異常検知を利用したブランドスイッチ分析

片岡　弘貴，森田　裕之

1. はじめに

スーパーマーケットやドラッグストアなどの小売業では，顧客の購買履歴を記録したPOSデータが広く利用されている．POSデータには，個々の顧客の膨大な販売実績情報が時系列に記録されており，CRMを実施する上で有効な示唆を与えてくれる．集計した売行き動向だけでなく，顧客が購買ブランドをスイッチするタイミングを把握することができれば，他社ブランドを購買していた顧客を自社ブランドに取り込む，もしくは，自社ブランドの顧客が他社ブランドへ流出することを未然に防ぐことが可能である[1]．

近年，顧客のニーズが多様化し，それに対応するため企業は多種多様な商品を販売しており，商品選択における顧客の自由度が増大している．これは顧客にとって，ブランドをスイッチする動機の1つとなるので，特に類似した特徴を持つ商品に対して，顧客が特定のブランドを購入し続けることは稀である．また，ブランドスイッチを定義する際，今まで継続して購買しているブランド以外の商品を1度でも購買すればスイッチしたと考えると，その定義は簡単であるが，実際にデータを分析すると，一度または数度，他ブランドを購買しても元のブランドに戻る顧客も少なくなく，許容性のあるスイッチの定義を行うことも実際には必要であろうと考えられる．しかしその一方で，分析する前に許容性を加味したブランドスイッチの状態を明確に定義することは，よほどブランドスイッチに対して明確な考えがない限り容易ではない．ブランドスイッチを事前に定義することができれば，教師付き学習タイプの分類モデルを適用することができるが，上述のように事前にクラスを定義することが困難であれば，教師なし学習タイプの分類モデルを考慮する必要がある．

かたおか　ひろき，もりた　ひろゆき
大阪府立大学 大学院経済学研究科
〒599-8531 堺市中区学園町1-1
受付 11.7.25　採択 11.11.5

本研究では，まず顧客が購買する商品ブランドの順序に着目し，シークエンシャルパタンを列挙する．次に列挙したシークエンシャルパタンを，その出現頻度と対象ブランドの出現割合から，どのくらい正常なシークエンシャルパタン（以下，正常パタンという）なのか，または異常なシークエンシャルパタン（以下，異常パタンという）なのかをスコア化する．最後に顧客の数回分の購買行動中に，どの程度の正常または異常パタンが存在し，それによって顧客が保持するスコアがどれくらいかを計算することで，顧客の異常性（ブランドスイッチしやすさ）を発見するモデルを提案する．

提供されたデータをもとに，3つの商品種類における各3種類の9ブランドを対象として計算機実験を行い，その有効性を確認した．またある期間のデータに適用して学習した正常または異常パタンの情報を利用して，それ以降の購買データについてもオンラインで異常性を評価し，顧客のブランドスイッチをある程度の精度で予測できることを確認した．

以下では2節で関連する研究についてまとめ，3節で基礎分析の結果を示す．4節では提案手法について説明し，5節で実験結果について示す．6節では，実際の利用を想定したシミュレーション結果について示す．

2. 関連研究

教師なし学習は，クラスタリング，異常検知，そして次元縮約の3つの分野などでよく利用される方法である[2]．このうち異常検知では文献[3]などに様々な手法が紹介されているが，連続値データを対象とした手法が主であり，離散値を対象とした手法は少ない．離散値を対象とした方法では，例えばコンピュータの不正使用などにおいてのアプリケーションがあり，IPアドレスやコマンドの利用状態の頻出パタンから異常を検知するアルゴリズムなどが提案されている[4][5]．これらの方法は，頻出パタンを利用しているため，購買履歴データの購買順番などが考慮されない．したが

って直接利用することは困難であるが，パタンの概念を利用した異常検知の研究としては関連する研究であるといえる．

本研究で提案するシークエンシャルパタンのスコア化の方法は，CAEP（Classification by Aggregating Emerging Patterns）[6]のアイディアが参考になっている．CAEPは教師付き学習モデルであり，事前に決定したクラスデータを与えて学習するタイプである．まず与えたクラスで出現するクラス別の頻出パタンの支持度を計算する．ここで支持度は，対象とするデータベース中に，ある頻出パタンが該当するトランザクション数の割合とする．そこで，対象とするクラスの支持度を分子に，他のクラスの支持度を分母にした割合を成長率と呼び，その成長率の値が一定値以上の頻出パタンを顕在パタンと呼ぶ．この顕在パタンから各クラスを予測するスコアを算出し，そのスコアを集約してクラス判定をCAEPは行っている．我々の手法は，シークエンシャルパタンを利用している点や，予め定められたクラスを必要とはしない点など相違点はあるが，スコアを利用するフレームワークは類似している．

また，異常検知におけるスコア化を利用している方法としては，最近，文献[7]が報告されている．これは鉄道のメンテナンスにおけるアプリケーションであり，いくつかのセンサーがモニターする値を離散化し，その離散値の順序関係を保持したままシークエンシャルパタンを集約するとともに，正常パタンと異常パタンを，それぞれ支持度とセンサー値の離散化した値の大きさからスコア化して集約した値で異常性を判定している．正常パタンと異常パタンをそれぞれスコア化し集約結果から異常性を判定するフレームワークは，本研究でも同じであるが，購買履歴から時系列情報を取り入れたシークエンシャルパタンを列挙し，そのシークエンシャルパタンの支持度と，対象ブランドの出現比率から正常値と異常値のスコア化を行っている点が，本提案でのオリジナルであるといえる．

3. 分析対象データの概要と基礎分析

今回提供されたデータ[1]は，複数のチェーン店のドラッグストアのPOSデータである．データ期間は2008年1月1日から2009年12月31日までの2年間のデータであり，JICFSコードで日用品に分類されるものの中から，口中衛生用品（歯磨き，歯ブラシ等），衣料用洗剤類（衣料用洗剤，柔軟剤等），インバス・ヘアケア（シャンプー等）の3分類のデータであった．また，それぞれ約60万，約70万，約46万レコードのデータであった．データ項目には，アイテムコードごとに，JICFS名，メーカー名，ブランド名が含まれているが，商品名や定価データは無く，販売時点での販売価格のみが利用可能である．

データの基礎集計から，3種類の合計売上高は，約6億4千万円で，口中衛生用品が約30%，衣料用洗剤類が約30%，インバス・ヘアケアが残りの約40%を占めている．各データにおける利用者人数は口中衛生用品が91,479人，衣料用洗剤類が89,322人，インバス・ヘアケアが77,468人存在するが，その35～40%は1度だけの購買履歴しか存在しない．

例えば衣料用洗剤類の中から，JICFSコード衣料用合成洗剤類を抜き出して，顧客の購買ブランド種類数を分析すると図1のように，購買回数が多いと複数のブランドを購買する傾向がある．

このように，購買頻度の大きい日常消費財は，長期的に見て複数のブランドを購買する傾向があることがわかる．商品種類にもよるであろうが，今回のデータでは長期間，顧客が唯一のブランドに固執することは少なく，ある程度他のブランドの購買を許容した状態で，ブランドスイッチを考察することが現実的に必要であることがわかる．

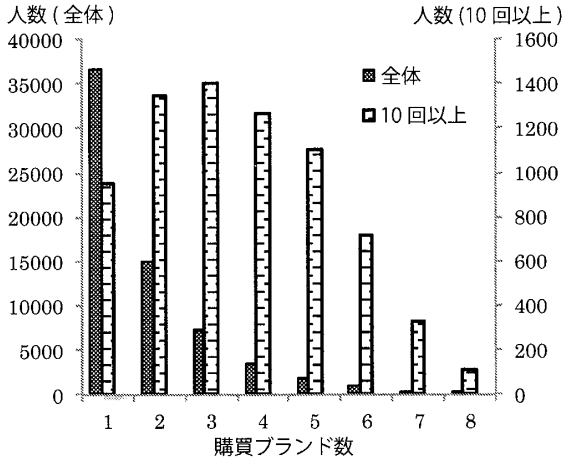

図1 購買回数による購買ブランド種類の差

[1] 平成22年度データ解析コンペティションにおいて，経営科学系研究部会連合協議会とカスタマー・コミュニケーションズ株式会社より提供されたデータ．

4. 分析の方針と手法の提案

本研究の目的は，あるブランドにロイヤルティのある顧客群から，ブランドスイッチと考えられる行動をする顧客を比較的短い購買回数で，事前に予測することである．ブランドスイッチを事前になるべく早く正確に予測することができれば，クーポン発行などで購買の継続を促すことや，スイッチ顧客の多い商圏を中心に，広告，宣伝活動を行うなど，適切なアクションを実行することが可能になる．

ブランドスイッチ分析で重要なポイントの1つは，その定義である．本研究では，特定のブランドにロイヤルティのある顧客が，単発で購買ブランドを変化することではなく，その後数回の購買においても他ブランド購買が継続するような状態をブランドスイッチと考える．そしてロイヤルティがある状態とは，ある回数の購買において，特定のブランドをk回（本論文ではkは1または2）以上購買する状態のこととする．前述の基礎分析で示したように，長期間購買する顧客は，今回対象となっている商品分類においては，複数のブランドを購買する傾向がある．したがって，ロイヤルティのある顧客がそのブランド以外の商品を購買した場合，それが，彼らの中で頻出するならば，特に危険視する必要はない．しかし，その行動が稀にしか起こりえない購買行動であるならば，何か異常な行動である恐れがあるので，注意すべきであると考える．そこで提案する手法には以下の要件が必要である．

A）顧客がどのブランドの商品をどのような順序で購買したかを表現できること

B）対象とした顧客群にとって一般的な購買行動と異常な購買行動を判別できること

これらの条件を満たすために，提案手法では顧客の購買ブランドのシークエンシャルパタンを利用する．シークエンシャルパタンは頻出パタンと異なり，同一アイテムの多重出現も含めて，どのアイテムがどのくらい出現したかというだけでなく，どのような順番で発生したということを表現することが可能である．例えば発生順序（購買機会）データを含む表1のような，ある顧客の購買履歴データが存在したとする．ここでは3回目の購買で，同時に複数のアイテムを購買している．

このとき列挙される頻出パタンは，｛A, AB, ABC, AC, B, BC, C｝となり，シークエンシャルパタンは，｛A, AA, AB, ABA, ABC, AC, B, BA, BC, C｝となる．

表1 購買履歴データの例

顧客ID	順序	アイテム
1	1	A
1	2	B
1	3	A
1	3	C

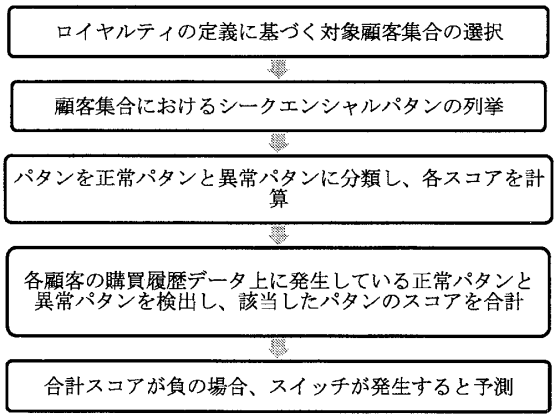

図2 提案手法のフロー図

このように頻出パタンでは，AAやBAなどは列挙されないため，発生順序や重複出現が分析上意味を持つ場合，シークエンシャルパタンで表現することが有用であることがわかる．このようなシークエンシャルパタンを利用した提案手法の流れは図2のようになる．

あるブランドにロイヤルティのある顧客群を対象として，シークエンシャルパタンを列挙した場合，その支持度が平均的であれば，当該顧客群に一般的な行動パタンであり，逆に支持度が小さいか，または大きければ，特殊な行動パタンと考える．加えて，列挙されたシークエンシャルパタンにおける当該ブランドの出現割合が大きければ，支持度の大小に関わらずロイヤルティを維持するパタンであろうし，その割合が小さければロイヤルティに反するパタンであると予想される．この仮定のもとに，各シークエンシャルパタンをスコア化する．

あるブランドにロイヤルティのある顧客群の購買行動データから列挙されたシークエンシャルパタンをp（$p \in P \mid P$は全ての列挙されたシークエンシャルパタンの集合）とし，$sup(p)$は支持度を示しているとする．ここである平均的なシークエンシャルパタン出現支持度SP_{th}を与えると，顧客に与える影響の度合い$D(p)$をSP_{th}からの乖離度を利用して式(1)のよう

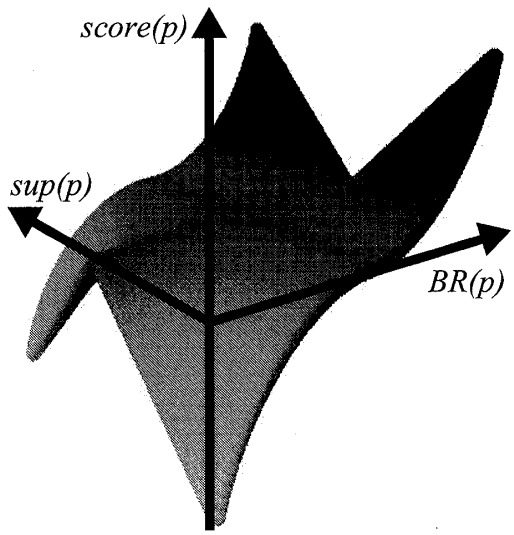

図3 $score(p)$ の分布の一例

に定める．

$$D(p)=|sup(p)-SP_{th}| \tag{1}$$

ここで SP_{th} は，$sup(p)$ が 0〜1 まで一様に分布していれば，その間のどこかになるが，シークエンシャルパタンの出現状況に依存する．そこで最小支持度を設定してシークエンシャルパタンを列挙し，それらの支持度の状況から設定する方が望ましい．

次に，あるシークエンシャルパタン p に含まれる対象のブランドの割合を $BR(p)$ とする．支持度と同様に，ある平均的な $BR(p)$ の基準値 BR_{th} を考える．また，与えられた BR_{th} からの乖離度が，ブランドの継続かスイッチに与える影響を示す値 $E(p)$ を考える．$E(p)$ はエントロピーの考え方を応用したもので，$BR(p)$ が BR_{th} のときに最小値 0 となるようにする．式(2)に基づいて $cBR(p)$ を算出し，求めた $cBR(p)$ を $cBR_1(p)$ として，$cBR_2(p)=1-cBR_1(p)$ とする．そして，$E(p)$ を式(3)に基づいて算出する．ただし，$BR(p)=BR_{th}$ のとき，シークエンシャルパタン全てが 0 の値となり，まったく評価されなくなる．ここでは，より一般的に b_{val} を導入して，任意の値（0 または比較的小さな正の実数）を与えることにする[2]．

$$cBR(p)=\begin{cases}\dfrac{BR(p)-BR_{th}}{2(1-BR_{th})}+\dfrac{1}{2} & (BR(p)\geq BR_{th}),\\ \dfrac{BR(p)}{2BR_{th}} & (BR(p)<BR_{th}),\end{cases} \tag{2}$$

[2] b_{val} は 0 でも問題ないが，予備実験の結果から，0 よりはごく小さな値を与えた方が良い結果を示す傾向がみられた．以降の実験においては，$b_{val}=0.01$ を設定して計算している．

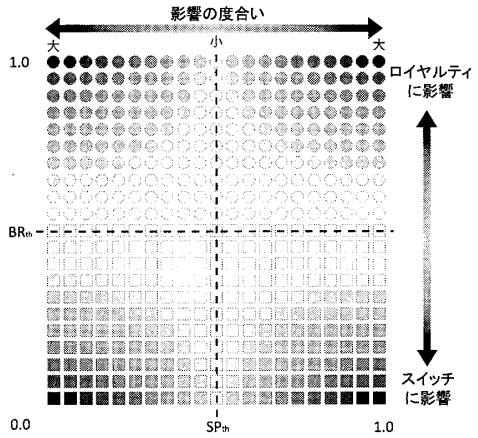

図4 パタンの $BR(p)$ と $sup(p)$ の関係

$$E(p)=1+\sum_{i=1}^{2} cBR_i(p)\times\log_2 cBR_i(p)+b_{val} \tag{3}$$

以上より，各シークエンシャルパタンのスコアと各顧客のスコアを，式(4)，(5)に基づいて算出する．

$$score(p)=\begin{cases}D(p)\times E(p) & BR(p)\geq BR_{th}\\ -1\times D(p)\times E(p) & BR(p)<BR_{th}\end{cases} \tag{4}$$

$$SC(u)=\sum_{p\in P_u} score(p) \tag{5}$$

ここで u は顧客 ID，P_u は，顧客 u のトランザクションに出現するシークエンシャルパタンの集合を表すとする．そして $SC(u)<0$ の場合，u は異常な状態と考える．支持度とブランドの割合によるスコアの関係のイメージは図 4 になる．○は正常なスコアであり，□は異常なスコアである．色が薄いほど 0 に近づき，色が濃いほどそれぞれ正常，異常に影響の強い値となる．

5. 計算機実験とその結果

提案手法を提供されたデータに適用するに際し，購買回数やブランドの競合状態から，分析対象ブランドとして，衣料用合成洗剤，柔軟剤，シャンプーに分類されるブランドの売上金額上位 3 位までを選択する．また，ロイヤルティ有無の条件，シークエンシャルパタン作成のための購買回数の確保，検証期間の確保のために，データ期間中，各商品種類を 12 回以上購買した顧客に限定する．次に，ロイヤルティ確立の条件を以下のように 2 種類設定する．

実験 1：初回の購買が対象ブランドである場合

実験 2：初回から 3 回目までの購買中 2 回以上対象ブランドを購買した場合

ロイヤルティ確立の条件が緩い実験 1 と，確立の条

件を厳しくした実験2の2つの異なる条件を設定した場合で，提案モデルの結果の違いを考察する．

まず実験1では，上述のロイヤルティ確立の条件に基づき，2回目から8回目までの7回の購買期間からシークエンシャルパタンを作成して，提案手法を適用する．適用手法によってブランドをスイッチすると予測される顧客が識別されるので，これを9回から12回目までの4回の購買で検証する．検証期間での予測結果との比較は，検証期間4回のうち対象ブランドをm回以下しか購買しない，という基準（以下，スイッチ基準と呼ぶ）で異常と，また$m+1$回以上購買すれば正常（異常以外を正常と考えることにする）と評価する．

計算機実験を行うパラメータSP_{th}およびBR_{th}については，与えられたデータからまったく学習できない場合，予備実験などの結果から初期値として，シークエンシャルパタンの$sup(p)$と$BR(p)$の分布のそれぞれ第3四分位点を与えることにする．また，この初期値を出発点としてパラメータを学習できる状況にあれば，変更することは可能である．もちろんその際も予測される異常の状況をどのように定義するかによって学習される値は異なるが，可能な範囲でパラメータを調整させればよい．以降の実験においては，上述の初期値をパラメータとして与えて計算した結果と，それぞれの基準でパラメータを調整した場合の結果を示し，その後パラメータを調整した場合のより詳しい計算結果を考察することにする．

表2は，実験1において，初期値のパラメータでの

表2 パラメータの違いによる比較結果[実験1]

スイッチ基準	学習無しF1値		学習ありF1値	
	平均	標準偏差	平均	標準偏差
0	0.495	0.061	0.650	0.059
1	0.558	0.056	0.757	0.039
2	0.597	0.060	0.808	0.037
3	0.600	0.076	0.852	0.036

表3 対象顧客数とパラメータ初期値[実験1]

	対象顧客数	初期値(SP_{th})	初期値(BR_{th})
アタック	735	0.007	0.500
アリエール	758	0.007	0.500
トップ	492	0.010	0.500
ハミング	743	0.011	0.600
レノア	701	0.010	0.500
ソフラン	541	0.015	0.500
ラックス	287	0.014	0.667
パンテーン	121	0.033	0.667
TSUBAKI	83	0.036	0.667

表4 提案手法と回数ベースのパラメータ[実験1]

パラメータ	提案手法		回数ベース
	SP_{th}	BR_{th}	回数
基準0	0.20	0.50	1
基準1	0.40	0.75	3
基準2	0.45	0.85	5
基準3	0.50	0.95	6

表5 各ブランドに対する計算結果[実験1]

ブランド名	スイッチ基準	提案手法F1値	回数ベースF1値	スイッチ基準	提案手法F1値	回数ベースF1値
アタック	0	0.59	0.60	2	0.83	0.84
アリエール	0	0.68	0.68	2	0.82	0.84
トップ	0	0.63	0.61	2	0.85	0.87
ハミング	0	0.59	0.53	2	0.75	0.70
レノア	0	0.68	0.54	2	0.79	0.74
ソフラン	0	0.56	0.55	2	0.78	0.77
ラックス	0	0.72	0.64	2	0.82	0.84
パンテーン	0	0.71	0.64	2	0.77	0.77
TSUBAKI	0	0.69	0.60	2	0.86	0.86
アタック	1	0.77	0.69	3	0.88	0.88
アリエール	1	0.78	0.76	3	0.85	0.85
トップ	1	0.78	0.75	3	0.91	0.91
ハミング	1	0.69	0.63	3	0.81	0.74
レノア	1	0.74	0.67	3	0.81	0.72
ソフラン	1	0.72	0.67	3	0.84	0.79
ラックス	1	0.79	0.75	3	0.86	0.85
パンテーン	1	0.73	0.73	3	0.83	0.81
TSUBAKI	1	0.81	0.74	3	0.88	0.89

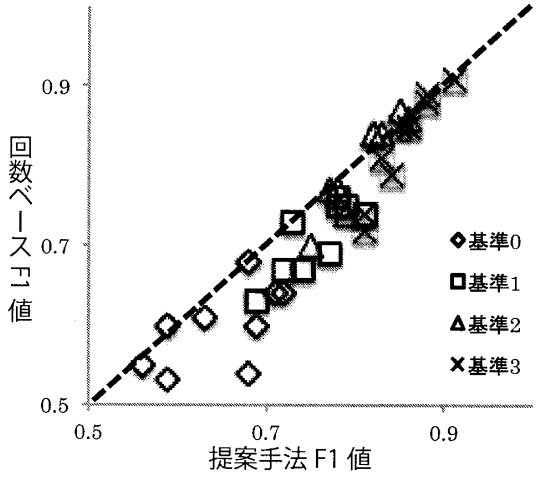

図5 提案手法F1値と回数ベースF1値の比較[実験1]

表6 パラメータの違いによる比較結果[実験2]

スイッチ基準	学習無しF1値		学習ありF1値	
	平均	標準偏差	平均	標準偏差
0	0.616	0.050	0.644	0.081
1	0.614	0.085	0.758	0.061
2	0.564	0.093	0.818	0.047
3	0.505	0.090	0.861	0.034

表7 対象顧客数とパラメータ初期値[実験2]

	対象顧客数	初期値 (SP_{th})	初期値 (BR_{th})
アタック	722	0.010	0.500
アリエール	604	0.010	0.600
トップ	418	0.014	0.500
ハミング	724	0.017	0.600
レノア	705	0.016	0.600
ソフラン	481	0.023	0.500
ラックス	286	0.017	0.667
パンテーン	117	0.034	0.667
TSUBAKI	72	0.042	0.667

表8 提案手法と回数ベースのパラメータ[実験2]

パラメータ	提案手法		回数ベース
	SP_{th}	BR_{th}	回数
基準0	0.25	0.45	0
基準1	0.40	0.70	1
基準2	0.50	0.75	2
基準3	0.50	0.95	3

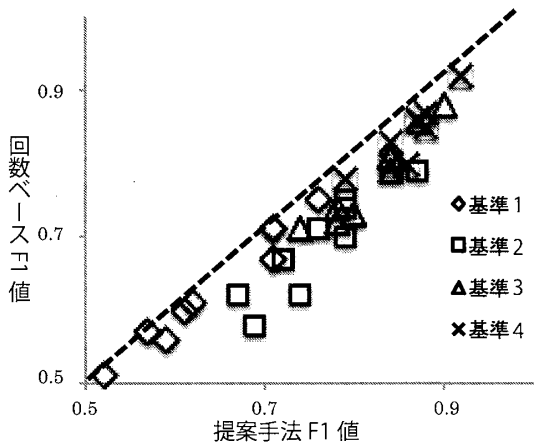

図6 提案手法F1値と回数ベースF1値の比較[実験2]

表9 各ブランドに対する計算結果[実験2]

ブランド名	スイッチ基準	提案手法F1値	回数ベースF1値	スイッチ基準	提案手法F1値	回数ベースF1値
アタック	0	0.52	0.51	2	0.78	0.72
アリエール	0	0.57	0.57	2	0.74	0.71
トップ	0	0.59	0.56	2	0.80	0.73
ハミング	0	0.61	0.60	2	0.78	0.74
レノア	0	0.71	0.67	2	0.84	0.81
ソフラン	0	0.62	0.61	2	0.84	0.81
ラックス	0	0.71	0.71	2	0.84	0.81
パンテーン	0	0.71	0.71	2	0.84	0.79
TSUBAKI	0	0.76	0.75	2	0.90	0.88
アタック	1	0.69	0.58	3	0.84	0.83
アリエール	1	0.67	0.62	3	0.79	0.78
トップ	1	0.74	0.62	3	0.87	0.86
ハミング	1	0.72	0.67	3	0.84	0.80
レノア	1	0.79	0.70	3	0.86	0.80
ソフラン	1	0.76	0.71	3	0.88	0.85
ラックス	1	0.79	0.74	3	0.87	0.86
パンテーン	1	0.79	0.75	3	0.88	0.87
TSUBAKI	1	0.87	0.79	3	0.92	0.92

計算結果と各スイッチ基準におけるF1値の平均値が，最良値になるように調整したパラメータを使った計算結果を比較したものである．ブランドごとに計算したF1値を，スイッチ基準ごとの平均と標準偏差を示している．ここでF1値は，スイッチと予測した顧客のうち的中した割合である精度（Precision）と，実際にスイッチした顧客のうち的中した割合である再現率（Recall）の重みを1対1とした調和平均である（最大値は1となる）．また表中のスイッチ基準0は，4回の検証期間中一度も購買がなかった顧客をスイッチと考えた場合を意味している．表2より，調整した結果の方がパラメータを初期値のまま調整しない場合より良い結果であるが，初期値を使ってもある程度の結果を示せていることがわかる．また，初期値の場合の標準偏差は小さな値を示していることから，ブランドによってF1値の変動がそれほど大きくないこともわかる．なお，表3は各ブランドにおける実験対象顧客数とパラメータの初期値である各第3四分位点の値を示している．

表10 実験1と実験2の比較

	実験1	実験2
平均改善値*	0.032	0.036
改善値標準偏差	0.039	0.030

(*改善値は提案手法F1値－回数ベースF1値)

次に表5は，表4で調整したパラメータによる各ブランドの計算結果を示している．表中の回数ベース

F1値は，シークエンシャルパタン作成の購買期間（7回の購買）に，対象ブランドを購買した回数によってスイッチ顧客を分類した場合のF1値を比較参考とするために表している．なお表4に示すように，回数ベースのF1値も，各スイッチ基準のF1値の平均値が最良値になるようにスイッチと判断する回数を調整している．

この結果から，提案手法，回数ベース共に学習期間があり，パラメータを調整できれば，提案手法は多くのブランドに対して，単純な回数ベースでのスイッチの予測よりも精度の高い予測が可能であることがわかる．

図5は各基準での提案手法F1値と回数ベースF1値を比較した散布図である．斜線は互いのF1値が等しくなるラインを表しており，分布した点が斜線より提案手法F1値側にあるということは，提案手法F1値が同じ基準において回数ベースF1値よりも結果が良いことを示す．全体として提案手法は，多くの基準，多くのブランドで回数ベースのモデルよりも良い結果を示していることがわかる．

次に実験2では，前述のようにロイヤルティ確立条件を最初から3回のうち2回以上対象ブランドを購買している顧客に限定し，4回目から8回目までの5回の購買期間からシークエンシャルパタンを作成する．実験1と同様に，まず初期値のパラメータを利用した場合と，調整したパラメータを利用した場合の結果を比較する．

実験1と同様に，第3四分位点を指定することによって，一定の結果を得られることがわかる．また表8のようにパラメータを調整することで，表9および図6の結果を得ることができる．

表10は調整したパラメータを利用した場合の実験1と実験2の集計結果の比較である．この結果からは，それほど大きな違いは確認できない．したがって今回のデータから定義した実験1と実験2に関しては，ブランドロイヤルティの確立の条件を厳密に決める必要がないことがわかる．次節では，実験1と2の結果に大きな違いがなかったので，若干全体的な結果の良かった実験2に限定して，また，洗剤ブランドのトップを対象としてシミュレーションを行い，結果について考察する．

6. 応用のためのシミュレーション

提案手法を実際に利用する際には，いくつかの方法が考えられるが，以下ではその1つを想定してシミュレーションを行う．データ期間2年間のうち，最初の1年半の購買行動履歴から，実験2と同条件でスイッチ予測の対象顧客を抽出し，提案手法により計算を行う．その際，最初の1年半で学習されたシークエンシャルパタンとそのスコアを保存する．そして，以降の半年間において実験2のロイヤルティ確立条件に該当する顧客を抽出し，それらの顧客の各購買時における集計スコアを，保存したシークエンシャルパタンとそのスコアから計算して，顧客のスイッチを予測する．具体的には，各顧客の4回目以後の購買のたびに，顧客のトランザクションデータから一致する学習されたシークエンシャルパタンを照合する．その照合されたシークエンシャルパタンが持つスコアの全体の集計値をある購買時点における顧客が保持するスコアと考え，それが負の値になった場合にスイッチすると判断する．スイッチと判断された場合，それ以降6回の購買で他ブランドを3回，または4回以上購買する場合は実際にスイッチしたと判断する．

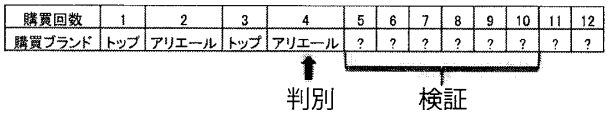

図7 判別と検証の一例

なお比較対象として，ある購買時点までに対象ブランド以外のブランドを購買した場合はスイッチと判断する単純なモデルを考える．

表11 各モデルでの予測結果

		提案モデル		単純モデル	
対象外ブランド購買回数		3回以上	4回以上	3回以上	4回以上
4回目	F1値	0.846*	0.880*	0.706	0.727
	Recall	0.786	0.846	0.857	0.923
	Precision	0.917	0.917	0.600	0.600
5回目	F1値	0.813*	0.828*	0.789	0.743
	Recall	0.765	0.857	0.882	0.929
	Precision	0.867	0.800	0.714	0.619
6回目	F1値	0.722	0.727	0.850*	0.757*
	Recall	0.684	0.750	0.895	0.875
	Precision	0.765	0.706	0.810	0.667

その結果をまとめたものが表11である．表中の＊がついているのは，提案モデルと単純モデルを比較した際，スイッチ基準ごとによいF1値の結果を表す．全体としては提案モデルのほうがよいが，6回目につ

いては単純モデルの結果がよいことがわかる．その原因を調べたところ2つの点が明らかになった．1つは，単純モデルは判断する購買機会を長くすれば，スイッチと予測する顧客数が単純増加する．つまり，1度も対象ブランド以外を購買しなかった顧客以外は，購買機会をとても長くすると，すべてスイッチと判断することになる．これはrecallを1にすることになるが，precisionは最終的に全顧客数に対するスイッチと決定される顧客数となる．今回，その顧客数が比較的多く，また6回では偶然precisionも高かったため，F1値が大きくなっていることが分かった．逆に，提案手法は購買機会によって判断が変わる点は有効であるが，顧客が商品を併買するとき，少し問題があることがわかった．

表12 併買により誤判別された顧客

ID	1回目	2回目	3回目	4回目	5回目	6回目
82477	トップ	トップ	ボールド	トップ ボールド	トップ	ボールド アリエール
111169	ボールド	ボールド	トップ	トップ ブルーダイヤ ニュービーズ	トップ ボールド	トップ

表12はそのような顧客の購買の例であるが，4回目などにいくつかの商品を併買していることがわかる．今回は，併買であっても，該当するシークエンシャルパタンが存在すれば，スコアを集計しているため，対象ブランド以外のブランドが併買されると，少し強すぎる異常スコアを集計してしまう傾向が確認された．このように，一般的には提案モデルのほうが有効であろうと考えられるが，併買が存在する場合には，今後，スコアの集計に少し工夫を行う必要があるかもしれない．

また図8は，表11で計算した結果のうち，実際にスコアが購買機会とともにどのように変化するかを表した一例である．これは，4回目の購買でスイッチと判別されたある顧客の例である．4回の購買中2回「トップ」を購買しているので，検証期間の6回中に3回以上「トップ」を購買するとも判断できるが，その購買の異常性（スコアの低さ）から，スイッチ顧客に判別されている．実際5回目以降に「トップ」を購買していないので，この判別は正しく機能していたことがわかる．

以上のように，提案モデルは，過去1年半のデータを用いて，その後の半年間において，図8の顧客のように，単純に対象ブランド購買回数だけでは判別できないような顧客を予測することを可能としている．また，このシミュレーションでは，学習期間において列挙したシークエンシャルパタンと，計算したスコアを用いるため，顧客の購買回数が増加してもオンラインで順次対応することが可能である．つまり，既存顧客の購買行動によるスコアの変化だけでなく，新規顧客に対しても購買データが追加されていく時点において，特定ブランドに対するスイッチを予測するスコアを計算し，判断することが可能である．したがって，もしこのモデルがレジシステムの一部に組み込まれていたとすると，ある顧客がレジを通過する際に，その顧客の対象ブランドに対するスコアを計算し，異常スコアを持てばその顧客がスイッチすると予測する．スイッチすると予測されれば，レジ通過中にあらかじめ定めておいたクーポンを発行するなどの手段によって，より早く未然にブランドスイッチを防止するといった応用が可能である．

購買回数	ブランド
1	トップ
2	アリエール
3	トップ
4	アリエール
5	アタック
6	ブルーダイヤ
7	さらさ
8	その他
9	アリエール
10	アタック

4回の購買までに出現したシークエンシャルパタン	スコア
アリエール アリエール	-0.396
アリエール トップ	-0.026
トップ アリエール	-0.024
トップ トップ	0.208
アリエール トップ アリエール	-0.101
トップ アリエール アリエール	-0.093
トップ アリエール トップ	-0.001
トップ トップ アリエール	-0.001
トップ アリエール トップ アリエール	-0.030

図8 特徴的なスイッチ顧客

7. まとめ

本稿では，ドラッグストアのPOSデータに対して，顧客の購買履歴データからシークエンシャルパタンを列挙して，その支持度と，含まれるブランドの割合の2つの指標からシークエンシャルパタンの正常と異常のスコアを決定し，そのスコアから顧客の異常性を分析した．異常をブランドスイッチと考えることで，ブランドスイッチに影響を与えるシークエンシャルパタンを発見し，将来的にブランドスイッチをする顧客の予測をするモデルを提案した．

今回の提案モデルでは，顧客の属性情報や，購買店舗のデータなど，顧客の嗜好をより正確に分類できる可能性のあるデータは，顧客の購買回数を確保する観点からあまり工夫できなかった．これを利用できれば，特徴的な顧客群を特定することが可能であるので，よ

り精度の高い結果を出すことも期待できる．また6節で考察したように，併買について，今回は特にスコア集計上の工夫を与えていなかったが，同じシークエンシャルパタンであっても，単に1商品の購買の結果として該当する場合と，いくつかの商品を併買する結果として該当する場合では少し状況が異なるのかもしれない．基本的なフレームワークの変更の必要はないかもしれないが，スコアの集計などに際しては併買かどうかによって修正をかけるほうが，より正確な予測に近づく可能性が考えられる．以上のような点を踏まえながら，今回は，1つのデータに対して適用した結果であるので，汎用性を持たせるためにも，少し傾向の異なるデータなどにも適用し，より汎用的で精度の高い手法として改善していきたい．

謝辞　本研究では経済産業省「情報大航海プロジェクト」の“パーソナル情報保護・解析基盤開発・改良と検証”において開発されたプログラムを一部使用している．また，国立情報学研究所 宇野毅明先生が開発されたプログラムも使用させていただいている．ここに深謝の意を表する．

参考文献

[1] 坂巻英一，齋藤俊則，多項ロジットモデルを用いた消費者のブランドスイッチ行動予測モデル構築法の提案，経営情報学会，第15巻，第2号，pp. 23-38, 2006.

[2] X. Zhu and A.B. Goldberg, Introduction to Semi-Supervised Learning, Morgan & Claypool, 2009.

[3] 山西健司，データマイニングによる異常検知，共立出版，2009.

[4] R. Vaarandi, Real-Time Classification of IDS Alerts with Data Mining Techniques, Proceedings of 2009 MILCOM Conference, pp. 1786-1792, 2009.

[5] R. Vaarandi and K. Podins, Network IDS Alert Classification with Frequent Itemset Mining and Data Clustering, Proceedings of the 2010 IEEE Conference on Network and Service Management, pp. 451-456, 2010.

[6] G. Dong, X. Zhang and L. Wong, Caep : Classification by aggregating emerging patterns, Proceedings of the International Conference on Discovery Science, pp. 30-42, 1999.

[7] J. Rabatel, S. Bringay and P. Poncelet, Anomaly Detection in Monitoring Sensor Data for Preventive Maintenance, Expert Systems with Applications, 38, pp. 7003-7015, 2011.