紳士スーツ	139
カバー・シーツ類	寝具雑貨	138
ハンドバッグ	コーチ	137
婦人ジャケット	マダムプレタ	136
男児衣料	マックレガー	136
カーペット	敷物	136
婦人靴	モードエジャコモ	131
ハンドバッグ	グッチ	130

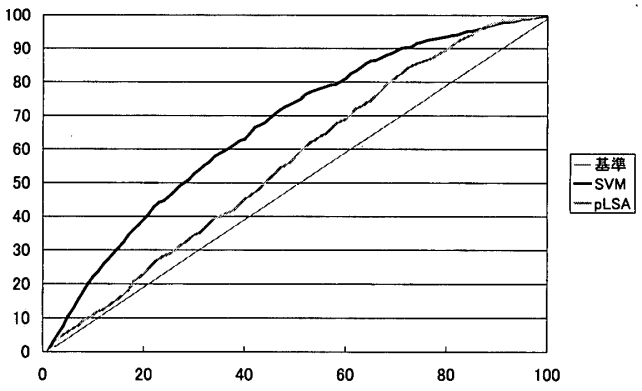

図2 キャラクタシューズのリフト図

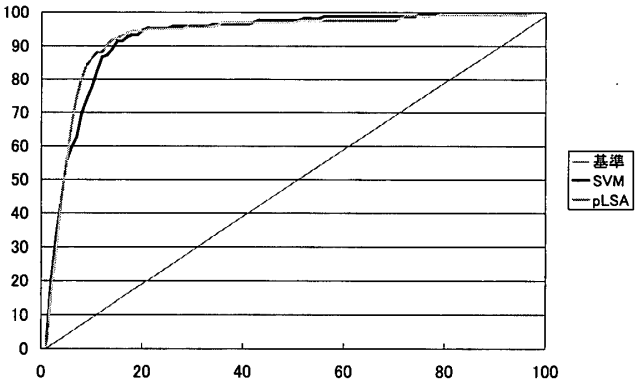

図3 ミセスカジュアルスカートのリフト図

表2　50商品の平均再現率の比較

	1%	2%	5%	10%	20%
1-SVM	11.0	17.6	30.0	42.9	60.7
$K=50$	5.8	10.9	24.1	40.8	58.5
$K=100$	7.2	13.7	29.7	43.9	57.7
$K=200$	8.4	15.5	31.4	41.4	53.6
$K=300$	9.1	16.4	29.2	37.4	51.3
$K=400$	8.4	14.6	25.1	33.1	48.2
相関係数法	3.7	4.9	6.3	8.2	12.2

うには1GB程度の主記憶を必要とするため，これ以上大規模なデータでの計算は行うことができなかった．それぞれの手法の詳細は参考文献を参照されたい．

図2には，最も購入顧客数の多かった「キャラクタシューズ」について，また，図3には，最も予測精度の高かった商品の一つ，売場名「ミセスカジュアル」での「スカート」のリフト図を示した．これらの図は，予測した親近性の降順に顧客を並べ，上位からある割合（横軸）の顧客を選び出したときの再現率，すなわち，

$$\frac{\text{選び出された顧客の中で}\ y\ \text{を購入した人数}}{\text{全顧客の中で}\ y\ \text{を購入した人数}}\times 100$$

の値を縦軸にグラフ化したものである．どちらも濃い黒線は1-SVMを用いた提案手法の結果，灰色の線が$K=200$とした場合のpLSAの結果である．

表2は，実験した50商品すべてに対する平均パフォーマンスを示したものである．親近性上位1%，2%，5%，10%および20%の顧客を選び出した場合の再現率を，手法のパラメータを変化させながら示した．提案手法である1-SVMを用いたものが他と比べ全般的に高い性能となっており，本手法の有効性を示している．

4. 売場別の分析

この節では，親近性の指標を百貨店における売場の分析に応用することを試みる．まず，前節までは商品ごとに行ってきた集計を売場（テナント）別のものに変え，顧客ごとに各売場での購入回数を要素とする購買ベクトルを算出する．なお，データでは，商品がセール（奉仕）品として販売されたことが識別できるよう分類がなされていた．そこで，同じ売場でもセール時と通常時では異なる売場として扱い分析を試みた．

顧客一人が利用した売場数の様子をヒストグラムにしたものが図4である．集計の結果，利用した売場が二つ以下の顧客数は全体の38%，顧客一人が利用した売場数の平均は約5.2店であり，全体の売場数約600と比べ少ないことが分かる．また，デシル分析を行うと，全体の9割の顧客は利用した売場数がわずか12店（2%）以下であった．そこで，まず，前節で述べた親近性の指標を各売場に対して計算し，（データ上では）利用したことのない売場に対しても親近性を付与する．その上で，求めた親近性指標を用い双対尺度法による分析を試みた．

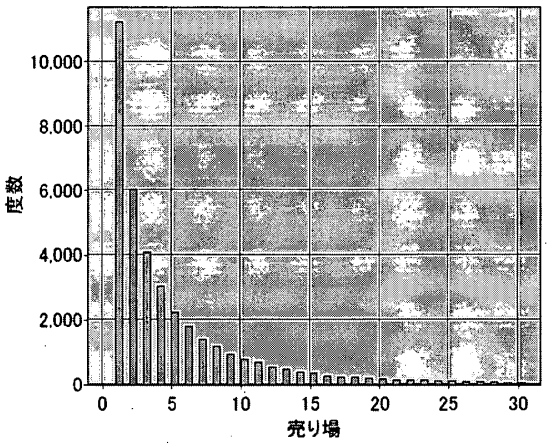

図4　売場数による顧客の分布

まず，1-SVMによる親近性指標の計算のため，前節と同様に最低30人以上の顧客が利用した売場を抽出し，これらの売場を3店以上利用したことのある顧客20,636人のデータを用いた．この結果，抽出された売場数は605，顧客一人当たりの平均利用売場数は約8.6店となった．1-SVMのカーネル関数は，購買ベクトルをノルム1に基準化した上で前節と同じ式(3.6)のRBFカーネルを用いた．パラメータは$2\sigma=605\times10$（商品数×10），またνの変化の間隔は$r=50$と設定した．

次に，1-SVMにより得られた親近性データを用い，双対尺度法により各売場間の関係や，顧客との関係を分析した．双対尺度法は，対応分析あるいは数量化III類とも呼ばれる分析法である．本研究の場合では，親近性データを継次カテゴリデータとして扱い分析をした．手法の詳細については参考文献[9]などを参照されたい．ここでは，百貨店の主要な売り上げを占めている，1階から3階までの婦人向けの商品を扱う71の売場（ブランド）のみを対象に分析した結果を報告する．この71の売場は，いずれも特定のブランドを扱うテナントである．また，売場の数が非常に少なくなったため，ほとんどの顧客は一つあるいは二つの売場でしか購買履歴がなく，一人当たりの平均もわずか2.31店である．したがって，このようなデータでは実際の購買回数や金額をそのまま用いた分析は困難で

あると考えられる．

双対尺度法では，親近性の様子を反映するよう顧客と売場双方に同一の尺度上の得点を与え数値化し，売場と売場，あるいは売場と顧客との関係を図示することができる．すなわち，親近性のパターンが似通った売場や顧客は近くに布置されるような図が得られる．得られた布置を示したものが，図5から図8である．なおこれらの図は，説明の都合上，売場名称の表示の有無が変えてあるだけで，いずれの図も同一の布置の様子を示したものである．また，図中の売場名称では，セール品を販売した際の売場名には先頭に「S」を付けて通常品の場合と区別してある．

図5には，軸の意味を説明する上で重要と思われる売場の名称を表示した．また，売場の位置している階も記号により区別して表示した．これを見ると，横軸の正方向に2階のヤング向けの「ヤングレディースブランド」のほとんどが，逆に負の方向には3階の年配向けの「レディースブランド」のほとんどが布置されている．このことから，横軸は年齢と解釈することができる．双対尺度法では，各顧客に対しても売場の布置と同一の尺度の得点が付与される．そこで，顧客を若者と年配に分けて売場の布置に重ねてプロットをしてみる．図6および図7はそれぞれ，1970年以降と1950年以前に生まれた顧客の布置を示した図である．若い年齢層の顧客は右側に，逆に年配の顧客は左側に布置されていることが分かる．このことからも，年齢により利用する売場が異なっていることが分かる．また，このことは提案する親近性が顧客の購買パターンを正しく反映できているためと考えることもできる．

一方，縦軸では，正の方向にルイ・ヴィトン等の高級ブランド品が，負の方向にセール品や安価なブランドが布置されているため，縦軸は高級・高価格あるいは逆にカジュアル度と考えられる．特に，図8には，各売場の通常時とセール時での布置の変化を示すための矢印を付けた．この図を見ると，2階のヤングレディースブランドでは，セールになると布置が下方向に

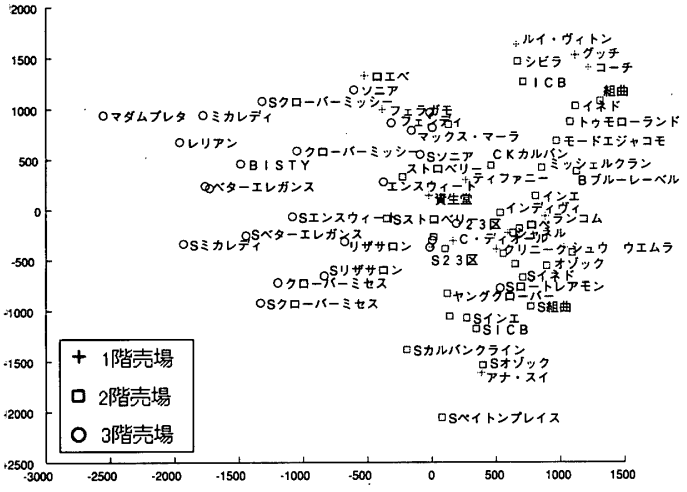

図5　売場の布置

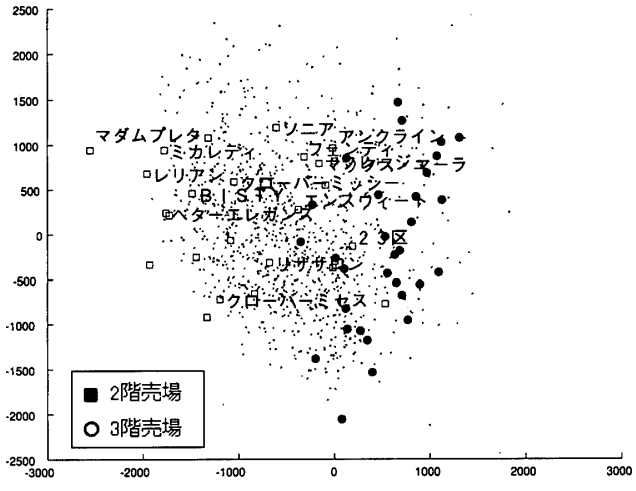

図7　1950年以前に生まれた顧客の布置

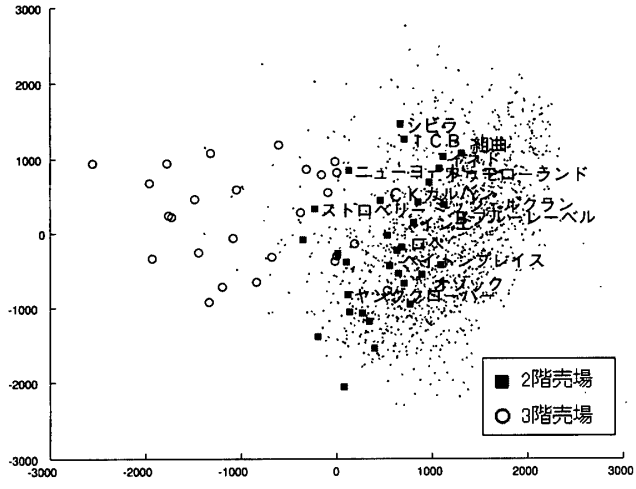

図6　1970年以降に生まれた顧客の布置

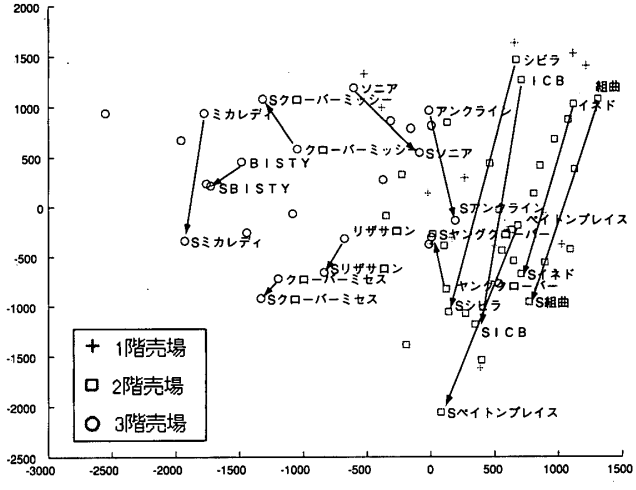

図8　セール時と通常時との比較

大きく移動している一方，3階のレディースブランドでは，下方向に移動するものの，その移動幅は2階のブランドと比較して小さいことが分かる．したがって，各ブランドはセール時になると，購買顧客層が変化するものの，その変化はヤング向けブランドと比べると年配向けのブランドはあまり大きくなく，固定客を摑んでいると思われる．

また，ヤングクローバー，クローバーミッシー，クローバーミセスなどの売場では，セール時の布置の変化が特に小さい．これらの売場は，いずれも大きな特別サイズを扱っており，したがって，セール時でも顧客層の変化は他の売場と比べ小さくなるためと考えられる．

5. おわりに

本研究は，百貨店におけるクレジットカード利用履歴データを分析する一手法として，未購買商品に対する買いやすさの指標となる「親近性」を算出する手法を提案した．数千種類以上の商品に対して，顧客一人当たりの購買商品がわずか数種類と極端に少ない本データの場合，未購買商品に対して一律に購買数を0として分析することは適切ではなく，また，標準的な多変量解析の手法を適用することも困難であると考えられる．このような場合，通常は商品あるいは顧客をいくつかの群に分類するなどして分析することが考えられる．しかし，ダイレクトメール（DM）発送のための顧客抽出といった応用も視野に入れるのであれば，本研究で用いた程度の商品分類や売場（テナント）別といった細かな分類での分析は，実用上非常に重要であると考えられる．節3.4で行った実験も，未購買商品のDM発送といったことを念頭において行った．このデータでの実験では，本手法の有効性を示すことができたと考えられる．

今後は，本手法に基づいたDM発送実験などを行い，さらなる有効性の検証が必要と考える．また，実験では購買顧客数の多い約1,700商品の購買履歴を用いたが，実務的にはどの商品の購買が親近性に寄与しているかといった分析を与える必要があろう．さらには履歴として用いる商品の選択方法や商品数による影響など，実証研究を通じて明らかにすることが必要である．いずれも今後の課題としたい．

参考文献

[1] C.-C. Chang, and C.-J. Lin: *LIBSVM : A Library for Support Vector Machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[2] T. Hofmann: *Latent Semantic Models for Collaborative Filtering*, ACM Transactions on Information Systems, 22, pp. 89-115, 2004.

[3] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl: *GroupLens : Applying Collaborative Filtering to Usenet News*, Communications of the ACM, 40, pp. 77-87, 1997.

[4] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl: *GroupLens : An Open Architecture for Collaborative Filtering of Netnews*, in Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, North Carolina, ACM, pp. 175-186, 1994.

[5] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson: *Estimating the Support of a High-dimensional Distribution*, Neural Computation, 13, pp. 1443-1471, 2001.

[6] U. Shardanand and P. Maes: *Social information filtering : Algorithms for Automating "Word of Mouth,"* in Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, vol. 1, pp. 210-217, 1995.

[7] J. Shawe-Taylor and N. Cristianini: *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.

[8] V. N. Vapnik: *The nature of statistical learning theory*, Statistics for Engineering and Information Science, Springer-Verlag, New York, 2000.

[9] 西里静彦：質的データの数量化，朝倉書店，1982.