日本オペレーションズ・リサーチ学会
2004年秋季研究発表会

百貨店における隠れた親近性の発掘

東京海上火災保険 (株) オウ ロ WANG, Lu
02702060 東京工業大学 * 吉原 亜弥 YOSHIHARA, Aya
01703730 東京工業大学 矢島 安敏 YAJIMA, Yasutoshi

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の分析を通じ, 顧客と商品の間に潜む隠れた親近性発掘手法の提案と, さらには算出した親近性を用いた顧客の購買行動の分析を行う.

研究に用いたデータは, カードでの購買履歴であることから, 高額商品の購買など顧客の行動の一部の記録であると考えられる. 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適当でない. そこで, 顧客の過去の購入商品の組合せを分析し, 未購買商品に対してもある種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.

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, 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, GroupLens project[3] や Ringo[5] などで, 盛んに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, これらの方法との比較検証を行った.

2. 1クラス SVM を用いた親近性の算出

顧客 i の商品の購入の様子を購入回数を要素とする N 次ベクトル $x_i \in \mathbb{R}^N$ で表す. ここで N は商品の種類である. このとき, ある商品 y に注目した場合の, 顧客と商品 y との親近性を考える. 商品 y を購入した顧客のベクトル x_i の集合を X_y とし, $X_y = \{x_1, x_2, \ldots, x_L\}$ と仮定する. 1クラスサポートベクターマシン (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することで, X_y に属すベクトルの特徴を抽出しようと考える. この際, 包含されない点が存在することも許容し, 包含されない点に対してはペナルティーを科して半径とともに最小化を考える. すなわち, 未知変数として球の半径の2乗を R, 球の中心を $a \in \mathbb{R}^N$ また, ペナルティーに対応した非負変数 ξ_i $(i = 1, 2, \ldots, L)$ を導入し, 以下の最適化問題:

$$(2.1)\quad \left|\begin{array}{ll} \text{最小化} & R + \frac{1}{\nu L}\sum_{i=1}^{L}\xi_i \\ \text{制約} & \|x_i - a\|^2 \le R + \xi_i,\ i = 1, 2, \ldots, L, \\ & \xi_i \ge 0,\ i = 1, 2, \ldots, L, \end{array}\right.$$

を考える. ただし, ν はペナルティー項 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.

集合 X_y の特徴をより良く反映させるため, 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 $\phi(x_1)$, $\phi(x_2), \ldots, \phi(x_L)$ に対して, 同様な最適化問題を考える. この場合には, (2.1) の双対問題:

$$(2.2)\quad \left|\begin{array}{ll} \text{最小化} & \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i\alpha_j\mathcal{K}(x_i, x_j) - \sum_{i}^{L}\alpha_i\mathcal{K}(x_i, x_i) \\ \text{制約} & \sum_{i=1}^{L}\alpha_i = 1, 0 \le \alpha_i \le \frac{1}{\nu L},\ i = 1, 2, \ldots, L \end{array}\right.$$

を最適化する. ただし, $\mathcal{K}(x_i, x_j)$ は空間 $\mathcal{F}$ の元 $\phi(x_i)$ と $\phi(x_j)$ の内積を表す関数である. 本研究では, RBF カーネル関数

$$(2.3)\quad \mathcal{K}(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/2\sigma\right)$$

を用いた. ここで, σ はあらかじめ定められる正のパラメータである.

ある ν に対して, 問題 (2.2) 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は,

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

として, $S(\nu) \equiv \{\, x \mid f(x) \le R^* \,\}$ と ϕ を使うことなく記述することができる.

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. ある ν に対して, 問題 (2.1) の最適解を ξ_i^*, 問題 (2.2) の最適解を α_i^* とする. ペナルティー項への重み ν に関して以下の定理 [4] が知られている.

定理 2.1 νL は球の外部となる点, すなわち $\xi_i^* > 0$ となる点の数の上限となる.

そこで, 適当な間隔で, $0 < \nu^1 < \nu^2 < \cdots < \nu^K < 1$ と K 個の ν を予め定め, それぞれに対応する領域 $S(\nu^1), S(\nu^2), \ldots, S(\nu^K)$ を計算する. その上で, $I(\nu, x)$ を $x \in S(\nu)$ ならば 1, それ以外では 0 を返す関数とし, 購買ベクトル $x \in \mathbb{R}^N$ に対して $R(x) = \sum_{k=1}^{K} I\left(\nu^k, x\right)$ を算出する. 多くの $S(\nu^k)$ に含まれる x ほど $R(x)$ は大きな値となることから, ある商品 y への親近度と考えることにする.

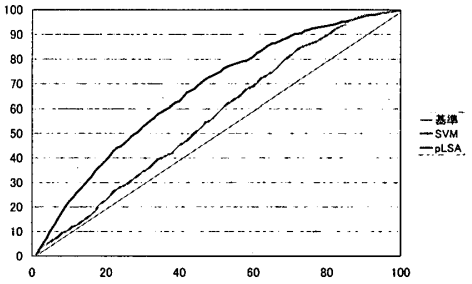

図 1: キャラクターシューズのリフト図

3. データを用いた検証と売場の分析

前節で述べた親近度の妥当性を百貨店のデータを用いて検証する. データは, 2 年間に 30 人以上の顧客に対して販売実績のある商品約 1600 を選び, かつ, これらの商品を 3 種類以上購入した顧客約 6 千人分を用いた.

実験は, まず顧客を 4 グループに等分し, 次のような 4-fold cross-validation を行った. 今, ある商品 y に注目した場合, 1 つのグループの顧客から商品 y に関する購買記録を削除し, テストデータを作成する. 残りの 3 グループの顧客の履歴をトレーニングデータとして用いて, テストデータに含まれる顧客の商品 y に対する親近度を算出する. 算出された親近度が実際の購買と一致しているかを, リフト図や再現性などで評価する. なお, 商品 y としては, 購入顧客数の多い 50 商品それぞれについて実験を行った. また, 提案手法との比較のため, 協調フィルタリングとして用いられている相関係数法 [2], および Hofmann の確率的潜在クラスモデル (probablistic Latent Semantic Analysis, pLSA)[1] を用いた. pLSA では潜在クラスの数を $K = 50, 100, 200, 300, 400$ と 5 通り設定して実験を行った. 例として, 図 1 には, 最も購入顧客数の多かった商品「キャラクターシューズ」でのリフト図を示した. 濃い黒線は 1-SVM の結果, 灰色の線が $K = 200$ とした場合の pLSA の結果である. この商品の場合には, 提案手法が既存手法を上回る予測性能を示している. 表 1 は, 実験した 50 商品の平均パフォーマンスを示したものである. 算出した親近度の上位 1%,2%,5%,10%および 20% の顧客を選んだ場合の再現率を, 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 を用いたものが全般的に他と比べ高い予測性能を示していることがわかる.

最後に, データを売場 (テナント) 別に集計したデータを用いて, 百貨店における売場の特徴の分析を試みる.

表 1: 50 商品の平均購買率の比較

	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

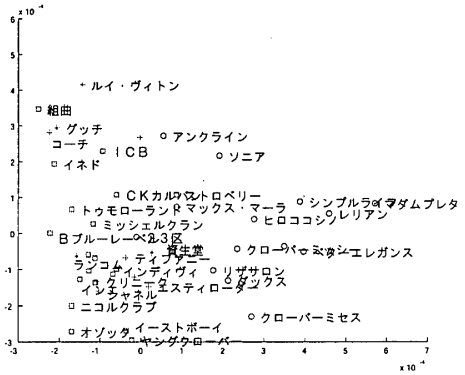

図 2: 対応分析の結果

集計の結果, 顧客 1 人当たり利用したことのある売場は平均 5.2 店となり, 全体の売場数約 600 と比べて極めて少ない. このようなデータの場合には, さまざまの分析をすることは困難であると思われる.

そこで, 前節で述べた親近度の指標を各売場に対して算出し, 利用していない売場に対しても親近度を付与する. 図 2 は, この親近度に基づき対応分析を行い, 売場を平面に布置した様子である. なお, 分析の詳細は発表の中で述べる予定である.

参考文献

[1] T. Hofmann, *Latent semantic models for collaborative filtering*, ACM Transactions on Information Systems, 22 (2004), pp. 89–115.

[2] 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 (1997), pp. 77–87.

[3] 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, 1994, ACM, pp. 175–186.

[4] 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 (2001), pp. 1443–1471.

[5] 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, 1995, pp. 210–217.