

# 特保茶飲料に関する購買トランザクションのバイクラスタリング: 併買パターンの抽出と実店舗における実証実験

Biclustering the purchase transaction data for tea drinks: Extracting interesting patterns and its experimental evaluation in the real supermarket store

石田悠真<sup>\*1</sup>  
Yuma Ishida

羽室行信<sup>\*2</sup>  
Yukinobu Hamuro

丸橋弘明<sup>\*3</sup>  
Hiroaki Maruhashi

加藤直樹<sup>\*4</sup>  
Naoki Katoh

宇野毅明<sup>\*5</sup>  
Takeaki Uno

<sup>\*1\*</sup><sup>\*2\*</sup><sup>\*3\*</sup><sup>\*4</sup>関西学院大学  
Kwansei Gakuin University

<sup>\*5</sup>国立情報学研究所  
National Institute of Informatics

The objective of this paper is to show the case study about a knowledge discovery in the consumer behavior database, and test the effectiveness of the method, biclustering, in conducting an experimental evaluation at the real supermarket store. The biclustering we used consists of two core methods, bipartite graph polishing and maximal clique enumeration. The experiment of selling the specified health drinks with non-healthy foods showed high sales gain, but cannot be confirmed in other experiments. We learnt difficulty to implement the datamining results in real world, and get insight of the reasons.

## 1. はじめに

データマイニング研究は, Agrawal らによって相関ルール列挙の効率的なアルゴリズムが提案されて以来, 多くの理論研究および応用研究が行われてきた. しかしながら理論研究においては過度に効率に焦点があたってしまったために, 応用上の課題に答えられないことも多い. 一方で, 応用研究においては, 様々なヒューリスティックを駆使して, 意味ある結果を導出しているが, それが現実の経営現場で本当に有効かどうかを検証されることはない. さらに, 近年の機械学習/AI ブームに押されて, データマイニング手法への注目はますます低下してきているのが現状であろう.

データマイニングは, 広義には, 人間にとって有用な仮説の導出を手助けする諸技術と定義できる. この定義の抱える根本的な課題は, 得られた仮説の有用性をどのように評価するかにある. 仮説を前提とする統計学は, 一貫して有意確率 (P 値) によって仮説の評価をしてきた. また機械学習では, 理論 (仮説) は重視せず一貫して未知データへの予測精度によって評価してきた. では, データマイニングが目的とする仮説の有用性はどのように評価すればよいのであろうか? 例えば相関ルールマイニングでは, 支持度や確信度といった数値的な評価基準が数多く提案されているが, これは一つの評価指標だけでは仮説の有用性を評価できないという事実の裏返しでもある. 著者らの考えでは, 仮説の有用性は, 数値的な指標に加えて, その仮説が利用される現場の納得感や施策への利用可能性といった主観的な評価を取り入れない限り評価できないと考える. そして得られた仮説をもとにして何らかの施策を実施し, その効果を確かめることで初めて数値的な評価が可能となるのである.

以上のような背景のもと, 本研究の目的は, データマイニング手法である 2 部グラフ研磨と極大 2 部クリーク列挙を用いたバイクラスタリングを消費者購買履歴データに適用することで有望な仮説を列挙し, そこから小売店の現場の人間と協働で有用な仮説を導出し, その仮説の有効性を店頭実験によって検証することにある. 大枠としての課題は, 小売店の現場からの要請で, 特定保健用食品の茶類 (以下「特保茶」と呼称する) をどのように販売すればよいかというものであり, バイクラスタリングにより, 特保茶との併買パターンおよび消費者のプロファイ

ルを分析することが可能となり, 現場で施策として落とし込める有用な仮説の導出が可能となる.

## 2. 手法

本研究で用いる手法は, 2 部グラフの研磨と極大クリーク列挙を用いたバイクラスタリング [3] である. クラスタリングとは目的変数を持たないデータ集合をデータ間の類似性に基づいていくつかのグループに分類する手法の総称である. 一般的なクラスタリングは, 一つのデータ集合のグルーピングに用いられるのに対し, バイクラスタリングは, 2 つの集合間の関係性データから, それら 2 つの集合を同時にグルーピングする. 本研究では, 購買履歴データにおける, レシート集合と商品集合の 2 つの集合の関係データを扱う.

バイクラスタリングが対象とする関係性データは 2 部グラフ  $G = (V \cup U, E)$  によって記述できる. ここで,  $V = \{v_1, v_2, \dots, v_n\}$  はレシート集合で,  $U = \{u_1, u_2, \dots, u_m\}$  は商品集合である (今回の実験では  $n = 9893, m = 510$ ).  $E$  は 2 つの集合間に張られた枝集合であり, あるレシート  $v_i \in U$  で商品  $u_j \in U$  が購入されていれば  $v_i, u_j \in E$  間に枝が張られる.

本研究で対象とするレシート集合  $V$  には, 特保茶が含まれるレシートのみを事前に選択しており, また商品集合  $U$  は, それらの特保茶以外の商品群として定義している. よってそのような 2 部グラフからバイクラスタを列挙することにより, どういった顧客 (レシート) がどのような商品集合を特保茶と併買する傾向があるかに関するルールを得ることが期待できる.

バイクラスタリングの方法はいくつか提案されているが, 本研究で利用する手法 [3] は, 大雑把に言えば,  $G$  上の密な部分グラフをバイクラスタとして抽出するものである. 密な部分グラフを抽出するにあたって容易に考えつく方法は, 極大 2 部クリークを抽出することであろう. 2 部クリークとは, 2 部グラフの頂点部分集合  $K \subseteq V, H \subseteq U$  に対して,  $K$  の任意の頂点と  $H$  の任意の頂点の間に枝があるとき, そのような  $K$  と  $H$  から誘導される部分グラフのことである. そして, ある 2 部クリークが他の 2 部クリークに含まれないとき, その 2 部クリークは極大 2 部クリークと呼ばれる.

しかしながら, 現実のデータベースにおいて極大 2 部クリークをバイクラスタとして抽出すると類似したクラスタが多数列挙されてしまうという問題がある. これは, 例えば, ある極大 2 部クリークから枝を一本削除するだけで, その極大 2 部クリーク

連絡先: 羽室行信, 関西学院大学, 西宮市上ヶ原一番町 1-155,  
TEL:0798-54-6572, hamuro@kwansei.ac.jp

クが2つの極大2部クリークに分かれるが、現実のデータにはノイズが多くついているために、そのような部分構造が多く現れ、結果として多数の類似したクリークが列挙されるのである。そこでこのようなノイズを除去する技術として2部グラフ研磨手法が提案されている[3]。

## 2.1 2部グラフ研磨

データ研磨の基本的なアイデアは、与えられたグラフの密な部分はより密に、疎な部分はより疎になるようにグラフを再構成することによって本質的な構造を失うことなくグラフ構造を明確化しようとするものである。2部グラフ  $G = (V \cup U, E)$  において、 $X(v)$  を  $v \in V$  に隣接する  $U$  の頂点集合とする。また、 $X(u)$  を  $u \in U$  に隣接する  $V$  の頂点集合とする。2部グラフの研磨アルゴリズムを Algorithm 1 に示す。ここで示すアルゴリズムは、効率の悪い方法ではあるが、理解のし易さを優先させている。

### Algorithm 1 2部グラフ研磨アルゴリズム

```

1: function BIPOLISHING( $G = (V \cup U, E)$ ,  $\sigma_1, \sigma_2$ )
2:    $V, U$ : 頂点集合,  $E$ : 枝集合,  $\sigma_1, \sigma_2$ : 類似度下限値
3:    $V', U', E' = \phi$ 
4:   for all  $v \in V$  do
5:      $S = \phi$ 
6:     for all  $v' \in V$  do
7:       if  $\text{sim}(X(v), X(v')) \geq \sigma_1$  then
8:          $S = S \cup \{v'\}$ 
9:       end if
10:    end for
11:    for all  $u \in U$  do
12:      if  $\text{sim}(X(u), S) \geq \sigma_2$  then
13:         $E' = E' \cup \{(v, u)\}$ ,  $V' = V' \cup \{v\}$ ,  $U' = U' \cup \{u\}$ 
14:      end if
15:    end for
16:  end for
17:  return  $G = (V' \cup U', E')$ 
18: end function

```

このアルゴリズムは、任意の2つの頂点  $u \in U$ ,  $v \in V$  について、枝を張るか張らないかを隣接関係から評価することで新たなグラフを再構成しようとするものである。まず、 $U$  への接続関係が頂点  $v$  と類似した頂点集合  $S$  を見つけ出す(6~10行目)。 $V$  をレシート集合、 $U$  を商品集合とすると、商品との結びつき(共起関係)がレシート  $v$  と類似したレシート集合が  $S$  となる。今度は逆に  $u \in U$  から  $V$  への接続を調べる。 $u$  の接続先の頂点集合 ( $X(u) \subseteq V$ ) と  $S \subseteq V$  との類似性を判断し(12行目)、類似していれば、頂点  $v, u$  を接続し、類似していなければ接続しない。このことにより、お互いの共起関係において類似したレシート  $v$  と商品  $u$  に枝を張り直すことが可能となる。なお、類似度の計算(7,12行目)には jaccard 係数 ( $\text{sim}(A, B) = |A \cap B| / |A \cup B|$ ) を利用した。

このようにグラフ研磨は、集合間の接続関係(お互いの共起関係)によってオリジナルの2部グラフのグラフ構造を変更するため、2部グラフ上で孤立した枝は省かれ、また密な部分グラフにおける少数の欠損枝を修復するなど、集合間の関係性の明確化およびノイズのクリーニングとしての効果が期待できる。

上記のアルゴリズムを利用し、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。そして、最終的に得られた2部グラフが研磨2部グラフである。

## 3. データ

本研究で利用したデータは、マーケティングリサーチを主要事業とする株式会社マクロミルが収集している消費者購買履

歴データ QPR である。QPR とは、全国3万人のモニタに専用のバーコードリーダーを貸与して日々の購買情報を収集した消費者購買行動履歴データである。データの対象期間は2016年10月1日から2018年9月30日までの2年間である。対象商品は飲食品とし、対象業態はスーパーマーケットのみとした。さらに、分析対象商品の特保茶とした。特保茶は健康効果に付加価値を持つ特定保健用食品の茶類である。研究で対象とする特保茶は、評価実験を行うスーパーマーケットの実店舗で販売されている商品のみとし、それらを全て同一の特保茶とみなして併買分析を行った。また、容量による購買傾向の差異を考慮し、本研究では500ml以下の容量の特保茶のみを対象とした。QPR データにおいて、特保茶を含んでいるレシート枚数は計14,976枚であり、モニタ数の合計は2,515人である。

## 4. 知識発見

### 4.1 バイクラスタの列挙

2部グラフ研磨では、2つのパラメータ  $\sigma_1$  と  $\sigma_2$  によって得られる研磨グラフが異なってくる。これらのパラメータの組み合わせにより、得られる極大2部クリークの数がどのように変化するかを示したのが表1である。研磨をしなかった場合に得られた極大2部クリーク数は8433個であった。各閾値を大きくするほど、得られる極大2部クリークの数が少なくなっていることがわかる。では、パラメータ  $\sigma_1, \sigma_2$  をどのようにして決めればよいであろうか? 本研究では、パラメータを一意に定めるのではなく一定の粒度の0.05刻みで網羅的に探索し、その結果得られた全クラスタの中から信頼性の低いクラスタをフィルタリングするという方法を採用することにした。

表1: 研磨の  $\sigma_1, \sigma_2$  の閾値と極大2部クリーク数の関係

| $\sigma_1$ | $\sigma_2$ | 0.1   | 0.2   | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|------------|------------|-------|-------|-----|-----|-----|-----|-----|-----|-----|
| 0.1        | 4,226      | 3,038 | 1,791 | 899 | 519 | 506 | 506 | 501 | 488 |     |
| 0.2        | 3,659      | 2,635 | 1,517 | 740 | 508 | 497 | 464 | 367 | 166 |     |
| 0.3        | 2,756      | 1,704 | 744   | 481 | 416 | 273 | 134 | 50  | 5   |     |
| 0.4        | 1,655      | 727   | 432   | 267 | 144 | 48  | 13  | 1   | 0   |     |
| 0.5        | 1,760      | 705   | 427   | 263 | 141 | 48  | 13  | 1   | 0   |     |
| 0.6        | 487        | 242   | 80    | 20  | 6   | 1   | 1   | 0   | 0   |     |
| 0.7        | 456        | 221   | 73    | 18  | 5   | 1   | 1   | 0   | 0   |     |
| 0.8        | 450        | 219   | 74    | 18  | 5   | 1   | 1   | 0   | 0   |     |
| 0.9        | 447        | 217   | 73    | 18  | 5   | 1   | 1   | 0   | 0   |     |

### 4.2 バイクラスタのフィルタリング

以上に示したグリッド探索の結果得られた39,377個のバイクラスタをフィルタリングするために、バイクラスタの良さを式(1)~(6)に示される6つの評価指標により定義した。

$$|C| \quad (1)$$

$$RR = |D|/|C| \quad (2)$$

$$RM = |N|/|M| \quad (3)$$

$$CRR = (|C \cap D|)/\min(|C|, |D|) \quad (4)$$

$$CRM = (|N \cap M|)/\min(|C|, |D|) \quad (5)$$

$$GR = \frac{|C|/|T|}{|C'|/|T'|} \quad (6)$$

ここで、 $C$  はバイクラスタを構成するレシート集合、 $M$  はバイクラスタに属するモニタ集合、 $D$  は実データ上でバイクラスタの商品群を含むレシート集合、 $N$  はそのモニタ集合である。また、GRは顕在パターン[1]における増加率(growth rate)であり、バイクラスタを構成する商品集合を含むレシートの出現確率  $|C|/|T|$  が、対照群(コントロール群)における出現確率  $|C'|/|T'|$  に比べてどれほど高いかを示す。ここで  $T$  は、レシートの総件数であり ( $|T| = 14,796$ )、 $T'$  は対照群におけるレシートの総件数を表す ( $|T'| = 185426$ )。対照群として

は、本研究では、特保茶を含まず、かつ特保茶以外の一般的なお茶を含むレシート集合とした。

$RR, RM, CRR, CRM$  は、いずれも、グラフ研磨を行うことで2部グラフ上の枝の付替えが発生することでアイテム-レシート構造が改変されるが、その改変の程度が一定の範囲内であるようなクラスタのみに限定しようというものである。 $RR(RM)$  は、バイクラスタを構成する商品を含むオリジナルデータ上のレシート (モニタ) とバイクラスタ上のレシート (モニタ) のサイズの違いを一定範囲に抑えようとするものである。さらに  $CRR(CRM)$  は、単にサイズのみを考慮するのではなく、共通するレシート (モニタ) の割合が高いクラスタに限定しようというものである。そして、 $GR$  は、対照群である一般のお茶に比べて特保茶に特徴的なクラスタに限定する効果がある。

以上の5つの評価指標について、式(7)~(10)で示される条件を設定し、2871個のバイクラスタを選択した。

$$|C| \geq 10 \quad (7)$$

$$2.0 \geq RR, RM \geq 0.5 \quad (8)$$

$$CRR, CRM \geq 0.9 \quad (9)$$

$$GR \geq 1.3 \quad (10)$$

さらに、異なる  $\sigma_1, \sigma_2$  で列挙されたバイクラスタには、同一アイテム集合で構成されるバイクラスタが存在し、そのような場合、それらのバイクラスタの中で最も  $RR$  が1.0に近いクラスタを選択し、最終的に136個のバイクラスタを得た。

### 4.3 得られた3つの知見

得られた136個のバイクラスタについてスーパーマーケットの現場の責任者と検討を重ねた結果、以下に示す3つの有用な知見を得るに至った。ここで有用とは、商品の併売についての意味的妥当性があり、顧客特性が実験実施店にも合致し、施策に結びつけやすいと評価した知見のことである。

#### 4.3.1 健康志向的な併買

特徴的であった併買パターン群の1つに200ml紙パックの健康志向飲料がある。その一覧を表2に増加率の降順に示す。野菜ジュースやトマトジュース、豆乳、青汁など様々な200ml紙パックの健康志向飲料が併買パターンとして挙がっていた為、特保茶と200ml紙パックの健康志向飲料の購買には相関があることが考えられる。

またモニタの特性について見ると、健康を意識してはいるものの、普段は料理をしない若い単身の男性がスーパーで弁当や惣菜を購入する際、食前や食間に、気軽に栄養を摂取できる200ml紙パックの健康志向飲料を飲み、食中や食後の外出時、勤務時の飲料として健康効果に付加価値を持つ特保茶を購入しているのではないかと推測した。

表2: 健康志向的な特保茶との併買商品一覧

| 名称              |       | GR   | M  | C   |
|-----------------|-------|------|----|-----|
| A社 紙パック野菜ジュース 1 | 200ml | 16.1 | 20 | 241 |
| A社 紙パック野菜ジュース 2 | 200ml | 11.1 | 18 | 243 |
| B社 紙パック野菜ジュース   | 200ml | 4.8  | 10 | 67  |
| A社 紙パックトマトジュース  | 200ml | 4.6  | 19 | 43  |
| B社 紙パック青汁       | 200ml | 3.4  | 12 | 23  |
| C社 紙パック豆乳       | 200ml | 2.1  | 11 | 15  |

#### 4.3.2 免罪符的な併買

健康志向的な購買とは逆に、非健康志向的であると思われる商品との併買パターンも多く、例えばスナック菓子やカップ麺、アルコール飲料などが挙げられる。これは、健康を意識してはいるものの不健康な商品の購買は止められない為に、健康効果のある特保茶を同時に購入することによって健康を維持しようという、いわば免罪符的な購買動機があると推測される。

表3: 免罪符的な併買パターン例

| 名称           |  | GR   | M  | C   |
|--------------|--|------|----|-----|
| A社 スナック菓子    |  | 4.94 | 14 | 86  |
| B社 カップ焼きそば   |  | 2.49 | 24 | 36  |
| C社 缶ビール      |  | 2.13 | 11 | 12  |
| D社 缶チューハイパック |  | 1.95 | 12 | 20  |
| E社 缶ビール      |  | 1.88 | 11 | 12  |
| F社 缶ビール      |  | 1.86 | 17 | 26  |
| G社 カップうどん    |  | 1.79 | 15 | 22  |
| H社 缶チューハイ    |  | 1.78 | 13 | 22  |
| A社 ポテトチップス   |  | 1.52 | 15 | 20  |
| たばこ (細分類)    |  | 1.43 | 35 | 241 |

#### 4.3.3 高カカオ商品との併買

最後に、高カカオチョコレートと特保茶の併買を取り上げた。これは、現場の目線から見て特に興味深いと評価された併買パターン群である(表4)。このクラスタの顧客は高齢の女性に特徴的であり、健康を意識する年配者が動脈硬化や認知機能の改善に効果があるとの報告もあるカカオと特保茶を摂取しているのではないかと推測した。

表4: 高カカオチョコレートとの併買パターン

| 名称           |  | GR   | M  | C  |
|--------------|--|------|----|----|
| A社 高カカオチョコ 1 |  | 2.83 | 18 | 24 |
| A社 高カカオチョコ 2 |  | 1.35 | 13 | 18 |

## 5. 店頭評価実験

以上、QPRデータの解析の結果、3つの有用な知見が得られたが、それらの知見に基づいて株式会社光洋が展開するスーパーマーケットMaxValu EX 西宮北口店において以下に示す3つの店頭実験を実施した。

### 5.1 健康志向的な飲料特設コーナー

健康志向的な併買に関する仮説から、特保茶と200ml紙パックの健康志向飲料を併設した特設コーナーを設置した(図1の左)。POPも作成し、弁当、惣菜を購入した顧客をターゲットとして、特保茶と200ml紙パックの健康志向飲料の併買を促した。設置期間は2019/1/1~1/31の1ヶ月間とし、期間を通じて同じ陳列方法で同一POPを利用した。



図1: 売り場の様子. 左から、健康志向飲料売り場、免罪符商品売場、高カカオ商品売場。

評価方法であるが、特設コーナーに設置した特保茶と紙パックの健康志向飲料の併買確率が過去のある期間に比べてどの程度上昇したかを評価した。比較する過去の期間は前年同月(2018/1)、および直近の3ヶ月間(2018/10,11,12)とした。結果は表5に示される通りである。併買確率対比は、前年比、直近3ヶ月比のいずれも1.0を下回っており、健康志向コーナーの設置については効果は認められなかった。現場担当の人間によると、本コーナーの周りには魅力的なコーナーが多く設置さ

れており、本コーナーが目立たなかったのかもしれないとの見解もあり、健康志向飲料を併設する効果については、今後、多様な要因を加味した実験が必要となるであろう。

表 5: 健康志向併買実験の結果。表中の併買確率対比とは、直近3ヶ月併買確率（もしくは前年1月併買確率）に対する当年1月併買確率の比である。

| 内容                      | 当年1月  | 直近3ヶ月  | 前年1月  |
|-------------------------|-------|--------|-------|
| レシート総数 (a)              | 57541 | 184138 | 56649 |
| 200ml 健康紙パック飲料 (単体購買回数) | 454   | 1,454  | 416   |
| 健康コーナー特保茶 (単体購買回数)      | 217   | 564    | 95    |
| 特保茶と紙パックの併買回数 (b)       | 6     | 25     | 6     |
| 併買確率 (b/a:単位は 1/100000) | 10.4  | 13.6   | 10.6  |
| 併買確率対比                  | -     | 0.77   | 0.98  |

5.2 免罪符的飲料の特設コーナー

免罪符的な併買に関する仮説から、非健康的と考えられる商品を購入する顧客をターゲットにして、健康志向飲料の購入を促進することを目的としたコーナーを設置した。多くの顧客に共通する客動線の最終地点付近（すなわちレジの近く）に、健康志向飲料がまとめて陳列されたコーナーを設置した（図1の中央）。非健康志向食品として、スナック菓子、アルコール飲料、たばこ、カップ麺（インスタント麺）、チョコレートの5つを想定した。設置期間は健康志向とコーナーと同様である。

評価方法は、非健康志向食品の購入を条件に、特設コーナーに設置した健康志向飲料の条件付き購買確率（確信度）によって評価した。比較期間は健康志向コーナーと同様である。なお、本実験で陳列した商品のうち、前年の1月にはまだ発売されていない商品があり、前年比による結果は、直近3ヶ月比に比べて相対的に過大評価していることに留意されたい。結果は表6に示される通りである。

前年対比で見ると、タバコ以外には効果が認められたが、直近3ヶ月比で見ると、スナック菓子、アルコール飲料、チョコレートとの併売に高い効果が見られ、その他の非健康志向食品については効果が認められなかった。

表 6: 免罪符的な併買実験の結果。セル内の4つの数字は、上から非健康食品単体のレシート数 (a)、特保茶との併買レシート数 (b)、確信度 (b/a \* 10000)、当年1月の確信度の対前年1月比および直近3ヶ月比を表している。

| 内容      | 当年1月  | 直近3ヶ月  | 前年1月  |
|---------|-------|--------|-------|
| スナック菓子  | 7,365 | 23,320 | 6,599 |
|         | 30    | 70     | 9     |
|         | 40.7  | 30.0   | 13.6  |
|         | -     | 1.36   | 2.99  |
| アルコール飲料 | 8,879 | 28,850 | 8,065 |
|         | 32    | 64     | 18    |
|         | 36.0  | 22.2   | 22.3  |
|         | -     | 1.62   | 1.61  |
| タバコ     | 1,265 | 3,993  | 1,468 |
|         | 2     | 6      | 2     |
|         | 15.8  | 15.0   | 13.6  |
|         | -     | 1.05   | 1.16  |
| インスタント麺 | 6,210 | 17,424 | 5,831 |
|         | 30    | 83     | 11    |
|         | 48.3  | 47.6   | 18.9  |
|         | -     | 1.01   | 2.56  |
| チョコレート  | 5,291 | 14,459 | 4,797 |
|         | 24    | 50     | 8     |
|         | 45.4  | 34.6   | 16.7  |
|         | -     | 1.31   | 2.72  |

5.3 カカオチョコと特保茶の併売コーナー

最後に、高カカオチョコと特保茶の併買を促進するために、ある特保茶に限定し、その冷蔵庫の前に高カカオチョコを併売した（図1の右）。実験として、3パターンの売り方を実施し、それらの結果を比較することで併売効果を確認することにし

た。3パターンの売り方とは、1) 高カカオチョコと特保茶を併売し、かつPOPを表示する、2) 併売のみ実施しPOPはつけない、3) なにもしない、の3つである。これら3パターンの売り方を、2018/12/21～2019/1/31までの42日間に均等ランダムに割り付けることで実験を実施した。結果は表7に示される通りで、高カカオチョコと特保茶の併売効果は全く認められなかった。高カカオチョコと特保茶の併売は、QPRのデータ上では高い頻度で観察されたクラスタであることから、実験を実施した地域の特性にも原因があるかもしれない、これは、今回のように訓練データとテストデータが全く異なるサンプルである場合の難しさであるとも言える。

表 7: カカオチョコとの併売実験の結果

| 内容           | 何もしない | 併売のみ | 併売+POP |
|--------------|-------|------|--------|
| 高カカオチョコ (単体) | 67    | 61   | 56     |
| 特保茶 (単体)     | 73    | 78   | 83     |
| 併買回数         | 0     | 1    | 0      |

6. おわりに

本稿では、2部グラフのデータ研磨と極大2部クリークを用いたバイクラスタリング手法を消費者購買行動履歴データに適用し、実際の小売店と協働して特保茶に関する有用な3つの仮説、すなわち、1) 健康志向的な飲料との併買、2) 非健康的食品との免罪符的併買、3) 高カカオチョコとの併買、を得ることができた。これらの仮説から、実際のスーパーマーケットで3つの売り場を新たに構成し、それらの仮説の有効性を検証したところ、2) については高い有効性が示された一方で、1)、3) については有効性が認められなかった。2) の仮説は、酒類やスナック菓子、カップ麺など、いくつかのカテゴリにおいて異なるバイクラスタが得られ、それらを一般化することで「免罪符的」という仮説を得るに至った。一方で3) については、ある一つのクラスタにのみ認められたものであり、十分な一般化ができていないとも考えられる。さらに訓練データが対象とするQPR上のモニタとテスト対象の店舗の顧客の性質の違いも相まって、有効性が認められなかったのかもしれない。

謝辞

本研究は、JST CREST(グラント番号: JPMJCR1401) の研究助成を受けている。株式会社マクロミルには消費者購買履歴データを提供いただいた。株式会社光洋の松下寿美子氏および内藤龍徳氏には仮説の導出から店頭実験まで多大な協力をいただいた。この場を借りて謝意を表する。

参考文献

[1] Dong,G. , Li,J., Efficient mining of emerging patterns: discovering trends and differences, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, p.43-52, 1999.

[2] 中原孝信, 大内章子, 宇野毅明, 羽室行信 (2016), 『データ研磨の2部グラフへの適用とTwitterからの意見抽出』2016年度人工知能学会 (第30回)

[3] 宇野毅明, 有村博紀, 浅井達哉 (2003) 『極大2部クリークの高速列挙法とデータマイニングへの応用』夏のLAシンポジウム 2003年7月