4G1-2

マイクロクラスタリングを用いた概念化とモデルの構築

Prediction model using Micro-clustering

中原 孝信 *1 Takanobu NAKAHARA
宇野 毅明 *2 Takeaki UNO
羽室 行信 *4 Yukinobu HAMURO

*1関西大学 データマイニング応用研究センター
Data Mining Applied Recerch Center, Kansai University

*2国立情報学研究所 情報学プリンシプル研究系
Principles of Informatics Research Division, National Institute of Informatics

*3*4関西学院大学 経営戦略研究科
Institute of Business and Accounting, Kwansei Gakuin University

In this study, we propose the method of using micro-clustering for prediction model to using POS data. An algorithm for graph clustering, micro-clustering, it is possible to extract a density group structure. Therefore, micro-cluster is grouping closely related items to be purchased in common. To build a classification model of health-conscious we use cluster for the explanatory variables. By utilizing the micro-cluster, it is shown that the classification accuracy and the validity of the interpretation is improved.

1. はじめに

近年ではデータの収集コストが安価になったことから，様々なデータを容易に取得・収集できるようになってきた．これまでに小売店で蓄積された POS データを対象にした研究は，ブランド選択に関する研究 [Guadagni 83] や，販売促進の効果に関する研究 [Gupta 88]，そしてデータマイニングを用いた購買行動に関する研究 [Hamuro 98],[Nakahara 05] など，数多くの研究が行われてきた．しかし，これらの研究は，同一店舗内の売上げデータを対象にしており，ブランド比較や商品比較は可能であるが，他店舗で行われた購買行動を把握することはできない．しかし，近年では携帯端末の普及により消費者がモニターとして，日々自分の購買した商品をスキャンすることでデータを蓄積するサービスが行われている．この方法で蓄積されたデータには，共通のモニター ID で複数の店舗を利用した情報が含まれているため，分析者は他店舗のデータを横断的に利用することが可能である．

本研究は，この店舗横断的なデータであるスキャンパネルデータを用いて，顧客の購買行動に関する特徴をマイクロクラスタリングにより概念化し，概念を利用した分類モデルを構築する．購買行動の特徴は，顧客が店舗や商品を選択する際に想起する店やブランドをマインドとして捉えることを目的にしており，正例と負例でマインドの違いを明らかにする．そして，意味解釈の妥当性と分類精度の向上という 2 つの観点から評価を行う．スキャンパネルデータは，経営科学系研究部会連合協議会が主催する平成 25 年度データ解析コンペティションで提供していただいた．

2. 手法

分類モデルを構築するにあたって，本研究では目的変数として健康志向の顧客群を正例とし，その他の顧客郡を負例と定義した．健康志向かどうかは，健康に関する食事関連の 4 項目のアンケート結果をスコアリングした．具体的には,「1 食でより多くの食材が摂れるように料理をする」,「1 汁 3 菜を意識して料理を作る」,「1 食あたりのカロリーや塩分・脂質・糖分・食物繊維などを意識しながら食事を作る」,「自分の健康・体調管理よりも，家族の健康・体調管理を意識して料理をする」. という 4 つの質問を対象にして，5 件法の回答から平均値を計算し，平均値以上であれば「健康志向」 そうでなければ「非健康志向」として定義した．

そして，説明変数として顧客マインドを設定し，健康志向と顧客マインドの関係をモデル化する．ここで，顧客マインドとしては店舗と商品の関係を設定することにした．健康志向の顧客群とそうでない顧客群が持つ商品に対する店舗のイメージを明示化しようということである．具体的には，説明変数として，顧客の店舗での商品の購入の有無を表した 2 値変数を設定する．例えば,「ダイエー」で「牛乳」を買った事がある顧客は,「ダイエー_牛乳」という変数の値が 1 となり，購入のない顧客の値は 0 となる．

店舗数を n，商品数を m とすると，説明変数は $n \times m$ 次元ベクトルとなる．しかし，今回扱うデータでは，カバーするサンプル数があまりにも少ない変数が多数を占めることになり，結果として精度の高いモデルが得られないという問題がでてくる．そこで，前処理として変数をクラスタリングすることを考えるが，その方法に本研究の特徴がある．

まず，顧客をトランザクション，そして店舗商品のペア変数をアイテムと考え，アイテムの共起頻度を計算し，変数間の類似度グラフを構成する．類似度としては様々なものを定義できるが，最終的に分類モデルを構築するという目的から，顕在パーン (emerging patterns) における増加率 (GR:Growth Ratio) を用いる．正例，負例のトランザクション集合をそれぞれ D_p，D_n とすると，2 つのアイテム a, b の負例に対する正例の増加率は以下の式で定義される．

$$GR_{D_d \to D_p}(a,b) = \frac{|Occ_p(a,b)|/|D_p|}{|Occ_n(a,b)|/|D_n|} \quad (1)$$

ここで $Occ_p(a,b)$ はアイテム a, b が共起するトランザクショ

連絡先: 中原孝信，関西大学，吹田市山手町 3-3-35，
TEL:06-6368-1121，nakapara@gmail.com

ン集合を表す．この式は，負例の共起確率に対する正例での共起確率の比であり，1.0 より大きければ，アイテム a, b は正例に特徴的な共起パターン (顕在パターン) であると言える．そして，増加率が 1.0 より大きい変数ペアを全列挙し，それらの変数間に枝を張る．

このように得られた類似度グラフでは，お互いに類似した変数群の枝密度は濃くなり，逆に類似していない変数群の枝密度は薄くなる．そこで，類似度グラフから，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，正例に特徴的な変数クラスタを構成することができる．同様の考えは負例に対しても容易に当てはめることができ，負例に特徴的な類似度グラフを構成しておく．

一般グラフのクラスタリングについては，ニューマンクラスタリング，グラフ分割，極大クリーク列挙など，これまでも様々な手法が提案されてきたが，どの手法も問題点を抱えており，決定打になっていないというのが現状である．

例えば，極大クリーク列挙では，現実データにおいては多くの場合，非常に多数の類似した極大クリークが列挙されてしまうという問題がある．列挙された極大クリークの類似関係を用いて，極大クリークを更にクラスタリングするという方法も提案されているが，列挙される極大クリークの数によっては計算量が問題となる．このような問題の多くは，そもそも対象とするグラフにノイズが含まれるために起こる問題とも考えられる．

そこで，最近著者らは，対象とするグラフをクリーニングする「グラフ研磨」手法を提案している [Uno 2014]．これは，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しておこうというものである．直感的には，枝密度の濃い部分グラフはより濃く，薄い部分グラフはより薄くするというものである．このような方法を適用することで，列挙されるクリークの数が劇的に少なくなることがわかっている．

研磨の方法は至ってシンプルで，全ての頂点ペアについて，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従って，新たなグラフを再構成する．全頂点ペアの計算は節点数の 2 乗の計算量が必要となるが，より効率的なアルゴリズムが存在する [Uno 2014]．

類似度としては様々な定義を用いることができるが，本研究では Jaccard 係数を用いる．グラフ上での 2 つの節点 u, v の Jaccard 係数 $sim(u, v)$ は，以下のとおり定義される．

$$sim(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{2}$$

ここで $N(u)$ は節点 u に直接接続のある節点集合を表している．そしてユーザが与えた最小類似度以上の類似度 δ を持つ変数ペアに枝を張ることでグラフを再構成していく．この類似度を用いてグラフを再構成すると，大雑把に言えば，共通節点の多い節点間に枝が張られ，少ない接点間の枝は切断される．これは，SNS における友達紹介のアルゴリズム (すなわち共通友達の多い友達は友達である可能性が高い) と同様なもので，グラフ構造のプリミティブな変化予測 (リンク予測) を行っているとも解釈できる．

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列挙された極大クリークを我々はマイクロクラスタと呼んでいる．

以上により得られたマイクロクラスタを説明変数として分類モデルを構築する．マイクロクラスタとしての変数は，マイクロクラスタを構成するアイテム数の 30%以上のアイテム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 値の変数である．

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)，p 個の説明変数 (マイクロクラスタ) ベクトルを $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ とすると，ロジスティック回帰モデルは式 (3) で表される．

$$\Pr(y = 1|\mathbf{x}) = f\left(\boldsymbol{\beta}^\top \mathbf{x} + \beta_0\right) \tag{3}$$

$f(\cdot)$ 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する．

回帰問題において $\boldsymbol{\beta}$ の推定には最小 2 乗法を利用するのが一般的であるが，説明変数の数 p がサンプル数に比べて多いとき，説明変数間の共線性が問題となり，異なる推定法が必要となる．この問題に対して様々な推定法が提案されてきたが，最小 2 乗法に $\boldsymbol{\beta}$ に対する罰則を与えた上で最小化問題 $\mathrm{argmin}\{||y - \boldsymbol{\beta}^\top \mathbf{x}||_2^2 + J(\boldsymbol{\beta})\}$($J(\boldsymbol{\beta})$ は罰則項) を解く罰則付き回帰が有効であることがわかってきた．その中でも $J(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1$ とした lasso，および $J(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_2^2$ とした ridge 回帰がよく利用される．ここで，$||\boldsymbol{\beta}||_q$ は q-ノルムで $||\boldsymbol{\beta}||_q = (\sum_{i=1}^p \beta_i^q)^{1/q}$ である．$\lambda \in [0, \infty)$ は正の定数であり，lasso においては $\boldsymbol{\beta}$ をどの程度疎に選択するかのトレードオフパラメータである．つまり λ が大きい場合には，$\boldsymbol{\beta}$ の多くの値が 0 となる．逆に λ が 0 の場合は通常の最小 2 乗法となる．ridge 回帰においては λ を，大きく設定しても回帰係数が 0 と推定されることはないが，推定値が全体的に小さく推定されることになる．

ridge 回帰は共線性への対処法として用いられるが，変数選択としては機能しない．一方で lasso は λ の値によっては多くの回帰係数が 0 となることから変数選択の有効な手法として注目されている．しかしながら一方で共線性のある変数が選ばれにくいといった問題も指摘される．そこで，両者の罰則を結合し，$J(\boldsymbol{\beta}) = \lambda_1|\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||_2^2$ とした elastic net がある [Zou 2005]．本研究では，ridge 回帰も lasso でも思ったようなモデル精度が得られなかったために elastic net を使うことにした [*1]．

3. 計算実験

本研究で利用するスキャンパネルデータは，2012 年 1 年間のデータで約 6500 人のモニターによる購買情報が含まれたデータであり，上述の方法で，健康志向と非健康志向を定義し目的変数として利用した．

3.1 マイクロクラスタの生成

マイクロクラスタリングを列挙する際に利用した顕在パターンの閾値は $GR = 1.0$ とした．また，グラフ研磨は，δ の値によって様々なグラフ構造が得られるため，最適な δ を一意に定めることは困難である．そこで，本研究では δ を 0.1 から 0.9 までの 0.1 刻みで動かし，各 δ でクラスタを列挙した．そし

*1 統計解析ツール R のパッケージ glmnet を用いている．そこでは，λ_1, λ_2 の調整を $(1-\alpha)/2||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1 (0 \leq \alpha \leq 1)$ のように調整パラメータ α によって実現している．α を 0 に近づければ ridge 回帰の罰則が強くなり，逆に 1.0 に近づければ lasso の罰則が優先される．α は試行錯誤の実験から 0.001 とした，

表 1: マイクロクラスタに関する各種統計量

δ	クラスタ数	節点平均	節点数	枝数	枝密度	重複度
0.1	25	12.760	307	14631	0.300	1.039
0.2	25	12.760	307	14631	0.300	1.039
0.3	25	12.760	307	14631	0.300	1.039
0.4	26	10.462	238	6894	0.141	1.143
0.5	26	8.615	217	3019	0.063	1.032
0.6	41	4.927	200	1230	0.026	1.010
0.7	42	3.857	156	349	0.007	1.038
0.8	38	3.289	125	248	0.005	1.000
0.9	17	2.176	37	23	0.000	1.000
ORG*2	405	27.916	313	6845	0.140	36.121

て列挙されたクラスタから完全に一致するクラスタを一意にすることで，多様なクラスタを生成した．表 1 は，δ を変えたときに得られたクラスタに関する統計量を示している．この結果は店と細分類のペアをアイテムとして扱った健康志向のケースであり，ORG は研磨を行わずに極大クリークを列挙した場合の各種統計量を示している．項目名「クラスタ数」は得られた極大クリークの数,「節点平均」は 1 つのクラスタに属する平均節点数,「節点数」「枝数」は研磨後のグラフにおいて，少なくとも 1 つの節点に接続のある節点数，及び枝の数である．「枝密度」は，完全グラフの枝数に対する実枝数の割合を表している．そして「重複度」は，1 つの節点が属するクラスタ数の平均をそれぞれ表している．

δ が 0.3 より小さい場合は，研磨の過程においての構造はそれぞれ異なっていることを確認したが，最終的に収束した構造は同じであった．また δ が 0.4 から 0.7 まではクラスタ数が増加している．これは，δ を増加させると，間接的な共起関係の弱い枝は削除されるため，小さいサイズのクラスタが生成されるからである．また，それ以上の δ になるとクラスタ数が減っているが，これは接続がなくなり，2 節点以上のクラスタではなく，単一の節点が増えていることが理由である．節点数は，本来 δ によらず一定であるが，単一の節点からなるクラスタは除外しているため ORG に比べて節点数が減少している．重複度は，ORG の約 36 から 1 へと減少しており，研磨の過程で，重複の大きい 2 つのクリークが併合され，逆に重複の小さい 2 つのクリークが分離するために重複率が下がっており，同様にクラスタ数 (極大クリーク数) も大幅に少なくなっている．一般に極大クリーク数は巨大になることが多く，クラスタリングにおける大きな問題となっているが，グラフ研磨によりその数が効果的に減少していることが分かる．

表 2 は，マイクロクラスタの例を示している．クラスタの要素を見ると，同じ店舗から構成されたクラスタや小売系のお店が集まったクラスタなど，顧客の購買関係を反映させたクラスタができており，意味解釈が比較的容易である程度顧客のマインドを表現したものとして解釈できる．計算実験において最終的に約 1,400 個のマイクロクラスタが得られた．

3.2 健康志向予測モデルの構築

得られた 1400 個のマイクロクラスタを説明変数に利用して，2 項ロジスティク回帰によって「健康志向」と「非健康志向」の予測を実施した．提案モデルの予測精度を評価するため，10-fold cross-validation で評価した結果，正答率は 70.69%であった．この正答率は，事前確率による予測 (0 と 1 の出現回数が多い方を予測結果とする方法) と比較して有意であった (有意確率=2.2e-16)．また，データ研磨を実施せずに類似度グラフから極大クリークを列挙して同様に 2 項ロジスティク回帰を実施した場合には，予測精度は 65.36%であった．データ研磨を行うことで正答率を約 5%改善することができた．学習後の重みベクトルの非ゼロ要素は元のクラスタ数 1400 個から 160 個まで削減した．

表 2: クラスタの抜粋

セブン＆ｉ系クラスタ
{ セブン＆ｉ系-その他水物, セブン＆ｉ系-蒲鉾, セブン＆ｉ系-その他畜産, セブン＆ｉ系-コンニャク, セブン＆ｉ系-冷凍農産素材, セブン＆ｉ系-キャンディ・キャラメル, セブン＆ｉ系-炭酸フレーバー, セブン＆ｉ系-油揚げ, セブン＆ｉ系-その他加工水産, セブン＆ｉ系-インスタントカレー }
ダイエー系クラスタ
{ ダイエー系-半生菓子, ダイエー系-生麺・ゆで麺, ダイエー系-食パン, ダイエー系-その他畜産, ダイエー系-牛乳, ダイエー系-菓子パン, ダイエー系-ヨーグルト, ダイエー系-豆腐 }
小売混合クラスタ
{ ダイエー系-加工食品, 西友系-加工食品, その他一般小売店-生鮮食品, セブン＆ｉ系-家庭用品, マツモトキヨシ-化粧品, その他 100 円ショップ（ダイソーなど）-化粧品 }

elastic net で選ばれた変数は共線性の高い変数同士も選ばれている．そこで，それらの中から意味を解釈するために 5%有意であった変数を, 代表的な変数として選択した．それらの変数を表 3 に示す．非健康志向に寄与する変数は，係数がマイナスになっている変数である．変数はクラスタになっており，横軸で区切られた範囲が 1 つのクラスタを示している．例えば，非健康志向を説明する変数として，表の上から「スーパーでコーラ」「スーパーでインスタント袋麺」「イオン系でスナック」を購入している．不健康の代名詞となるような食品群が出現している．また，NEWDAYS-食品から始まるクラスタは 10 個のアイテムから構成されているが，その多くが「コンビニ・自販機で規制食品」を買っており，さらに「その他スーパー」が集まったクラスタは，加工食品・菓子，ツマミ系といった商品を購入している，このように意味解釈が可能な非健康志向の典型的な購買を示している．

一方で健康志向の特徴としては，ほぼ全てのクラスタに出現している店舗は,「スーパー」または「ドラッグストア」で「セイジョー」「ダイエー」「生協」「マツモトキヨシ」「サンドラッグ」「ケーヨー」「ヤオコー」などを贔屓にしている．また,

*2 ORG は研磨後ではなく類似度グラフに対する各種統計量である．

ローソンではスナックではなく，生菓子を購入しているなど，健康志向の特徴を表すクラスタが出現していた．

表 3: 有意なマイクロクラスタ

その他スーパー-コーラ	-0.194018
その他スーパー-インスタント袋麺	-0.198015
イオン系-スナック	-0.243502
NEWDAYS-食品	-0.260298
その他屋外の自販機-食品	-0.260298
サンクス-食品	-0.260298
セブンイレブン-日用品	-0.260298
デイリーヤマザキ-食品	-0.260298
ミニストップ-食品	-0.260298
住宅街の道路沿いの自販機-食品	-0.260298
家電量販店-文化用品	-0.260298
楽天市場-文化用品	-0.260298
職場（オフィス）の自販機-食品	-0.260298
ファミリーマート-飲料・酒類	-0.26732
ローソン-飲料・酒類	-0.26732
その他スーパー-乳製品	-0.313709
その他スーパー-チョコレート	-0.407512
その他スーパー-チーズ	-0.407512
その他スーパー-和惣菜	-0.407512
その他スーパー-漬物	-0.407512
その他スーパー-畜肉ソーセージ	-0.407512
その他スーパー-米菓	-0.407512
その他スーパー-納豆	-0.407512
ローソン-生菓子	0.169906
その他の小型食品スーパー-調理品	0.179944
その他 100 円ショップ（ダイソーなど）-菓　子	0.181586
その他スーパー-スープ	0.181586
その他スーパー-ホームメーキング材料	0.181586
その他スーパー-ラッピングフィルム	0.194462
その他スーパー-水	0.194462
その他一般小売店-その他農産	0.194462
その他スーパー-ビール	0.218102
その他スーパー-マカロニ	0.218102
その他一般小売店-珍　味	0.224946
マツモトキヨシ-住居用洗剤類	0.224946
サンドラッグ-衣料用洗剤類	0.226391
その他スーパー-油揚げ	0.245447
その他スーパー-チーズ	0.257239
その他スーパー-畜肉ソーセージ	0.262252
その他スーパー-わかめ	0.286723
その他スーパー-その他食品	0.290387
その他ディスカウントストア-家庭用品	0.290387
ケーヨー D2-日用雑貨	0.290387
セイジョー-日用雑貨	0.290387
セイムス（SEIMS）-日用雑貨	0.290387
ヤオコー-加工食品	0.290387
サミット-加工食品	0.295597
セブン＆ｉ系-珍　味	0.317744
クリエイト-菓　子	0.357944
その他の小型食品スーパー-日用品	0.378628
その他ホームセンター-食品	0.378628
セイジョー-日用品	0.378628
ダイエー系-日用品	0.378628
L-楽天市場-日用品	0.378628
生協の個人宅配-食品	0.378628
その他スーパー-漬物・佃煮	0.387985
その他スーパー-乳製品	0.520824
その他スーパー-加工肉類	0.520824

4. おわりに

本研究では，グラフ研磨を用いたマイクロクラスタリングを用いて，得られたクラスタを予測問題に実施した．そして，消費者の店舗別の購買行動を考慮することで，健康志向と非健康志向で店に対するマインドの違いを明らかにした．また，消費者のマインドを概念化したクラスタを用いることで，ある程度高い予測精度を得ることができ，グラフ研磨を利用したマイクロクラスタリングが予測問題に有効であることを示した．

今後の課題は，他の志向に対するマインドの概念化とモデル化を実施することで，健康志向以外の志向に対しての有効性を確認し，また，他のクラスタリング手法との比較などを行っていく必要がある．また，獲得した消費者のマインドを用いたマーケティング施策を展開することを目標にさらなる研究を進めていきたいと考えている．

謝辞

本研究の一部は，科学技術振興機構 CREST，及び ERATO 湊離散構造処理系プロジェクト，文部科学省の科研費若手研究 (B) 4730375，科研費基盤研究 (B) 25285127 の研究助成を受けている．

参考文献

[Guadagni 83] Guadagni, P. M. and Little, J. D. C., “A logit [odel of brand choice, calibrated on scanner data”, *Marketing Science*, 2, 1983, pp. 203–238.

[Gupta 88] Gupta, S., “Impact of sales promotions on when, what, and how much to buy”, *Journal of Marketing Research*, 25, 1988, pp. 324–355.

[Hamuro 98] Hamuro, Y., Katoh, N., Matsuda, Y. and Yada, K., “Mining pharmacy data helps to make profits”, *Data Mining and Knowledge Discovery*, 2, 1998, pp. 391–398.

[Nakahara 05] 中原孝信, 森田裕之,「百貨店のクレジット購買データを用いた関連購買による顧客特徴分析」, オペレーションズ・リサーチ, Vol50, No.7, 2005, pp. 488–494.

[Uno 2014] 宇野毅明, 中原孝信, 前川浩基, 羽室行信「データ研磨によるクリーク列挙クラスタリング」情報処理学会アルゴリズム研究会報告書, 2014-AL-146(2), pp. 1-8, 2014.

[Zou 2005] Zou, H. and Hastie, T., “Regularization and variable selection via the elastic net”, *Journal of the Royal Statistical Society B*, 67, 2005, pp. 301–320.