3E3-1

広告の売り上げパターンへの影響に関するデータマイニング

Data mining on effect of advertisement to sales patterns

城戸 健太郎[*1] 鷲尾 隆[*1] 矢田 勝俊[*2] 元田 浩[*1]
Kentarou Kido Takashi Washio Katsutosi Yada Hiroshi Motoda

[*1]大阪大学産業科学研究所知能システム科学研究部門高次推論方式研究分野
Department of Advanced Reasoning , Division of Intelligent Systems Science , Osaka University. The Institute of Scientific and Industrial Research

[*2]関西大学商学部
Faculty of Commerce , Kansai University

Basket analysis is a method to search for co-occurrence patterns of items. Recently, QFIMiner has been proposed where co-occurrence patterns consisting of not only symbolic items but also numeric items are searched for. However, the temporal changes of the frequency of co-occurrence patterns cannot be detected by the current approaches. Because the temporal changes of the co-occurrence patterns are often observed in our real society, some approach to detect the changes in an accurate manner meets practical needs of some applications. In this paper, we propose a method to detect the statistically significant differences of the support of co-occurrence patterns across two time points. Moreover, the performance evaluation of the proposed method was performed by using actual data in marketing field.

1. はじめに

データマイニングにおいて，主要な解析フレームワークの１つに事象共起パターン分析がある．事象共起パターン分析の活用範囲は広く, 非常に膨大なデータから多頻度の共起パターンを探索し, そこから得られた情報を活用することで, 商品開発や経営戦略に役立たせることが可能となる．事象共起分析の代表的な手法としてバスケット分析 [Berry,Linoff99] が挙げられるが, バスケット分析では数値アイテムを含むデータを解析できないという問題点がある．現実世界では, 数値を含むデータが多く, バスケット分析を用いて解析を行うことができるデータの種類は限られたものになる．この問題点を解決すべく, 我々は定量的相関規則導出手法（QFIMiner）[光永 05] と呼ばれる手法を開発した．これを用いることで, 数値アイテムを含むデータに対し, 定量的事象共起パターン分析が可能となる．数値の直接的な扱いが可能になることで, データから得られる情報は飛躍的に拡大する．

ここで, 事象共起パターンや定量的事象共起パターンの分析によって非常に有用な情報を得ることができる可能性がある一方で, データの性質が時間に伴い変化していき, 単に解析を繰り返すだけでは, 有用な情報を得られない場合がある．現実のデータでは, 事象共起パターンや定量的事象共起パターンが時間変化することが多いが, 従来の手法ではそれら時間的変化を的確に捉えることが困難である．

そこで, 本研究では, 数値を含まないデータ, ないしはデータ中の数値を無視する場合には, バスケット分析を適用し，２つの時期にまたがってマイニングした多頻度アイテム集合のサポートの統計的有意差検定 [稲垣等 92][篠崎 95] による時間変化分析を行う方法を開発する．また, 数値情報も対象とする場合には，２つの時期にまたがってマイニングした定量的多頻度アイテム集合のサポートの統計的有意検定による時間変化分析を行う方法を開発する．さらに, バスケット分析による多頻度アイテム集合と QFIMiner から得た定量的多頻度アイテム集合の比較による多頻度アイテム集合の特徴分析方法の開発を行い性能評価実験を行う．

連絡先: 大阪産業科学研究所
〒 567–0047 大阪府茨木市美穂ヶ丘 8-1
E-mail: k-kido@ar.sanken.osaka-u.ac.jp

2. バスケット分析と QFIMiner の概要

バスケット分析は，１つ以上のアイテムの集合であるトランザクションで構成されるデータベースを対象とする．例えば, スーパーで各顧客が購入した商品の情報は以下のようにデータベースに保存されている．

顧客$_1$ ： { 精肉，青果，菓子 − − − 惣菜 }
− − − − −
顧客$_n$ ： { 食品，青果，鮮魚 − − − 菓子 }

ここでは各顧客の購入物が記号アイテム，顧客単位のデータがトランザクションに相当する．アイテム集合 I の支持度はアイテム集合 I の出現頻度を表し, 全トランザクションの中でそのアイテム集合を含むものの割合を示す値である. 即ち

$$\sup(I) = \frac{I \text{ を含むトランザクション数}}{\text{全トランザクション数}}$$

である. バスケット分析では支持度に閾値を設け, それらの値以上の支持度を取り出す. その閾値を最小支持度と言う．また, 最小支持度以上の支持度を有するアイテム集合を多頻度アイテム集合と呼ぶ.

現実のトランザクションには記号アイテム以外にも,「りんご 5 個」「牛肉 200g」というような数値データを含んだアイテムが混在していることが多い．QFIMiner は, 従来のバスケット分析では困難であった, 数値を含むデータを自動的に解析することを可能にした．この手法は数値と記号アイテムからなるトランザクションの集合であるデータセット D から定量的多頻度アイテム集合 (QFI) を探索する．トランザクションの数値部分は,D の全属性空間の部分集合 S に含まれる軸平行な超直方体領域を表す．QFIMiner は, トランザクション分布の密度評価のために前もって決めた格子や窓を用いるのではなく,DB-SCAN[Ester 96] と同様な密度の定義を使用する．この手法は, 適度な密度の閾値を用いることでクラスタを見逃す可能性を大きく減らすことができる．QFIMiner では, 一次元の部分空間のクラスタの探索からはじめ,（k-1）次元のクラ

スタを結合して k 次元の部分空間 S からクラスタの候補 $\hat{C}^S$ を順次生成していく幅優先探索アルゴリズムを用いる. これは SUBCLU[Kailing 04] と似ているが,QFIMiner は, 幅優先部分空間クラスタリングを標準的な Apriori アルゴリズムに埋め込むことで数値アイテムと記号アイテム両方からクラスタを導出することができる. 即ち, 数値アイテムと記号アイテムで構成される属性部分空間に存在する, 最小支持度（minsup）以上のトランザクションに支持されたクラスタが完全探索できる.

QFIMiner を多頻度アイテム集合の探索に用いる時の主要な利点には次が挙げられる,

- 多数のアイテムを含むデータから有用なアイテムの組み合わせを自動的に完全導出できること.
- 数値を含むデータを扱える事で, 探索できるデータが従来のバスケット分析に比べ飛躍的に多種多様になること.
- データの数値部分を解析する場合において, 値の異なる数値アイテムを記号アイテムのように区別して頻度計算するのではなく, 類似した値を持つ数値アイテムは同じクラスタに属するとして計算するため, 多頻度アイテム集合探索の際, 数値を領域で捉えることが可能となり, 定量的多頻度アイテム集合の探索が可能となる.

これらの理由により,QFIMiner を本手法に適用することでバスケット分析を適用した際に得られる情報とは異なる有用な情報が得られる可能性が高い.

3. 提案手法

3.1 バスケット分析の時間変化の検出方法と比較方法

対象データが数値を含まない, ないしはデータ中の数値を無視する場合には, 各期間にバスケット分析を適用する. バスケット分析後, 比較期間における出力結果より, 同じ共起パターンのサポートの値を比較し, 統計的有意差を生じた多頻度アイテム集合を検出する. 統計的有意差が存在するかどうかのを調べるために, 比率の差の検定を行う.

２組の独立なトランザクションデータセット D_1,D_2 において, ある事象共起パターンの支持度（頻度）が p_1 と p_2 であるとする. ここで比率の差の検定のために仮説 $H_0 : p_1 = p_2$ を立てる. ここで,x_1 はト D_1 の母数 n_1 から確率 p_1 で起こる共起事象パターンの出現数なので $P(X = x) = {}_{n_1}C_x(p_1)^x(1-p_1)^{n_1-x}$, x=0,1,2, ⋯ ,n となり２項分布 $B(n_1, p_1)$ に従う. このとき x の期待値と分散は以下のようになる

$$E(X) = \sum_{x=0}^{n_1} x {}_{n_1}C_x p_1^x (1-p_1)^{n_1-x} = n_1 p_1$$

$$V(X) = E(X^2) - \{E(X)\}^2 = n_1 p_1 (1-p_1)$$

となる. ゆえに $\hat{p_1}$ の期待平均と期待分散はそれぞれ $(p_1, \frac{p_1 q_1}{n_1})$ となる. ただし $q_i = 1 - p_i$ 同様にして $\hat{p_2} = \frac{x_2}{n_2}$ の期待平均と期待分散はそれぞれ $(p_2, \frac{p_2 q_2}{n_2})$ となる. よって $\Delta\hat{p} = \hat{p_1} - \hat{p_2}$ の期待平均は $p_1 - p_2$, 期待分散は伝播公式より $\frac{p_1 q_1}{n_1} + \frac{p_1 q_2}{n_2}$ となる. ただし σ_1, σ_2 はそれぞれ $\hat{p_1}, \hat{p_2}$ の期待標準偏差とする. 以上のことより D_1, D_2 の母数 n_1, n_2 が十分に大きいとき $\Delta\hat{p}$ は近似的に正規分布 $N(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2})$ に従う. ここで仮定 H_0 が真で $p_1 = p_2$ のときは, この共通の値を p とおくと $\Delta\hat{p}$ は近似的に正規分布 $N(0, (\frac{1}{n_1} + \frac{1}{n_2})pq)$ に従う. ただし $q = 1 - p$ とする.$\Delta\hat{p}$ の期待標準偏差を $\sigma = \sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}$ とおく. ここで, 標準偏差の２倍の外れ値を棄却域と設定し,$u(\frac{\alpha}{2}) = 2.00$ とおく. また, $\int_{-\sigma}^{\sigma} P(\Delta\hat{p})d\Delta\hat{p}$ を $\Delta\hat{p} = -\sigma$ 〜σ まで 2σ の長さ定積分すると, 正規分布表より, 値が 0.955 となるゆえに $\Delta\hat{p}$ が両側 $\Delta\hat{p} < -\sigma$ 及び $\sigma < \Delta\hat{p}$ になる確率は $1 - 0.955 = 0.045$ となる. よって有意水準を $\alpha = 0.045$ としたとき,p_1 と p_2 の差が統計的に有意な差を生じる条件式は

$$|\hat{p_1} - \hat{p_2}| > u(\frac{\alpha}{2})\sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})} \quad \cdots (1)$$

で表される. これより, バスケット分析の出力結果を期間で比較した際にサポートの差が（１）の式を満たしていれば有意差が存在すると判定できる.

分析の流れは以下のようになる. まず前処理後のデータを適当な期間で区切り, その後, 各期間でバスケット分析を適用する. 複数の期間に亘る結果に同一の多頻度アイテム集合が存在した場合, それらの支持度を比較し, 統計的有意差を生じているかを検証し, 生じていれば出力する. バスケット分析を適用する際の閾値を α として, 多頻度アイテム集合を検出する際には, 任意の期間 A では, 多頻度アイテム集合ではないアイテム集合が, 他期間 B においては支持度が閾値を僅かに上回る多頻度アイテム集合である時, 期間 A では, そのアイテム集合は多頻度アイテム集合として導出されず, その支持度が不明なため, 期間 B で多頻度アイテム集合として導出されても, 支持度に統計的有意差を生じているかが分からない. しかし, 閾値を $\alpha - \beta$ に設定することで, ある期間 A においては多頻度アイテム集合ではないアイテム集合が, 他期間 B では多頻度アイテム集合である場合, その支持度が α 以上ならばそのアイテム集合は α 以上の支持度を持ち, かつ統計的有意差を生じるものであると判別できる. そこで, 支持度 α 以上のアイテム集合を検出したい時, 支持度 α 付近での統計的有意差を β とすると, バスケット分析を適用する際の閾値は $\alpha - \beta$ に設定する.

3.2 QFIMiner の時間変化の検出方法と比較方法

データの数値属性も対象とする場合,QFIMiner を適用する. この場合, 基準とする期間データに QFIMiner を適用し, その期間の定量的多頻度アイテム集合と同じアイテム集合の支持度を他期間について調べる. その際, 各アイテム集合に対し, 基準となる期間の定量的多頻度アイテム集合と同じ数値領域で支持度を求める. ここで, 他期間についても QFIMiner を適用すると, 基準となる期間の出力結果と異なる数値領域を有する定量的多頻度アイテム集合が得られてしまう可能性がある. 数値領域が異なると, 定量的多頻度アイテム集合間で支持度の差の検定を適用できない. 以上の理由により, 本手法では基準となる期間の定量的多頻度集合の結果と同じ条件で, 他期間の支持度を調べる. 各期間における支持度を求めた後, 期間同士の結果から支持度を比較し, 統計的有意差を生じた定量的多頻度アイテム集合を検出する. 統計的有意差の検出方法としては, バスケット分析の際と同様の方法を取る. この際の最小支持度は, 先述した同様の理由により $\alpha - \beta$ を適用する.

4. 手法性能検証

4.1 人工データの概要

提案手法の性能を人工データを用いて調査する. 性能検証の方法は数値アイテムを含む２つのデータ間で, 支持度に有意差を持つ定量的多頻度アイテム集合を検知できるかを見る. ただ

し, その際用いた人工データは１万個のトランザクションの中に, 特定のパターンのアイテム集合を埋め込んだデータであり, 片方のデータ A に埋め込む個数は固定し, もう一方のデータ B の埋め込み個数を変化させて検証を行う. ここで, 支持度 60% 付近での有意差を a%, 支持度 5% 付近での有意差 b% とすると, 埋め込み個数の組み合わせは, (6000,6000),(6000,6000-(10000×$\frac{a}{100}$)),(6000,6000-(10000×$\frac{2a}{100}$)),(500,500),(500,500-(10000×$\frac{b}{100}$)),(500,500-(10000×$\frac{2b}{100}$)) となる. また, 上記の埋め込み個数の組み合わせそれぞれにおいて, 数値アイテムの値に正規ノイズを加えて, 埋め込みパターンに数値アイテム空間内で一定の広がりを持たせた. これにより, 多頻度パターンが数値アイテム空間上で一定の広がりを有するクラスタを形成している場合に関する支持度の有意差検知性能を評価する. ここでは, 加える雑音の標準偏差を 0.01 と 0.1 の２種類に変化させ, 即ち埋め込み多頻度パターンの数値アイテム空間上での広がりを変化させ, 検証を行う. なお, 標準偏差は数値アイテムの値域の幅に対する割合である. １つの実験条件設定に対して 10 回分析を行い 10 回中何回有意差があると検知したかを調べる. (ただし, 本評価実験において a,b の値は a=1.39%, b=0.616% として埋め込み個数を決定した.)

4.2 シミュレーション評価結果

各実験条件設定において, 有意差を持つ定量的多頻度アイテム集合が 10 回中何回得られたかを表１及び２に示す. これより, 本提案手法は数値アイテムの値域の 10% にあたる標準偏差で数値アイテムの値にノイズを加えたデータに対しても, 実用に耐えうるだけの支持度の統計的有意差検出性能を示すことが分かる.

表 1: 標準偏差 $\sqrt{\sigma^2}$=0.01 の場合

埋め込み個数の組み合わせ	有意差を検出した回数
(6000,6000)	0/10
(6000,6000-(10000×$\frac{a}{100}$))	10/10
(6000,6000-(10000×$\frac{2a}{100}$))	10/10
(500,500)	0/10
(500,500-(10000×$\frac{b}{100}$))	10/10
(500,500-(10000×$\frac{2b}{100}$))	10/10

表 2: 標準偏差 $\sqrt{\sigma^2}$=0.1 の場合

埋め込み個数の組み合わせ	有意差を検出した回数
(6000,6000)	0/10
(6000,6000-(10000×$\frac{a}{100}$))	7/10
(6000,6000-(10000×$\frac{2a}{100}$))	10/10
(500,500)	0/10
(500,500-(10000×$\frac{b}{100}$))	8/10
(500,500-(10000×$\frac{2b}{100}$)	10/10

5. 評価実験

5.1 マーケティング実データの概要

提案手法により数値属性を含むマーケティング実データの分析を行った. 本データは 2005 年 3 月 1 日から 2005 年 6 月 30 日までのある店舗での購買履歴である. 4 月 13 日から 5 月 3 日までの３週間ある菓子のテレビ CM を流し,CM を流すことにより菓子を含む購入物の共起パターンがどのように変化したかを部門レベルで分析する. ここで, 部門とは同種の商品をひと括りにした集合体である. 例として, 南瓜や人参などの野菜と林檎や檸檬などの果物は青果という１つの部門に含まれる. 対象とする数値情報は１度の買い物の際に購入した各アイテムの購入個数である. バスケット分析の際には, 数値情報を除き, 多頻度アイテムの共起パターンを分析し, QFIMiner に適用する際には数値情報を加えて分析する. 分析するデータとしては, データの収集期間中に着目した菓子品目を１度でも購入した顧客を対象とし, 元々のデータから対象購入者の菓子を含む購買データを取り出したものとする. 前処理後のデータを４つの期間に分け, それぞれの期間において, バスケット分析および QFIMiner を適用し, 出力結果を期間ごとに比較しどのような変化が起こったかを調べ, それらの変化より CM が顧客の購買物の共起パターンにどのように影響を及ぼしたかを考察する. また, 考察対象店舗としては, データ６店舗中来客数の最も多かった店舗とした. 期間の設定は以下のように行った

期間１;CM を流す前 (3/1〜4/12)

期間２;CM が流れている期間 (4/13〜5/3)

期間３;CM の影響が最も表れる期間 (CM を流した後, 最も着目した菓子品目の売り上げが大きい期間)(5/4〜5/17)

期間４; それ以降の期間 (5/18〜6/30)

期間同士の比較は以下のように行った

期間１と期間２の比較;CM の効果にどれほどの速攻性があるかを比較結果より考察できる

期間１と期間３の比較;CM の効果が最も顕著に比較結果より考察できる

期間１と期間４の比較; この比較により CM の影響で長期的な購買共起パターンがどのように変化したかを考察できる

5.2 分析と考察

バスケット分析における最小支持度は 0.3 とした. これは, 過半数にいかないまでも無視できない多くの人数, つまり主要な割合の消費行動の変化を知るためである. ４つの期間それぞれにおいてバスケット分析を行い, それらの結果を比較し, 有意差の有無を調べると期間１と期間３の比較において有意差を生じた多頻度アイテム集合が検出された. 有意差の生じた共起パターンとサポート値を表２に示す. バスケット分析で有意差の生じた共起パターンの QFIMiner での分析結果を表３に示す.

表 3: 有意差の生じた共起パターン

多頻度共起パターン	期間１	期間３
精肉, 青果, 食品, 菓子, 日配	0.374	0.426
精肉, 青果, 菓子, 日配	0.487	0.540
精肉, 食品, 菓子, 鮮魚, 日配	0.313	0.357
精肉, 食品, 菓子, 日配	0.414	0.469
精肉, 菓子, 日配	0.541	0.603
精肉, 青果, 食品, 菓子, 鮮魚, 日配	0.288	0.330

表 4: 有意差の生じた定量的共起パターン

多頻度共起パターン	期間１	期間３
精肉 [1〜5], 青果 [1〜9], 食品 [1〜6], 菓子 [1〜9], 日配 [1〜6]	0.267	0.281
精肉 [1〜5], 青果 [1〜9], 菓子 [1〜9], 日配 [1〜6]	0.389	0.413
精肉 [1〜5], 食品 [1〜6], 菓子 [1〜9], 鮮魚 [1〜5], 日配 [1〜6]	0.207	0.228
精肉 [1〜5], 食品 [1〜6], 菓子 [1〜9], 日配 [1〜6]	0.302	0.324
精肉 [1〜5], 菓子 [1〜9], 日配 [1〜6]	0.442	0.481
精肉 [1〜5], 青果 [1〜9], 食品 [1〜6], 菓子 [1〜9], 鮮魚 [1〜5], 日配 [1〜6]	0.183	0.195

表 5: $\frac{QFIMiner\text{ での支持度}}{\text{バスケット分析での支持度}}$ の比

多頻度共起パターン	期間１	期間３	支持度比の差
精肉, 青果, 食品, 菓子, 日配	0.714	0.660	-0.054
精肉, 青果, 菓子, 日配	0.800	0.765	-0.035
精肉, 食品, 菓子, 鮮魚, 日配	0.661	0.639	-0.023
精肉, 食品, 菓子, 日配	0.729	0.691	-0.038
精肉, 菓子, 日配	0.817	0.798	-0.019
精肉, 青果, 食品, 菓子, 鮮魚, 日配	0.635	0.591	-0.044

表３より, 考察店舗における有意差を生じた多頻度共起パターンを見ると, それらはすべて最も大きな多頻度共起パターン｛精肉・青果・食品・菓子・日配｝の部分集合である. 従って, この最大の多頻度共起パターンの購入者の変化が他のすべての有意差の生じた共起パターンに大きく影響を及ぼしている. このことより, 期間１と期間３の比較における有意差を生じさせた原因は｛精肉・青果・食品・菓子・日配｝の購入者であることが分かる. また, ｛精肉・青果・食品・菓子・日配｝を１度に購入する客は１人暮らしではなく, 家族と同居していると推測される. また, 購入アイテムの多様性から主婦であり, １度にこれだけの量を購入する必要性があることから, 家族の中に子供もいるだろうと推測される. ゆえに, 期間１と期間３の比較において着目される購入者は家族を持った主婦, 特に子供のいる主婦であろうと考えられる. このことより,CM がターゲット層である子供を持つ主婦層に的確に影響を及ぼすことができたことが分かる. また表３を見ると,QFIMiner の結果より期間１と期間３のサポートの差を比較すると, バスケット分析の時よりも, すべてのアイテム集合において差が小さくなっていることが分かる. QFIMiner の個数の領域はすべてのアイテムにおいて１〜数個である. これより QFIMiner によって多頻度アイテムとされたものは個数の面から, 少量購入のアイテム集合を示す. 表５にバスケット分析の結果である記号アイテムのみの支持度と, QFIMiner の結果である各記号アイテムに少量の購買個数領域を加えた時の支持度の比を示す. これにより期間１と期間３の比較において, 期間３では少量の購入者の割合が減少し, 大量購入者（まとめ買い）が増えたことが分かる. よって, 考察店舗では CM がターゲットである子供を持つ主婦層に影響を及ぼし, 菓子購入者が増加した. また, 単に菓子購入者が増加しただけでなく購入の仕方も少量購入から大量購入へと変化している. しかし, 使用したデータの店舗周辺の情報, 他の商品が菓子に与えた影響等は本評価実験において, データがなかった為, 考慮することができなかった. そのため, 今回のデータのみで得られた考察はあくまでも仮定の範囲を出ない.

6. 結論

本研究では時間とともに変化する数値アイテムを含む共起パターンの分類手法を提案した. 提案手法は最小支持度を上回る範囲ではすべての多頻度アイテム集合を抽出し, それを元に時間軸上での有意差の生じる多頻度アイテム集合を出力することが可能である. 提案手法は実際のスーパー６店舗での数ヶ月にわたる購買データを用いて性能評価実験を行い, データがどのような変化をし, 外部からの影響をどのように受けたかを考察することを可能にした. しかし, データに適用する際に期間を区切る判定基準の開発, 有意差検定においてデータの分布を仮定するために生じる誤差, 共起パターン変化分析を行う際のデータの前処理の必要性等, 課題が残っている.

参考文献

[1] Mitsunaga Y.,Washio T.,Motoda H.: 適応的密度基準に基づく部分空間クラスタリングを用いた定量的多頻度アイテム集合のマイニング, 大阪大学大学院工学研究科通信工学科修士学位論文 (2006)

[2] Ester,M.,Kriegel,H.-P.,Sander,J.,and Xu,X.: A Density-Based Algorithm for Discoverring Clusters in Large Spatial Databases with Noise, In Proc.2nd Int.Conf.on Knowledge Discovery and Data Mining,pp.226-231(1996)

[3] Kailing,K.,Kriegel,H.-P.,and Kroger,P.: Density-Connected Subspace Clustering for High-Dimensional Data,Proc.Fourth SIAM International Conference on Data Mining(SDM'04),pp.246-257(2004)

[4] 稲垣宣生, 山根芳和, 吉田光雄:統計学入門, 裳華房 (1992)

[5] 篠崎信雄:統計解析入門, サイエンス社 (1995)

[6] Michael J.A. Berry,Gordon Linoff:SAS インスティチュート, 江原淳, 佐藤栄作:共訳データマイニング手法, 海文堂 (1999)