3J1-NFC1a-2

日常購買行動に関する大規模データを融合したベイジアンネットユーザモデル

Bayesian Network User Modeling by Large Scale Data Fusion in Retail Service

石垣司[*1] 竹中毅[*1] 本村陽一[*1]
Tsukasa Ishigaki Takeshi Takenaka Yoichi Motomura

[*1]産業技術総合研究所　サービス工学研究センター
Center for Service Research, National Institute of Advanced Industrial Science and Technology

The present paper describes a knowledge extraction method from Bayesian network user model based on the fusion of large scale dataset with respect to behaviors in retail service. The categories of customers and items are estimated from ID-POS data and questionnaire data by using a latent class model extended the PLSI method. We construct the Bayesian network model from the variables that extracted from these large scale dataset and demonstrated the knowledge extraction concerning the customer's behaviors.

1. はじめに

現在，サービスの科学的・工学的な研究が盛んになってきている [内藤 09]．現状のサービス産業の品質は熟練したサービス提供者の経験と勘への依存度が大きく，その生産性の低さが問題となっている．その中でも小売サービス業に注目すると，オーバーストア問題によって生じる，クーポン戦略や低価格化戦略などの過当競争による利益率の低下が問題視されている．この状態を継続することで小売サービス業者の価値と提供可能なサービスレベルが低下し，長期的には一般の生活者が享受できる価値も低下する．そのため，低価格化戦略に頼らないサービス品質の向上が課題となっている．商品の価格を下げずに来店率や顧客満足度を向上させることができる FSP(Frequent Shoppers Program) や CRM(Customer Relationship Management) のための仕組みが必要とされている．

そこでは，顧客個人やある特徴を持ったカテゴリに属する顧客に注目するマイクロマーケティングの観点が重要となってきている．マイクロマーケティングでは，顧客を売上額や居住地域などの何らかのカテゴリ毎にセグメント化する [中村 08]．各顧客カテゴリに適した販売促進や高付加価値化のための施策を実施することで，顧客満足度の向上と持続可能な需要創造を志向する．一方，商品管理に関してカテゴリマネージメント [Nielsen05] の観点も顧客満足度の向上にとって重要である。現状では各商品に対して属人的に大分類，中分類，小分類などの階層的なラベルを付与し，商品管理に利用している業者が多い．しかしながら，それらの商品分類の多くは流通業者の都合で設定されており，必ずしも顧客にとって意味のある分類とはなっていない．カテゴリマネージメントでは商品を顧客・消費者のニーズに基づいてセグメント化し、それに基づいたサービス設計を実施する。そこでは、商品カテゴリの定義が施策の成否に関して重要な要因となる。そのため適切な顧客と商品のカテゴリを自動的に発見し，顧客の要求と提供可能なサービスレベルのマッチングを図るためには，顧客のライフスタイルやパーソナリティを考慮した生活者視点での商品分類が必要となる

連絡先: 産業技術総合研究所サービス工学研究センター、〒 135-0064 東京都江東区青海 2-41-6、ishigaki-tsukasa@aist.go.jp

そこで本論では，日常購買行動に関する大規模データを融合的に利用した顧客行動の計算モデルを構築し，そのモデルからの知識発見を行う方法について述べる．ここでは以下の 2 つの手順で計算モデルの構築を実行する．

1 潜在的な顧客カテゴリと商品カテゴリを ID-POS データと顧客アンケートデータを統計モデル内で融合的に利用することで同時分類を行う

2 その分類結果と ID-POS とアンケートデータから得られる知見よりベイジアンネットワークモデルを構築し，その確率構造モデルから知識発見を行う

本論では流通量販店の 1 年間の ID-POS データと，その会員への 4000 人規模のアンケートデータを用いる．ここでは潜在的な顧客カテゴリと商品カテゴリを仮定し，顧客，顧客カテゴリ，商品カテゴリ，商品の関係を統計モデルで表現し，顧客パーソナリティやライフスタイルを考慮した顧客と商品の同時カテゴリ分類を実行する．その後，その分類結果や顧客や商品毎の特徴量を ID-POS データに付与し，各特徴量間の関係をベイジアンネットワーク [Pearl01, 本村 06] によりモデル化する．それにより，顧客の購買履歴データとアンケートデータに基づいた顧客行動の計算モデルを構築し，そのモデルから定量的な知識発見を行う．

2. 日常購買行動に関するデータ

2.1 大規模 ID-POS データとデータ抽出

本論では，兵庫県を事業エリアに約 150 店舗を展開する流通量販店で記録された 2008 年 10 月 1 日から 2009 年 9 月 30 日の期間における購買履歴データを利用する．この期間での全データのトランザクション数は約 6.7 億件（669511467 件）である．また，本論で対象とするデータはポイントカードの提示により顧客の購買記録とポイントカードの ID が関連付けて記録された ID-POS データである．本 ID-POS データにはポイントカード ID，店舗，利用日時，各商品に対して 1 対 1 対応の商品コード，購買数，購買価格の情報が含まれている．また，商品コードと関連付けることができる流通量販店独自の商品大分類・中分類・小分類・商品名のマスタが利用可能で，それぞれの分類数は 41，154，950，364397 である．ここでは MySQL を用いて ID-POS 用のデータベースを作成し，データ抽出環境を構築した．

2.2 顧客アンケートデータとその解析

2.2.1 アンケート内容

顧客のライフスタイルやパーソナリティを把握するため，顧客へのアンケート調査を実施した．2009 年 12 月に同流通量販店の会員約 17000 人に対しダイレクトメールを用いてアンケートを送付し，その内 3965 名から回答を得た．アンケートの項目数は全 35 問であり，デモグラフィック特性として年齢・性別・家族構成・家族人数・職業を設定した．ビッグ 5 法 [Goldberg90] を用いパーソナリティを把握するための質問項目を設定した．また，来店頻度，食生活，健康と食への意識，ダイエットへの意識，消費傾向などの質問項目も設定した．これらは消費者のライフスタイルや価値観に着目し，様々な先行研究を元に設計したものである [山本 01, 吉田 01]．回答は質問項目への当てはまりの強度順に“ 良く当てはまる ”から“ 全く当てはまらない ”までの 4 カテゴリの選択式である．4 つの選択肢に対し強度順に 4，3，2，1 の得点を与えている．

2.2.2 因子分析による顧客の消費・生活因子抽出

因子分析によりアンケートデータから顧客の消費・生活因子の抽出を行った．ここでは Kaiser の正規化を伴うバリマックス法を用いた．その結果，特徴的な 6 つの軸が抽出された．ここでは，それぞれの因子を以下のように特徴付け，6 つのライフスタイルに関わるカテゴリを抽出した.「こだわり消費派」: 高くても健康に良いものを選び，産地への関心，こだわりのブランドがある.「家庭生活充実派」: 料理が好きで食事も生活も充実している．気分も安定している.「アクティブ消費派」: 外向的で，新商品や話題の商品は試しに買ってみる．ただ無駄遣いは多い.「節約消費派」: チラシを見てお得な商品を買う．安ければ少々遠い店にも行く．高い商品は買わない.「堅実生活派」: 几帳面で家計簿をつけ，無駄遣いはしない．毎日の献立はスーパーに行く前に決める.「パパっと消費派」: スーパーでの買い物はできるだけ早くすませたい．お弁当を作ることがある．以上の 6 因子を顧客の消費・生活因子として顧客のライフスタイルカテゴリと定義する．これにより各顧客に対して，6 つの因子の得点を付与することができる．ここでは，アンケートの各質問項目に対して，因子の絶対値が最大値をもつライフスタイルカテゴリを，その質問項目が所属するライフスタイルカテゴリと設定した．そして，各ライフスタイルカテゴリに所属するアンケート回答の平均値を，各顧客の消費・生活因子得点として定義する．

3. 顧客－商品カテゴリの同時分類

3.1 潜在顧客・商品カテゴリのモデリング

ここでは X 人の顧客と Y 個の商品を対象とし，顧客 i と商品 j を表す変数をそれぞれ $x_i(i=1,\cdots,X)$ と $y_j(j=1,\cdots,Y)$ とする．また，潜在顧客カテゴリ数を U，潜在商品カテゴリ数を V とし潜在顧客カテゴリ k と潜在商品カテゴリ l を表す変数をそれぞれ $u_k(k=1,\cdots,U)$ と $v_l(j=1,\cdots,V)$ とする．ここでは顧客，商品，潜在カテゴリ間の関係を

$$p(x_i,y_j,u_k,v_l)=p(u_k)p(x_i|u_k)p(v_l|u_k)p(y_j|v_l) \tag{1}$$

としてモデル化する．このモデルは顧客 i の商品 j の購買数を N_{ij} とすると，その対数尤度は

$$\begin{aligned} L &= \sum_i^X \sum_j^Y N_{ij} \log p(x_i,y_j) \\ &= \sum_i^X \sum_j^Y N_{ij} \log \left\{ \sum_k^U \sum_l^V \right. \\ &\qquad \left. p(u_k)p(x_i|u_k)p(v_l|u_k)p(y_j|v_l) \right\} \end{aligned} \tag{2}$$

となる．

この潜在クラスモデルは EM アルゴリズムによる反復計算で対数尤度を最大化するパラメータを推定することが可能となる．このモデルにおいて推定すべきパラメータ数は U 個の $p(u)$，$X\times U$ 個の $p(x|u)$，$U\times V$ 個の $p(v|u)$，$Y\times V$ 個の $p(y|v)$ である．各パラメータに対して初期値を乱数で与えると，式 (2) の変形から潜在変数の条件付き確率は以下のように計算できる．

$$\begin{aligned} p(u_k,v_l|x_i,y_j) &= \frac{p(x_i,y_j,u_k,v_l)}{p(x_i,y_j)} \\ &= \frac{p(u_k)p(x_i|u_k)p(v_l|u_k)p(y_j|v_l)}{\sum_k^U \sum_l^V p(u_k)p(x_i|u_k)p(v_l|u_k)p(y_j|v_l)} \end{aligned} \tag{3}$$

また，ラグランジュの未定乗数法から各反復計算ステップの式 (4) の条件付き確率を最大化するパラメータは以下のように求めることができる．

$$p(x_i|u_k)=\frac{\sum_j^Y \sum_l^V N_{ij}p(u_k,v_l|x_i,y_j)}{\sum_j^X \sum_j^Y \sum_l^V N_{ij}p(u_k,v_l|x_i,y_j)} \tag{4}$$

$$p(v_l|u_k)=\frac{\sum_i^X \sum_j^Y N_{ij}p(u_k,v_l|x_i,y_j)}{\sum_j^X \sum_j^Y \sum_l^V N_{ij}p(u_k,v_l|x_i,y_j)} \tag{5}$$

$$p(y_j|v_l)=\frac{\sum_i^X \sum_k^U N_{ij}p(u_k,v_l|x_i,y_j)}{\sum_j^X \sum_j^Y \sum_k^U N_{ij}p(u_k,v_l|x_i,y_j)} \tag{6}$$

$$p(u_k)=\frac{\sum_j^X \sum_j^Y \sum_k^U N_{ij}p(u_k,v_l|x_i,y_j)}{\sum_j^X \sum_j^Y \sum_k^U \sum_l^V N_{ij}p(u_k,v_l|x_i,y_j)} \tag{7}$$

この反復を尤度が収束するまで実行することで各パラメータを推定することができる．

ここでは，顧客はアンケート回答者を対象とする．また，商品は 1 年間の売上個数の上位 1000 商品を対象とする．つまり，$X=3965$，$Y=1000$ である．また潜在クラス数は AIC 等の情報量規準により決定することが可能である．

3.2 アンケートに基づく制約条件の導入

3.1 節の EM アルゴリズムを実行することで提案モデルのパラメータ推定が可能となるが，本問題では 1 初期値依存性，2 計算時間についての困難が発生する．本モデルの尤度関数は単峰ではなく，乱数により設定する各パラメータの初期値に依存して尤度が局所最大化される．そのため，実行毎に異なった推定結果が与えられる．PLSA モデルでもこの問題が生じるが，本モデルでは 2 層の潜在変数を仮定しているため，さらに初期値に対して結果が敏感に反応する．そのため，複数回の試行を実施し，その中で最大の尤度をとる結果を採用するなどの処理が必要となる．また，推定すべきパラメータは数万パラメータと多い，EM アルゴリズムは 1 次収束のためある程度の反復回数が必要，などの理由から計算量は少なくない．加えて，2 つの潜在変数 u と v に対するクラス数 U と V を AIC により決定するためには 2 次元のグリッドを探索する必要がある．また，初期値依存性のため，その 2 次元グリッド探索を複数回

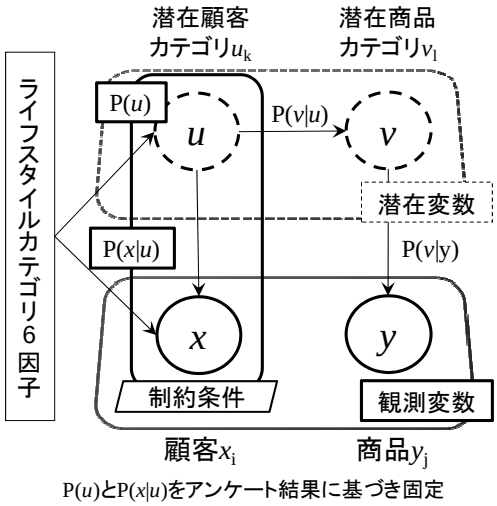

図 1: 制約条件を用いた提案モデル

実行する必要がある．この処理を全て実行することは実業務での使用の観点から現実的ではない．

そのため，アンケートの解析結果を制約条件としてモデル内に取り込む．ここでは次の仮定を制約条件とする．

仮定 顧客の購買行動はそのライフスタイルやパーソナリティに影響を受ける

ライフスタイルとパーソナリティに関連するアンケートの質問項目は全 20 問あり，各顧客 x_i に対して 6 因子に関連する質問項目の回答の平均値を $\boldsymbol{r}_q^i = \{r_1, \cdots, r_q \cdots, r_6\}$ とする．この平均得点を正規化した値を制約条件としてパラメータ $p(u_k|x_i)$ と $p(u_k)$ へ与え，パラメータを固定する．

$$p(u_k|x_i) = \frac{r_i^k}{\sum_k r_i^k}, \tag{8}$$

$$p(u_k) = \frac{\sum_i r_i^k}{\sum_i \sum_k r_i^k}, \tag{9}$$

$$p(x_i) = \frac{\sum_j N_{ij}}{\sum_i \sum_j N_{ij}} \tag{10}$$

そのモデルの概念図を図 1 に示す．また，ベイズの定理からパラメータ $p(x_i|u_k)$ が計算できるため，この制約条件を EM アルゴリズム内に代入することができる．これにより，初期値依存性と EM アルゴリズムの収束までの反復回数の低減が期待できる．また最適な潜在クラス数の探索空間を L に関してのみ行えばよい．

3.3 カテゴリ分類実験と結果

各分類実験は Mac OS X，プロセッサ 2 × 2.93GHz Quad-Core Intel Xeon，メモリ 32GB 1066Hz DDR3 の PC 内で Python2.6.4 の 64bit 版により動作させた．また，$U = 6$ とし $V = [2, 5, 10, 15, 20, 30, 50]$ について各 3 回異なる初期値で AIC の値を計算した．その結果として平均的に $V = 10$ が最適な値であると決定された．そこで $V = [6, 7, 8, 9, 10, 11, 12, 13, 14]$ について各 5 回異なる初期値で AIC の値を計算した結果，$V = 12$ が平均的に最適であると判断されたため，ここではその値を用いる．また，以下に示す分類結果は $U = 6, V = 12$ に対して異なる初期値で 30 回計算を実行し，最も尤度が高かったものを示している．

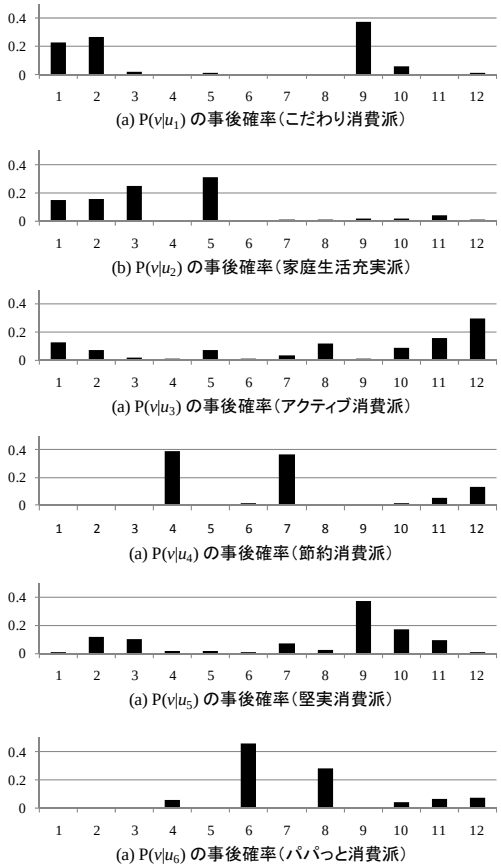

図 2: $p(v|u_k)$ の条件付き分布

図 2 に $p(v|u_k)$ 条件付き分布を示す．図中の横軸は各カテゴリ番号，縦軸は条件付き確率を示す．u_1，u_2，u_4，u_6 は数個の大きな条件付き確率をとるカテゴリが存在する．u_3，u_5 に関しては比較的複数のカテゴリで小さな条件付き確率値をとっていることがわかる．

ここで特定の特徴的な商品について分類の妥当性を探る．商品の中に野菜見切り品と果物見切り品が存在する．この 2 商品は節約消費派が最も高い条件付き確率を示している潜在商品カテゴリ 4 に属している．また，10 個詰めたまごは全 19 商品の内，1 年間の平均単価が高い 5 商品が，こだわり消費派が高い条件付き確率を示している潜在商品カテゴリ 1，2，9 に属している．また，平均単価が一番安い商品は節約消費派が高い条件付き確率を示している潜在商品カテゴリ 7 に属している．料理をすると回答している家庭生活充実派が高い条件付き確率を示している潜在商品カテゴリ 1，2，3，5 には調理済みの惣菜品がほとんど分類されていない，等の程度の妥当性を見ることができる．

制約条件の導入のパラメータ推定計算上の効果についてみる．以下の結果はそれぞれの条件において各試行を 20 回行った値を用いている．制約条件を用いない場合と用いた場合での推定結果の対数尤度の平均値と分散を表 1 に示す．制約条件ありの方が対数尤度の値が低くなっている．しかしながら，その分散の値は約 26 分の 1 にまで減少している．そのため，推定結果の初期値依存性は低減していると考えられる．また，制約条件を用いない場合と用いた場合での EM アルゴリズム収束までの平均反復回数を表 2 に示す．ここでは EM アルゴリズムの各反復ステップで対数尤度の向上が対数尤度の値の $10^{-4}\%$ を下回った時を収束と判定した．その結果，制約条件な

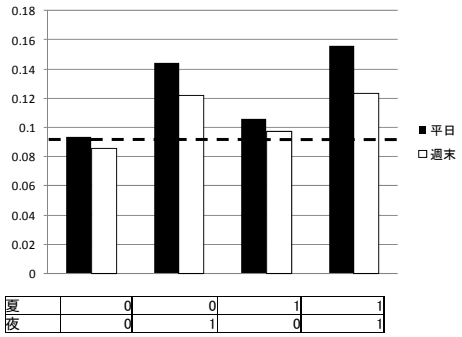

図 3: 交互作用を含む確率推論の一例

しでは EM アルゴリズムの収束までに約 500 反復必要であったが，制約条件ありでは約 100 回の反復で収束した．

表 1: 推定結果のばらつき

	制約なし	制約あり
対数尤度の平均	-61231738	-61426404
分散	54123074	2070215

表 2: 計算収束までの平均反復回数

	制約なし	制約あり
平均反復回数	512	103

4. 確率構造のモデリングと知識発見

4.1 確率構造モデルの構築

ここでは対象となっている約 420 万件のトランザクションデータに対してベイジアンネットワークモデルを構築した．各商品に対して購入した顧客に対し，合計購入数，合計金額，購入平均単価，特定保健用食品（トクホ）購入回数，プライベートブランド購入回数，国産品購入回数，健康食品購入回数，お手軽品購入回数，高級品購入回数，ダイエット的商品購入回数，お買い得商品購入回数に関しては対象店舗の顧客に対して ABC 分析を行い，その結果をラベルとして付与した．また，6 種類のライフスタイルカテゴリ属性と 12 種類の潜在商品カテゴリ属性のラベル，20 問のアンケート項目，デモグラフィック属性（年齢・性別・家族構成など），状況変数（購買の時間帯，平日か休日，季節など）も付与した．そのデータに対して Greedy search により AIC の意味で最適になるような確率構造の探索を行った．

4.2 確率構造モデルからの知識発見

構築したベイジアンネットワークからの知識発見を行った．その一例を図 3 に示す．ここではお手軽品購入パターンの状況依存性についての確率推論の結果である．図中の破線はお手軽品の商品購買履歴の事前確率である．お手軽品は平日の夜に購買率が上がり，かつ夏の方がさらに購買率が上がることが読み取れる．

このような定量的な知見がベイジアンネットワークモデルから複数抽出することができる．その他の顧客ライフスタイルカテゴリや潜在商品カテゴリに対して抽出された知識や知見については発表時に報告する．

5. 考察

本報告で作成した確率構造モデルを用いて，状況依存性を考慮した顧客の購買行動に関する定量的な確率推論を実行することが可能となっている．このモデルから顧客行動の予測を行い，その行動に適合したサービス提供を実施できる可能性がある．

パラメータ推定に関して，PLSA による最尤推定は過学習の問題が生じることが知られている．そのため，提案モデルに関しても同様の問題が生じる可能性があるため，TEM[Hofmann01] などのアニーリングを利用したパラメータ推定法の実験も行う必要がある．

6. むすび

本報告では，流通量販店の ID-POS データとアンケートデータを融合的に用いた顧客と商品の同時カテゴリ分類法とそのベイジアンネットワークによる計算モデル化からの知識発見について述べた．今後は LDA などの知見を取り入れたパラメトリックモデルへの拡張が課題である．また，実店舗への介入による得られた知見の効果測定も一つの課題としてあげられる．

謝辞

本報告で用いた ID-POS データと顧客アンケートデータは生活協同組合コープこうべから提供を受けた．ここに感謝の意を表する．また，本研究は経済産業省サービス研究センター基盤整備事業の委託研究費を受けている．

参考文献

[Goldberg90] L.R. Goldberg, “An alternative ”description of personality”: The Big-Five factor structure”, Journal of Personality and Social Psychology. Vol.59, No.6, pp.1216-1229, 1990.

[Hofmann01] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis”, Machine Learning, Vol.42, No.1-2, pp.177-196, 2001

[本村 06] 本村陽一，岩崎弘利：ベイジアンネットワーク技術，東京電機大学出版局, 2006

[内藤 09] 内藤耕（編），“サービス工学入門”，東京大学出版会，2009

[中村 08] 中村博（編），“マーケット・セグメンテーション”，白桃書房，2008

[Nielsen05] A.C. Nielsen, A. Heller, “Consumer-Centric Category Management: How to Increase Profits by Managing Categories Based on Consumer Needs”, Wiley, 2005

[Pearl01] J. Pearl, “Causality: Models, Reasoning and Inference (2nd ed.)”, Cambridge University Press, 2009

[山本 01] 山本真理子（編）心理測定尺度集〈1〉—人間の内面を探る“ 自己・個人内過程 ”，サイエンス社，2001

[吉田 01] 吉田富二雄（編），心理測定尺度集〈2〉人間と社会のつながりをとらえる“ 対人関係・価値観 ”，サイエンス社，2001