属性データとWEB閲覧履歴データが混在したオンラインショップデータに対する行動予測モデルの構築

大阪府立大学大学院
経済学研究科　博士前期課程2年
西口　真央

目次

- 研究動機
- 当該データの業界分析とデータの基礎分析
- 課題解決のための分類予測問題と研究目的
- 分析手法の基礎的内容と関連研究
- 提案手法
- 計算機実験と結果
- まとめ及び今後の課題
- 主要参考文献一覧

研究動機

近年、情報通信技術（ICT）の進化に伴い、利用されるデータの形式も多様化している。そしてそれらが複合的に蓄積され、分析される環境が整ってきている。例えば、BtoCにおけるオンラインショップでは、主に、顧客の年齢や職業などの顧客属性データ（以下、属性データという）と、各顧客がどのページを閲覧したか、そして何を購買したかを記録している閲覧・購買履歴データ（以下、ログデータという）が蓄積可能となり、マーケティングへの活用が課題となってきている。

このうち属性データは、従来より様々な解析手法で利用されてきている、トランザクションキーとカテゴリ的または数値的なデータ形式であるが、ログデータは、トランザクションキーに順序キーが加わった2つのキーとカテゴリデータを持つデータ形式である。そのため、順序キーを無視すれば、順序に関係なくどのページを閲覧したかというデータとして従来の手法を適用することは可能であるが、順序が持つであろう説明力を失うことになる。一方、ログデータを対象とした分析手法として、頻出系列パタンを列挙してクラス予測モデルを作成するCAESP[1]などが考えられるが、属性データも同時にモデルに組み込むことは、そのままでは難しい。

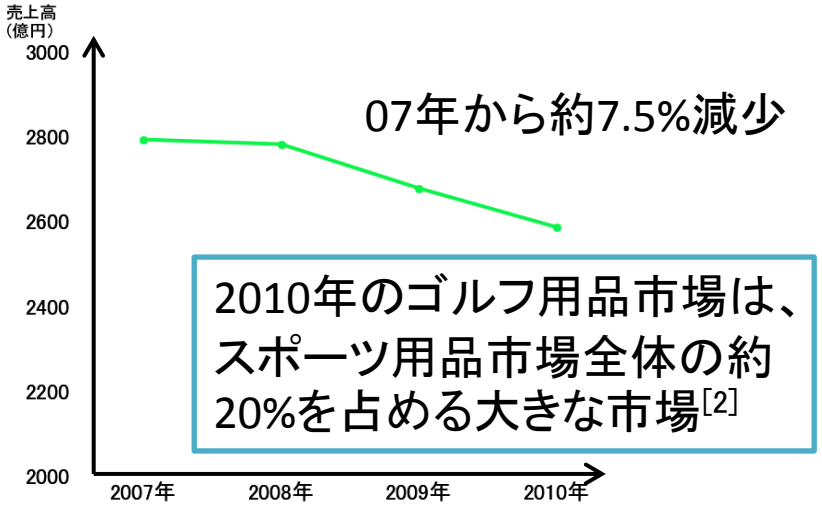
この問題を解決するため、本研究では、実際のあるオンラインショップを対象に、属性データとログデータが混在したデータセットとして与えられたとき、これら形式のデータを統合的に処理することが可能なクラス予測モデルを提案する。また当該データに手法を適用して、実用的な結果が入手可能であることを示す。

業界分析とデータの基礎分析

以下では、当該データが関係するゴルフ用品市場の現状と動向を分析した後、実際のデータに関する基礎集計を行う。2つの分析結果から、分析目的として2回以上の顧客の購買経験が重要であることを確認し、それを促進するための2つの分類予測問題を設定する。

利用可能なデータから、この分類予測問題に適した解法を提案し、計算実験を行うとともに、計算結果からマーケティング上の重要な示唆をマイニングすることが目的であることを示す

ゴルフ用品市場の現状と動向



ゴルフ用品市場売上高推移

2010年の国内B to Cの電子商取引（EC）市場規模は約7.8兆円で、年々増加傾向[3]

＜ショップオンライン化のメリット・デメリット＞

メリット	デメリット
商品を24時間265日、地理的制限なく販売可能	顧客が商品の実物を確認できないため、高価な商品の購買意思決定には不向き
アクセスログを蓄積することで、誰が、いつ、どのページを閲覧し、どの商品を購入したかなどの詳細な行動データを取得可能	Face to Faceのコミュニケーションによる感情データなどは得られない

市場規模は大きいが、飽和状態にあり、縮小傾向にある

EC市場は増加傾向で、そのメリットを活用する必要がある

ゴルフオンライン市場のシェアを適切に確保する必要がある

適切な顧客のセグメントを識別し、必要なプロモーションを実施することが重要

提供データ概要

あるゴルフショップサイトの会員の属性データ、ログデータ、商品受注データ

データ期間	2010年7月1日～2011年6月28日（約1年間）
会員数	アクセス有り=4053人（うち購買経験あり1352人）、アクセスなし=12063人
イベント数	約156万件(セッション数:約12万件) (セッション内のページ遷移がイベント)
分類	中分類数：5分類、小分類数：48分類（クラブ、用品・小物、ウェアなど）
商品数	3516点（受注実績が確認可能なもの）

＜使用可能な属性データの項目＞

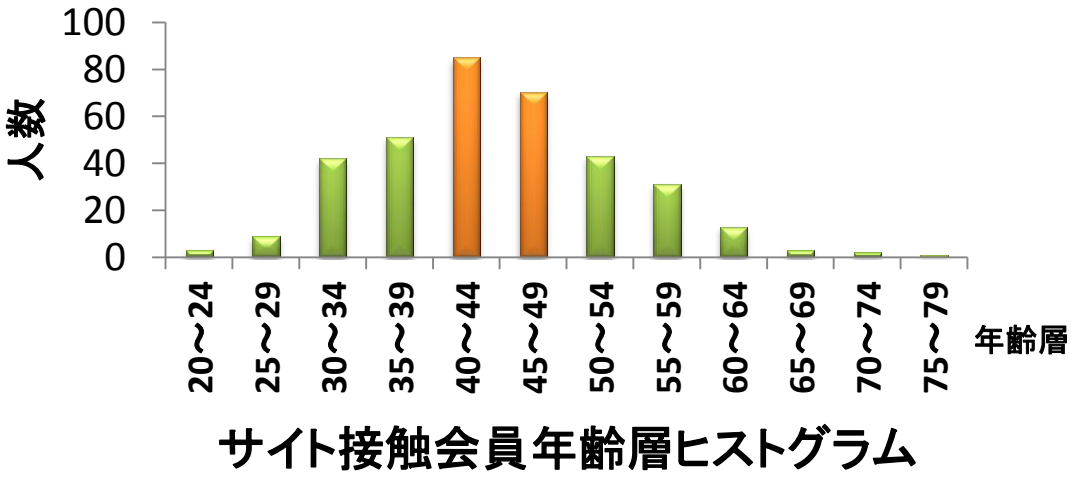
項目	説明
顧客ID	顧客の識別番号。今回はトランザクションキーとして使用
年齢	顧客の年齢。20～70歳
性別	顧客の性別。男性、女性の2値変数
ハンディキャップ	顧客のゴルフスコアハンディキャップ値。無記入者はスコア0
メルマガ購読の有無	メールマガジンの定期購読者であるかどうかの2値変数

＜使用可能なログデータの項目＞

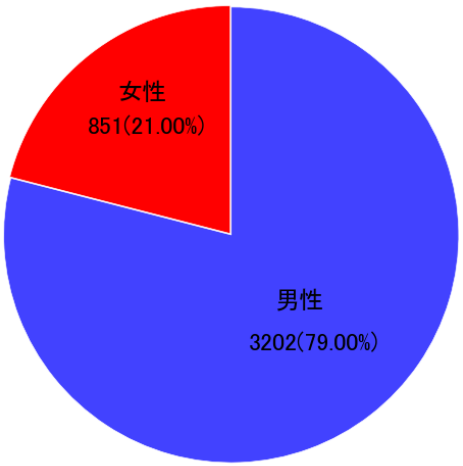
項目	説明
顧客ID	顧客の識別番号。今回はトランザクションキーとして使用
セッションID	一度ログインしてから、ログアウトするまでを1セッションとする
受注ID	受注の識別番号
アクセスページ	閲覧ページ名（コンテンツは不明）
時刻	イベント（ページの移動）が発生した時刻ごとに記録される
購買フラグ	受注が完了有無

※平成23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株式会社ゴルフダイジェスト・オンライン(GDO)より提供されたデータ

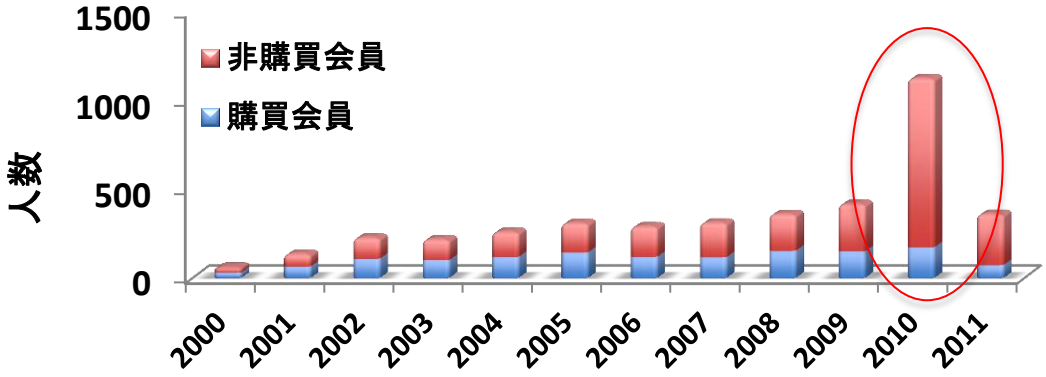
属性データの基礎分析



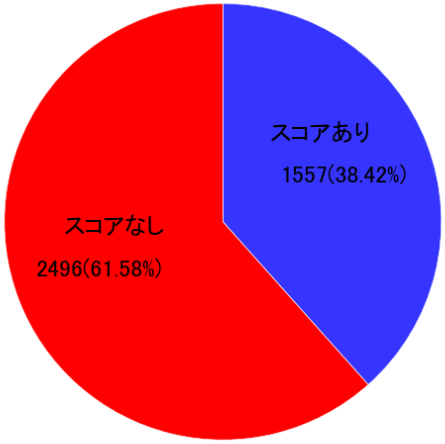
サイト接触会員年齢層ヒストグラム



サイト接触会員男女比



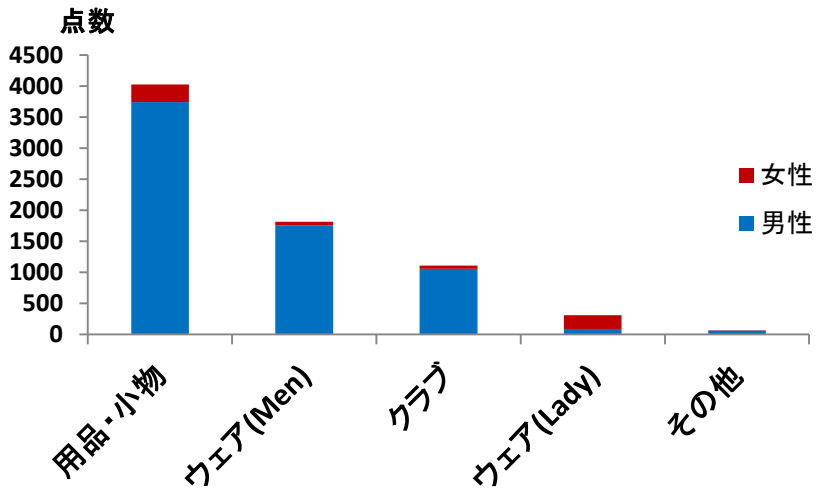
年別アクセス会員登録数



スコアハンディキャップの有無

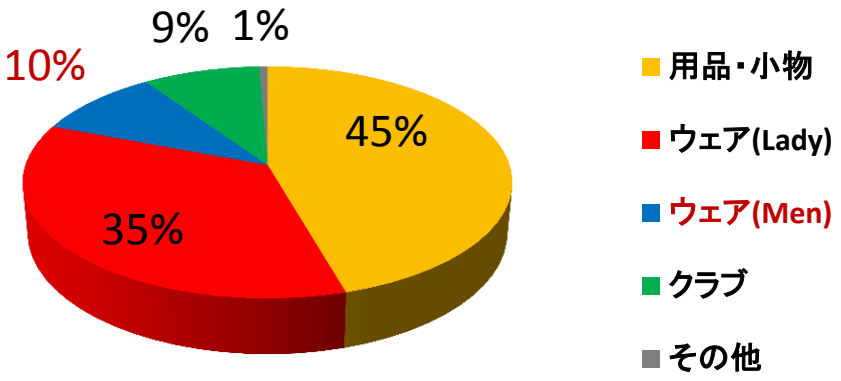
- 顧客の中心は30代後半から40代
- 2010の登録会員が多い（←お試し利用者）
- 約80%が男性
- 過半数はスコアなし顧客（←ゴルフ初心者or低関与）

購入商品の基礎分析

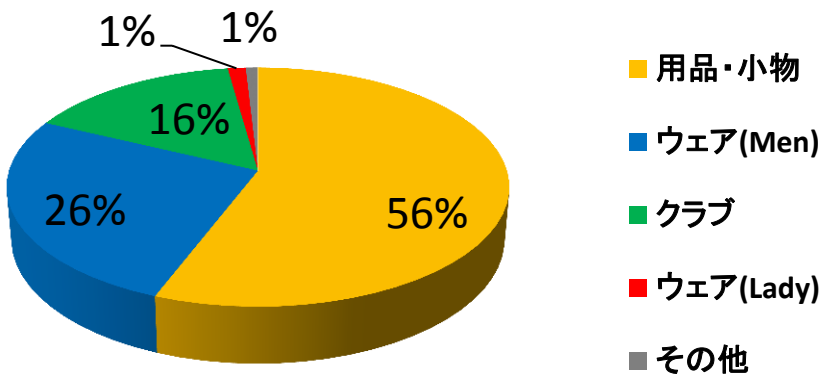


男女別購入商品内訳

- 用品・小物の購入が多い
  –少額商品だが、継続購買を狙えるチャンス
- 女性でメンズウェアを購入する割合は、男性がレディスウェアを購入する割合より明らかに高い
  –プレゼント利用など、女性の購買目的には興味深いポイントが存在している

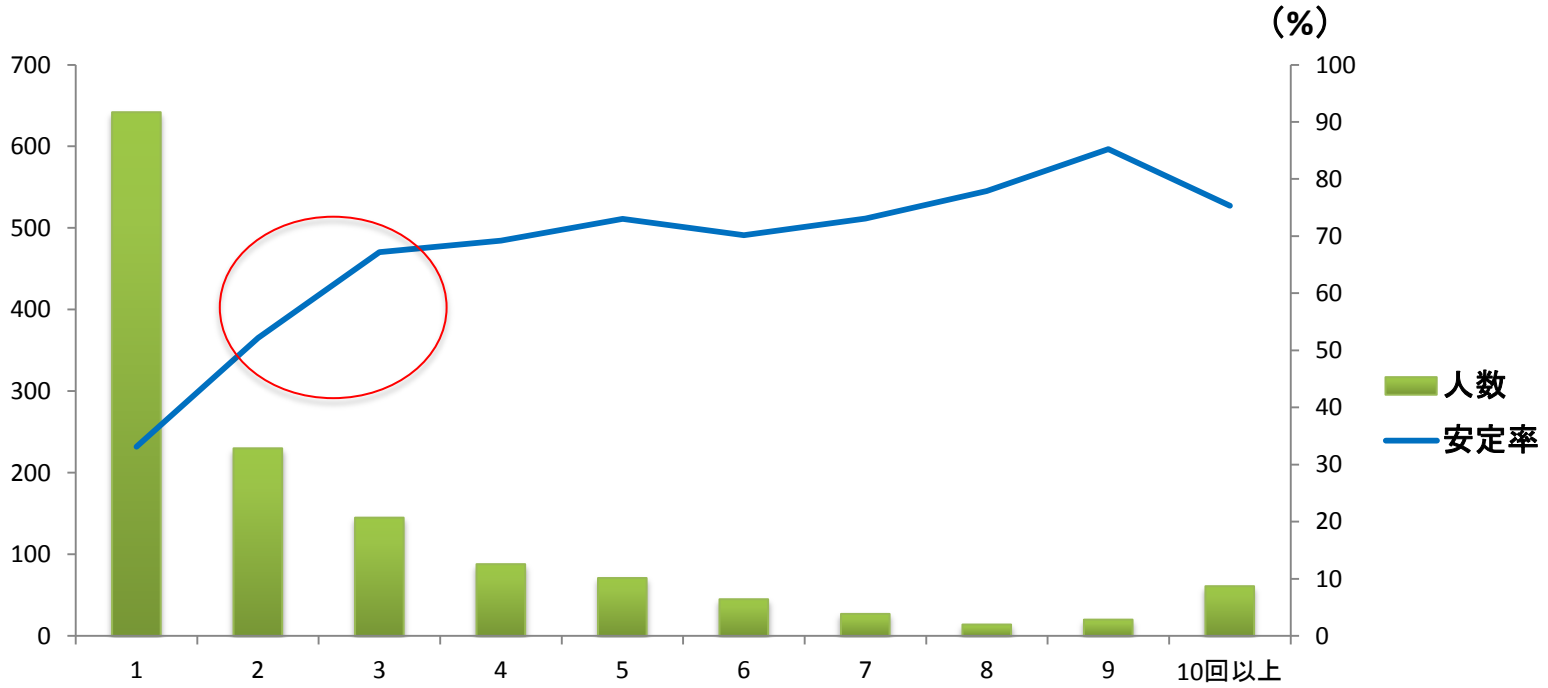


女性会員購入商品内訳割合



男性会員購入商品内訳割合

購入実態の基礎分析



購入回数分布と安定率推移

安定率=k回以上購入者数/(k-1)回以上購入者数

1回しか購入経験のない顧客と、2回購入経験のある顧客では、それ以上の利用回数の割合に大きな差がある

さしあたって2回のオンライン購入を経験していただき、その利便性と安心感を享受していただくことが、リピータ獲得に重要

分析目的の設定

継続利用顧客を増大することで、
リテール事業の売上改善を目指す

安定購買顧客層の目安を50%と考えると、2回以上の購買を達成してもらうことが1つのハードルとなっている

いかに2回購買経験をしてもらうか、またそのために1回目の購買をどうやって経験してもらうかが分析上重要な課題となる

1回しか購入していないユーザに2回目の購入経験を促すプロモーション

アクセスはしているが、購入経験のないユーザに、購入を経験してもらうためのプロモーション

＜当該データに対する課題解決のステップ＞

課題解決のための分類予測問題と研究目的

前述の課題解決ステップのため、以下の2つの分類予測問題を設定する

分析Step1:未購入クラス or 購入クラス

継続してサイトにアクセスのみしているユーザを対象に、
その後商品購入に至る顧客か、至らない顧客かを分類予測する

分析Step2：1回購入クラス or 2回以上購入クラス

1度購入経験したユーザを対象に、その後2回以上の購入に至る顧客か、1回のみにとどまる顧客かを分類予測する

集計の結果、巡回ページ数が多くても購入には至らないことが確認されたため，
実際の行動パタンとユーザ属性を説明要因としてクラスを予測方法が必要

↓

頻出パタンマイニングや系列パタンマイニングを応用し、複合形式のデータに対する新たなクラス予測モデルを提案する

↓

実際のゴルフ用品オンラインショップのデータを用いて、上記分類問題に手法を適用し、提案手法の有効性を検証するとともに、予測に寄与したパタンから、マーケティングを行う上で着目すべき点についての考察を行う

分析手法の基礎的内容と関連研究

以下では、本提案に関係する基本的な用語等の簡単な説明と、もっとも関係すると思われる関連研究を取り上げてその類似点や独創的なポイントについて説明する。

属性データ及びログデータの形式

＜属性データの例＞

顧客ID	アイテム
1	男性
1	30代
1	会社員
2	女性
2	20代
2	学生
…	…

＜閲覧履歴（ログ）データの例＞

顧客ID	順序	アイテム
1	1	ページ7
1	2	ページ3
2	1	ページ10
2	2	ページ3
2	3	ページ7
2	4	ページ10
…	…	…

属性データは顧客IDがキーであるのに対し、ログデータは顧客IDと閲覧順序という2つのキーが存在している系列データであり、データ形式が異なる

- 各形式のデータに対しては、様々な手法が提案されてきた
- また、順序キーを無視して扱うことで他の手法を適用している事例は様々存在
- これら2つの形式のデータを、そのまま利用した手法はあまり存在しない
- 2つの形式のデータを統合的に扱った研究の1つとして羽室ら[4]の研究がある。

頻出パタンと系列パタン

＜属性データの例＞

TID	アイテム
1	男性
1	30代
1	会社員
2	女性
2	20代
2	学生
…	…

TIDはすべて同一のクラスに所属しているとする。

共通するアイテムセット（パタン）とは

（例） x=（男性　会社員） この場合　（会社員　男性）
であったとしても同じ
xがどれくらい頻出しているかを表現するために、サポートを定義

$$\mathrm{support}_{D_1}(x) = \frac{\mathrm{count}_{D_1}(x)}{|D_1|}$$

・$|D_c|$:　D_cに属するTID件数

ここで$\mathrm{count}_{D_c}(x)$はデータベースD_c出現するTIDの数
サポートが規定の基準値以上であれば頻出パタンと呼ぶ

＜閲覧履歴（ログ）データの例＞

TID	順序	アイテム
1	1	ページ7
1	2	ページ3
2	1	ページ10
2	2	ページ3
2	3	ページ7
2	4	ページ10
…	…	…

共通する系列パタンとは

（例）y=＜ページ7　ページ3＞　この場合
＜ページ3　ページ7＞ は別の系列パタンである

サポートの考え方は、頻出パタンと同じである。
サポートが規定の基準値以上であれば頻出系列パタンと呼ぶ

クラスに特徴的なパタンの列挙

2つの異なるクラスが存在し、TIDはいずれか1つのクラスに所属しているとする。

各クラスに特徴的なパタンは、一方のクラスに大きく頻出し、他方のクラスにはなるべく出現しないパタンである。これを定義する考え方は大きく以下の2つ存在する

＜増加率定義＞

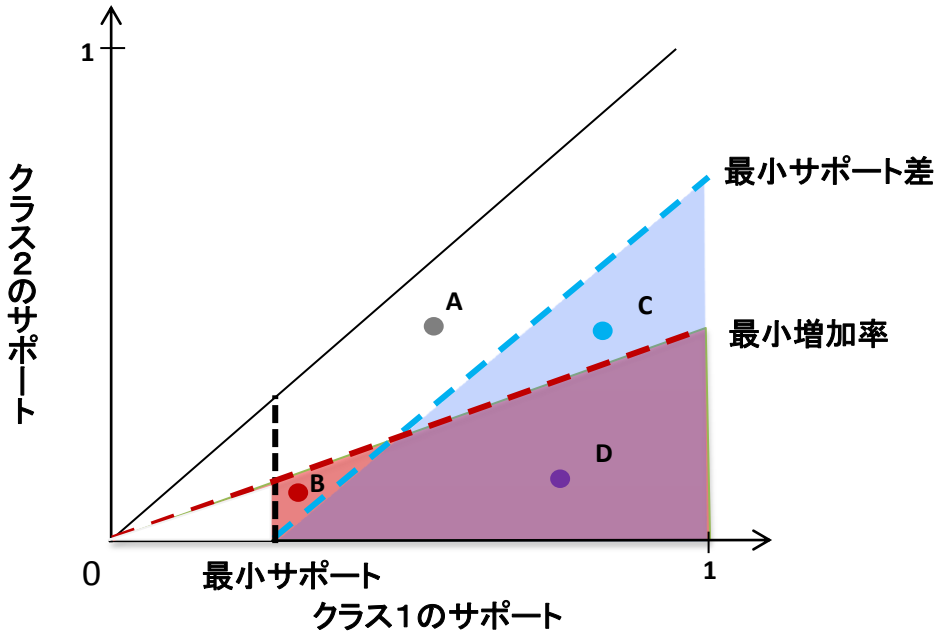
$$\mathrm{GR}_{D_1}(x) = \begin{cases} \dfrac{\mathrm{support}_{D_1}(x)}{\mathrm{support}_{D_2}(x)} & , \quad \mathrm{support}_{D_2}(x) \neq 0 \\ \infty & , \quad \mathrm{support}_{D_2}(x) = 0 \end{cases}$$

＜サポート差定義＞

$$\mathrm{Diff}_{D_1}(x) = |\mathrm{support}_{D_1}(x) - \mathrm{support}_{D_2}(x)|$$

最小増加率以上の頻出パタンを顕在パタン(**EP**)、最小サポート差以上の頻出パタンをコントラストパタン(**CP**)と呼び、それぞれの対象が系列パタンである場合、顕在系列パタン(**ESP**)、コントラスト系列パタン(**CSP**)と呼ぶ

EP（or ESP）とCP（or CSP）の関係



●：EPではないCPであるパタン

●：CPではないEPであるパタン

●：CPかつEPであるパタン

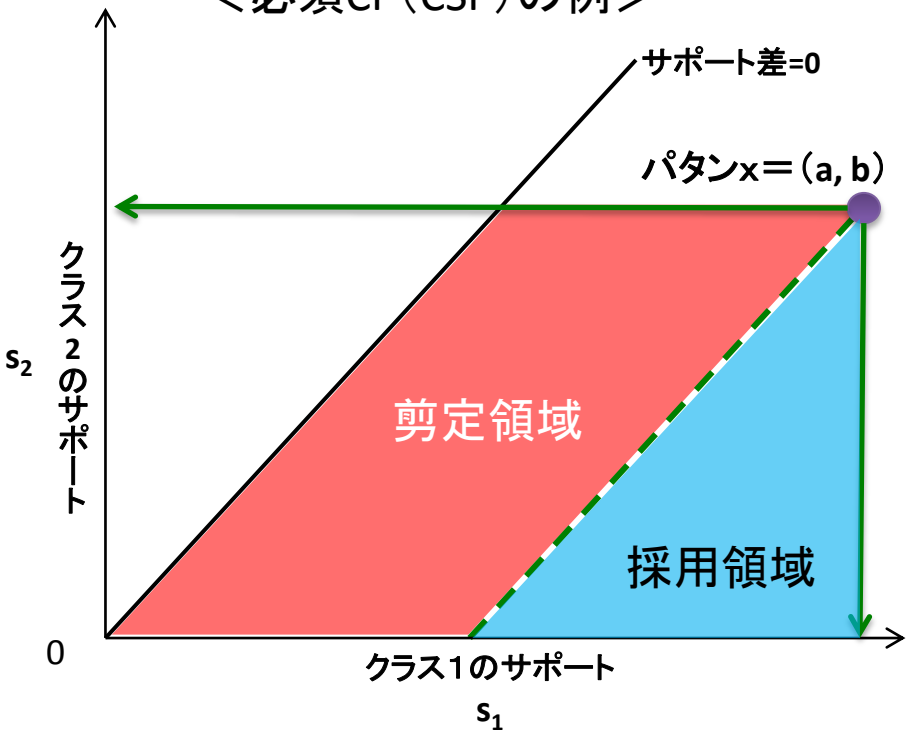
●：CPでもEPでもないパタン

定義の違いから各クラスのサポート空間にプロットしてみると、上図のような領域のずれが発生する。

	メリット	デメリット
CP	LCM[5],およびLCM_seq[5]が直接、CPを出力可能であるため、高速に処理することが可能、CPがカバーしているTIDに対して、何らかの予測を行う方法では、未分類に陥るTIDを減少させることができる	両方のクラスに一定程度存在するパタンも含まれるため、分類予測モデルを作成する場合に、予測誤差を生じる可能性がある
EP	片側のクラスにのみ顕著なパタンを最優先するため、パタン自体の説明力は強い	TIDをカバーする範囲が小さくなりやすく、未分類のTIDを発生させる危険性が存在する

必須EP（ESP）と必須CP（CSP）

＜必須CP（CSP）の例＞



パタンx=(a, b)がCPの場合、パタンy=(a, b, c)もCPのとして列挙されることがある。

このとき$support\boldsymbol{D_c}(y) \leqq support\boldsymbol{D_c}(x)$となるが、$Diff\boldsymbol{D_c}(y) < Diff\boldsymbol{D_c}(x)$であるならば、xですでにカバーされていると考えられるので、あえて残す必要はあまりない。

そこでこのように、yがxに含まれ、かつ上記のサポートの条件をみたす場合、yを取り除くことを剪定と呼ぶ。

図中ではyが青の領域のいずれかに存在すれば、そのままCPとして採用され、赤の領域に存在すれば剪定される。CSPも同様である。

EP（ESP）の場合は$GR\boldsymbol{D_c}(y) < GR\boldsymbol{D_c}(x)$**であれば**同様に、選定される。
これら剪定後に採用されたパタンを必須EP（ESP）、必須CP（CSP）と呼ぶことにする

EP（ESP）を用いたクラス分類手法

EPを用いた分類手法として、Classification by Aggregating Emerging Pattern[6]（以下、CAEP）が挙げられる

$$\mathrm{score}(d,c) = \sum_{e \subseteq d, e \in E(c)} \frac{\mathrm{GR}_{D_c}(e)}{\mathrm{GR}_{D_c}(e)+1} \cdot \mathrm{support}_{D_c}(e)$$

事後確率

$$\mathrm{norm\_score}(d,C) = \frac{\mathrm{score}(d,C)}{\mathrm{base\_score}(C)}$$

c : クラス
d : レコードデータ
e : 顕在パタン
E(c): あるクラスcの顕在パタン集合
score(d,c): cにおけるdの集約スコア
base_score(d,c): cの集約スコアの中央値
norm_score(c): cにおけるdの
基準化スコア

パタン集合である予測レコードdに対して、顕在パタンのスコア値を用いたクラス投票を実施

各クラスのnorm_score（d,c）が最大値をとるクラスを予測クラスとする

EPではなくESPを用いて、同様の手順で分類モデルを構築する手法が、Classification by Aggregating Emerging Sequential Patterns[1]（以下、CAESP）である

羽室らの提案手法[4]

＜羽室らの提案手法の特徴＞

取り扱うデータ	あるWEBサイトのユーザアクセスログデータ
モデル構築手法	アクセス日の区分やアクセス時刻区分などをアイテムとしてEPを、アクセスログのページ遷移をアイテムとしてESPを列挙した後、これらEPとESPを組み合わせた、統合化EPを使用し、CAEPにより予測モデルを構築している。

＜パタン列挙時の工夫点＞

時間幅制約	異なる時刻のトランザクションであっても、ユーザが指定した時間内に発生したトランザクションを同一とみなす
タクソノミの導入	サイトのディレクトリ構造などを利用。比較実験の結果から、タクソノミの導入により予測精度の向上が確認されている
必須EPによる剪定	必須EPの定義によりパタンの剪定を行うことで、多くの冗長なパタンを剪定している

羽室らの研究では、タクソノミや必須EPなどの工夫を凝らし、EPとESPを用いて、実用的な結果を導くことに成功している

既存研究との違い

羽室らの研究の問題点と、本研究との相違点は以下の通りである

＜問題点＞

- 論文では詳細なアルゴリズムが不明で、特にパタン列挙についての方法が明記されていない
- 計算時間についての記述がなく、パタン計算にどのくらいのコストが必要か不明 （パタン列挙には通常大きな計算コストが必要のため）
- 扱う予測問題が難しいこともあるが、予測の精度自体は良いとは言えない

＜相違点および改善点＞

- 時間幅制約は、我々は今回は分析期間中のデータを1つのトランザクションとしてとらえるため、採用していない
- タクソノミに関しては、いくつかの定義で導入を試みたが、計算時間上の困難が生じたことや、予測精度の向上につながらなかったため、今回は採用していない
- Visual Mining Studioのルールベース予測を利用することで、モデル構築に使用したパタンの視覚化などにも対応し、より実用的である

アソシエーションルール[7]

＜信頼度定義＞

$$\mathrm{Confidence}(\mathrm{A} \rightarrow B) = \mathrm{Support}(A \rightarrow B)/\mathrm{Support}(A)$$

＜リフト値定義＞

$$\mathrm{Lift}(\mathrm{A} \rightarrow B) = \mathrm{Confidence}(A \rightarrow B)/\mathrm{Support}(B)$$

＜Conviction定義＞

$$\mathrm{Conviction} = \mathrm{Support}(\mathrm{A}) \times \mathrm{Support}(B')/\mathrm{Support}(A \rightarrow B')$$

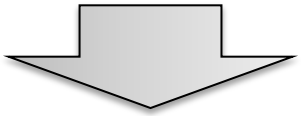
任意の最小サポート、最小信頼度、
最小リフト値、最小Convictionを満たす
ルール*A*→*B*をアソシエーションルールと呼び、
結論部*B*がクラスラベルであるルールを
クラスアソシエーションルールと定義する

- *A*:ルールの前提部
- *B*:ルールの結論部
- *B'*:B以外の事象
- *A*→*B* : ルールAならばB
- Support(*A*→*B*) : *A*→*B*のサポート
- Confidence(A→*B*): A→Bの信頼度
- Lift(A→*B*): A→Bのリフト
- Confidence(A→*B*): A→Bのconviction

Visual Mining Studioでは、大規模なデータに対してアソシエーションルールやクラスアソシエーションルールを効率的に列挙可能

クラスアソシエーションルールを ベースとしたクラス予測モデル

パタン集合である予測レコードdに対して、
クラスアソシエーションルールを用いたクラス投票を実施



各クラスの予測値※を計算し、
最も大きなスコアをとるクラスを予測クラスと分類する

※本研究では、モデル構築にはVisual Mining Studioを使用しており、予測値を判別する判別関数は、信頼度が最も高いルール、信頼度が同じ場合にはサポートの大きいルール、信頼度・サポートともに同じ場合にはルールの長さの短いルールを採用し、その信頼度とした最大値関数か、最大値関数にサポートの重みを考慮した平均値関数のうち、予測精度の高かった判別関数を採用することとする

このルールベース予測は、アソシエーションルールをそのまま
予測に用いるため、モデルの解釈が容易であるというメリットがある

提案手法

以下では、具体的な提案手法の構築方法を述べる。まず、全体の流れを説明した後、計算時間を短縮するためのパタンの列挙方法や、アソシエーションルールを列挙する際の入力データの作成方法を説明する。

非系列データと系列データを統合的に利用したクラス予測モデルの概要

属性データから入力データを作成

⇩

CP（またはEP）の列挙

⇩

ログデータから入力データを作成

⇩

CSP（またはESP）の列挙

⇩

CP（またはEP）とCSP（またはESP）のパタンを1つのアイテムに置き換え、元データをその変換アイテムに置換したデータに変換

⇩

変換したデータを入力として、長さ1～2のクラスアソシエーションルール※を列挙

⇩

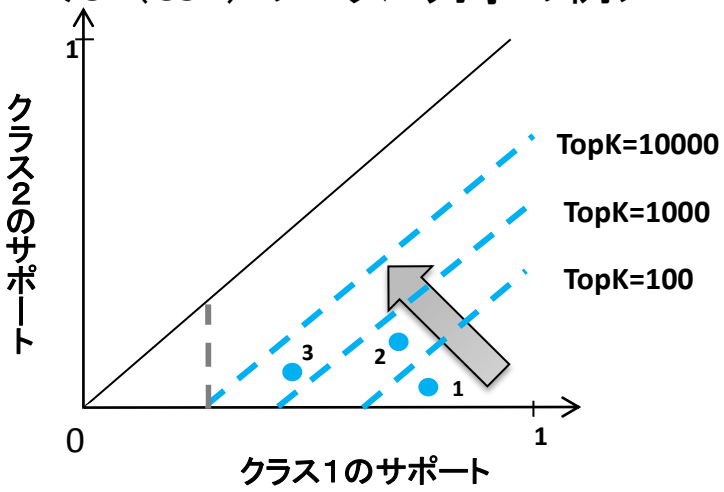
ルールベース予測により、予測モデルを構築※

※クラスアソシエーションルールの列挙および予測モデルの構築には、Visual Mining Studioを使用

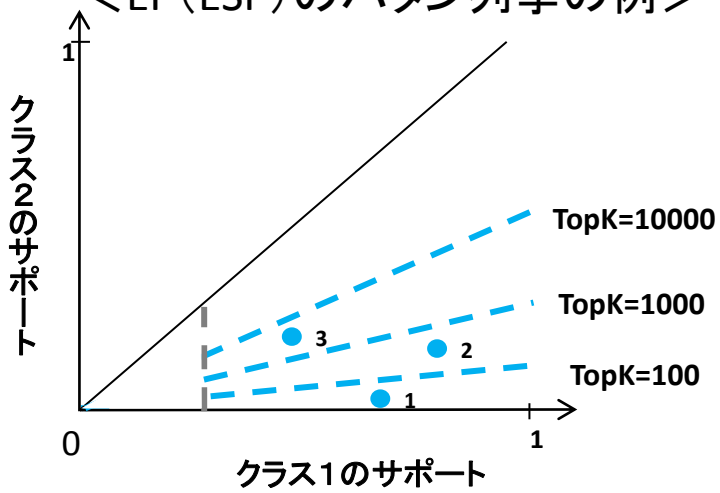
パタンの列挙方法

- 任意の最小サポートを満たすCP（またはCSP）のうち、サポート差の大きなパタンから採用していき、任意の強力な列挙パタン数の上限（topK）を満たした時点でパタンの探索を停止する（同じサポート差のパタンが複数個存在する場合は、topK個以上のパタンが列挙される）。この方法は、LCM[5]（またはLCM_Seq[5]）との相性がよく、高速なCP（またはCSP）の列挙が可能となる。
- EP（またはESP）の場合も同様の考え方で、増加率の高いパタンからtopKの上限まで採用する。

＜CP（CSP）のパタン列挙の例＞



＜EP（ESP）のパタン列挙の例＞



また、アソシエーションルール列挙の際、CP（またはEP）のみ、
もしくはCSP（またはESP）のみで構成されたルールは、冗長なパタンのため、除外する

ルールベース予測モデルの構築方法

＜CP（またはEP）入力データ＞

顧客ID	アイテム
1	男性
1	30代
1	会社員
…	…

＜CSP（またはESP）入力データ＞

顧客ID	順序	アイテム
1	1	ページ1
1	2	ページ7
1	3	ページ3
…	…	…

CP_2={男性, 会社員}

CSP_7={ページ1, ページ3}

顧客ID	アイテム
1	CP_2
1	CSP_7
…	…

CP（EP）とCSP（ESP）をアイテムとして、ルール列挙用の入力データを作成

Visual Mining Studioの
クラスアソシエーション分析により
各クラスのルールを列挙

不要な（冗長な）パタンの除外

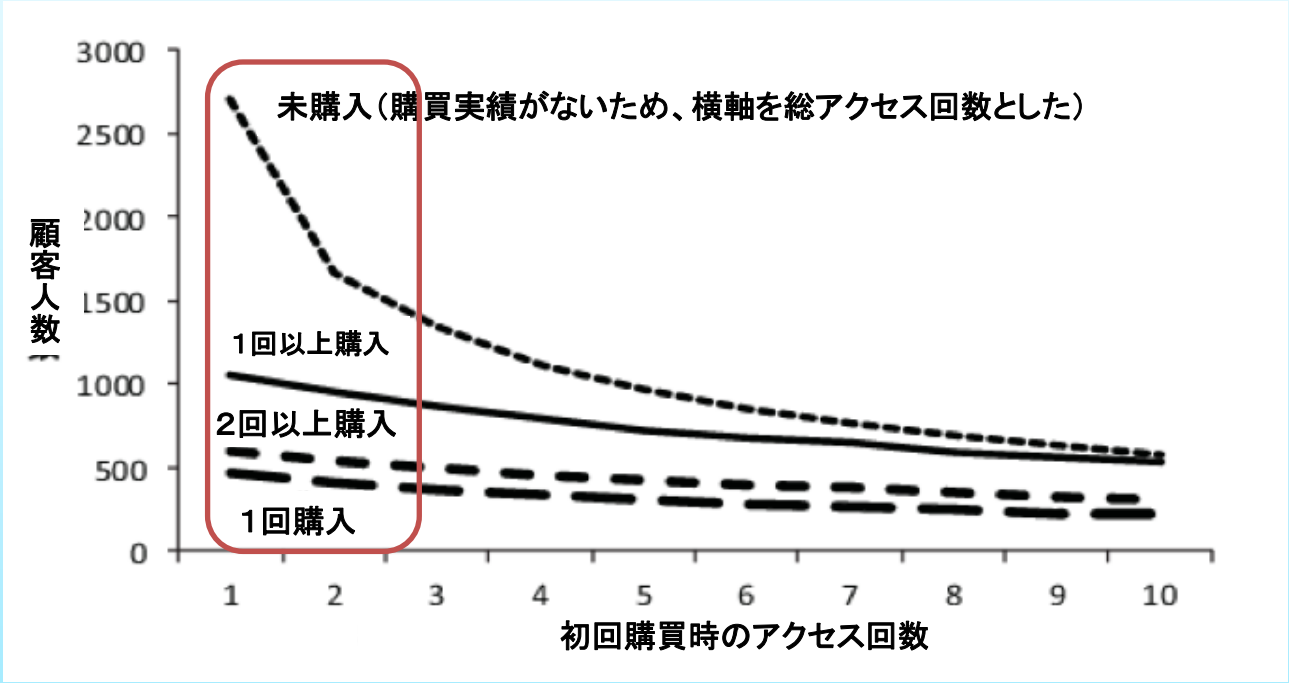
Visual Mining Studioのルールベース予測により、予測モデルを構築

ルール列挙用入力データ作成以降のVisual Mining Studioにおける処理の例

計算機実験と結果

実際のゴルフ用品オンラインショップのデータを用いて、提案手法の有効性を確認する。まず、モデル構築の際の分析期間の設定やクラス定義、使用する説明変数および計算時のパラメータの設定を述べ、結果の評価値を示す。得られたルールをネットワークグラフにより視覚化し、ルールの考察を行う。

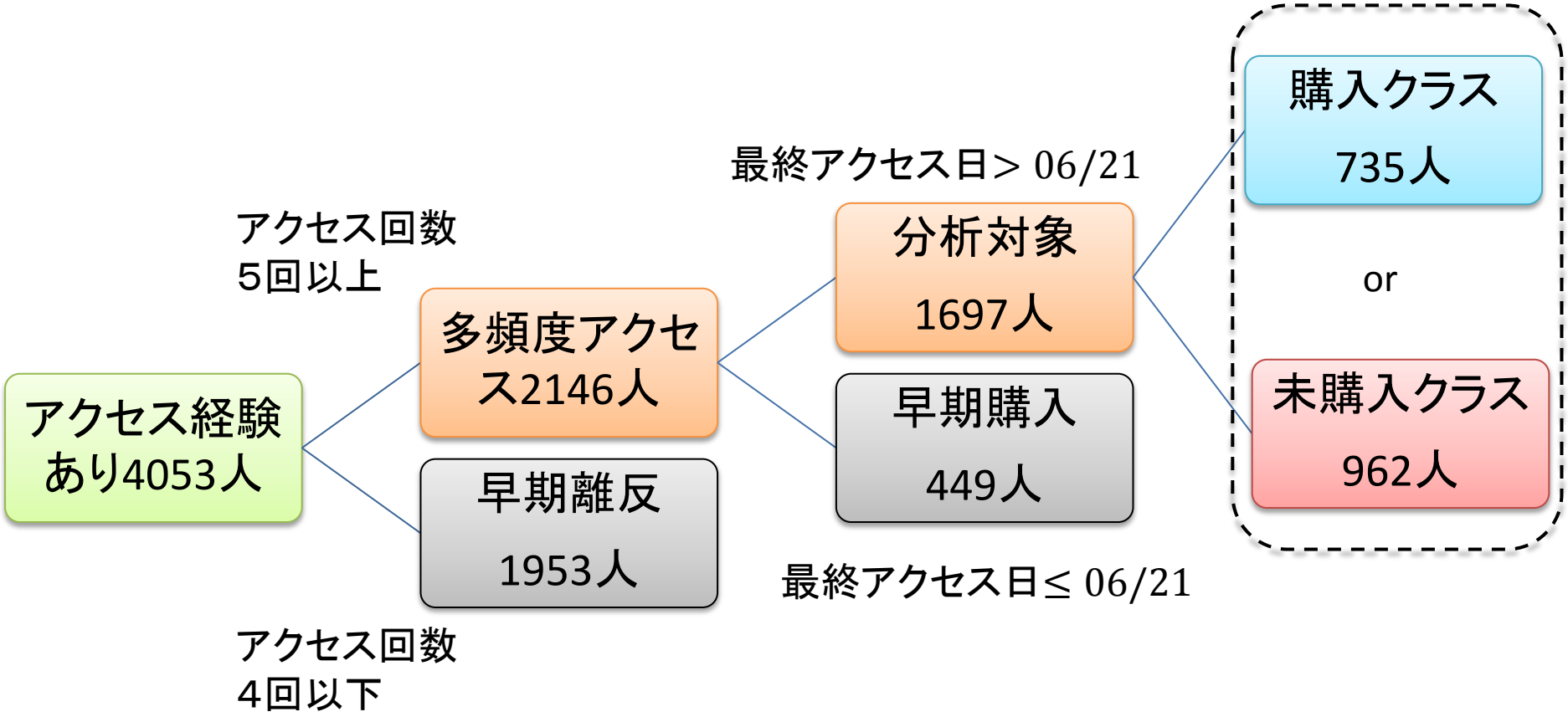
分析期間の決定



初回購入時のアクセス数と未購入者の総アクセス数

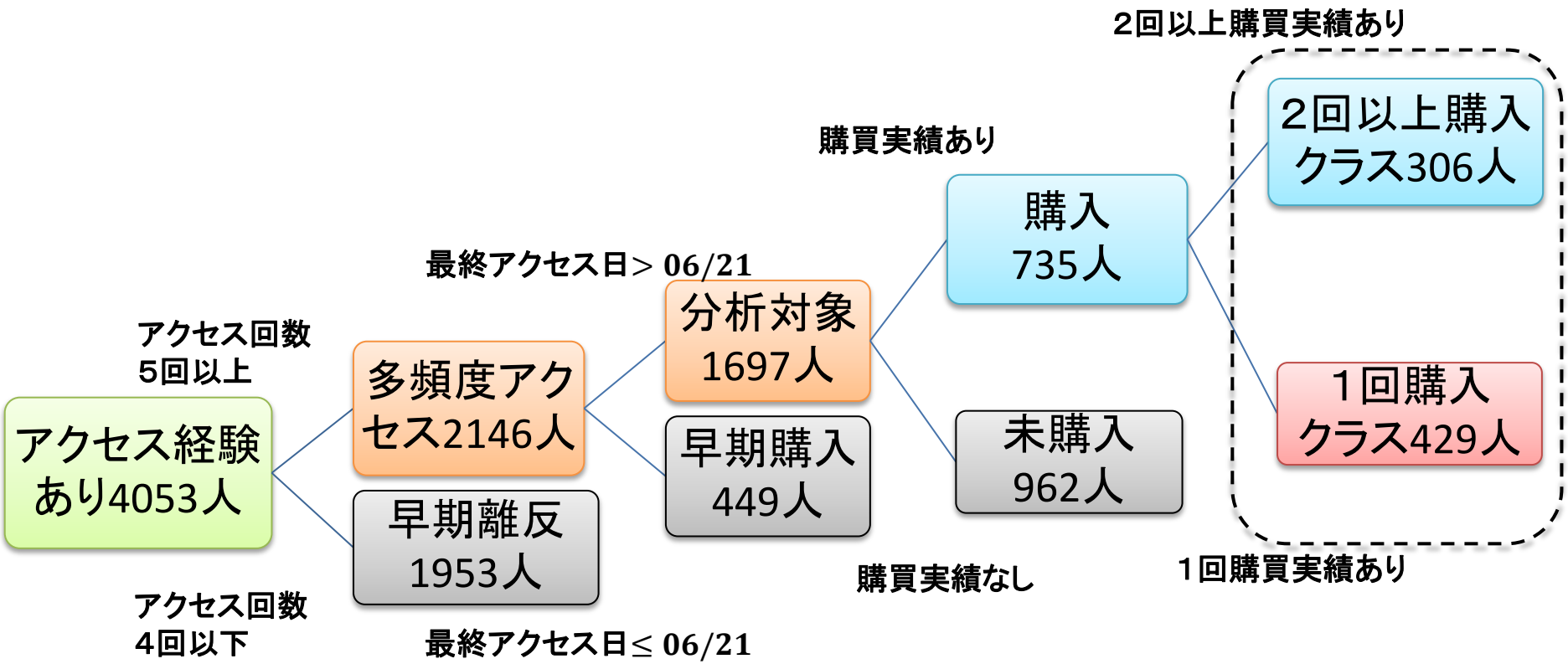
お試しアクセスで休止してしまう顧客が確認できるが、今回はページへの一定アクセスが存在する顧客を想定している。初回から5回のセッションを、モデルを構築するトレーニング期間とする。

分類問題1のクラス定義



アクセス回数が分析期間以上存在し、かつ最終アクセス日が提供データ期間終了日の1週間前以内（離脱していない）である顧客が分析対象

分類問題2のクラス定義



購入顧客の中でも、1回しか購入していないか、2回以上購入しているかによってクラスを決定する

説明変数およびパラメータ設定

＜属性データの説明変数＞

変数	カテゴリ
会員登録年	2010年以前登録 or 2010年以降登録
性別	男性 or 女性
年齢	5歳刻み
ハンディキャップ	スコア登録の有無
メルマガ購読	メルマガ購読の有無
トップページ閲覧	トップページを閲覧回数5回以下 or 6回以上

＜ログデータの説明変数＞　　アイテム単位=1ページ，レコード単位：ユーザID

変数	カテゴリ
ページ	• ページのURL情報をそのまま説明変数として使用 • トップページは、多くの顧客が訪問していたため、適当にカテゴリ化を行ったうえで属性データとして使用 • トップページのポップURLは、比較的サポートが高くかつ解釈困難なため、除外

＜パタン列挙時のパラメータ設定＞

パタン種類	パタン長
CSP（またはESP）	2～4
CP（またはEP）	1～7

＜アソシエーションルール列挙時のパラメータ設定＞

分類問題	最小サポート件数	最小信頼度	最小lift値	最小conviction	パタン長
1	3	50%	1.2	0.6	1～2
2	3	50%	1.2	0.9	1～2

検証内容と結果の評価値

- ３種類のシードによるテストサンプル法※により、モデルの検証を行う
- 同じ入力データ・同じパラメータ設定・パタン数により、CP&CSP、EP&ESPのそれぞれのパタンを用いたルールベース予測によりモデルをそれぞれ構築し、各パタンの効果を比較検証する
- topK=100とtopK=300の２つの設定値によりモデルを構築し、topKと各パタンの関係を検証する
- 予測精度は、accuracyの他に、precisionとrecallの加重調和平均であるf1値により評価し、３つのデータセットにおける各評価値の平均値、最大値、最小値を掲載する。さらに未分類数（予測不能ID）やモデルに使用したパタン数の観点からも考察を行う

$$accuracy = \frac{\mathrm{a} + \mathrm{d}}{\mathrm{a} + b + \mathrm{c} + d + e + f}$$

$$precision = \frac{a}{a + c}$$

$$recall = \frac{a}{a + b + e}$$

$$f1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

	予測されたクラス			
実際のクラス		購入クラス	未購入クラス	未分類数
	購入クラス	a	b	e
	未購入クラス	c	d	f

※テストサンプル法は、訓練データ6割、検証データ4割の割合でランダムに選択

計算結果 分類問題1

topK	100		300	
パタン種類	CP&CSP	EP&ESP	CP&CSP	EP&ESP
Accuracy_平均	0.569	0.463	0.561	0.579
購入クラスf1_平均	0.565	0.543	0.494	0.605
未購入クラスf1_平均	0.564	0.521	0.585	0.551
Accuracy_最大	0.591	0.489	0.578	0.599
Accuracy_最小	0.548	0.448	0.531	0.545
購入クラスf1_最大	0.635	0.543	0.636	0.615
購入クラスf1_最小	0.548	0.490	0.361	0.595
未購入クラスf1_最大	0.598	0.575	0.643	0.598
未購入クラスf1_最小	0.406	0.503	0.482	0.470
未分類数	1	176	0	5
計算時間(sec.)	3	430	6	611
使用パタン数	132	362	286	535

- 全体の考察としては、極端に良い結果とは言えないものの、topK=100においてはCP&CSPの、topK=300においてはEP&ESPの結果で各クラスのf1値が0.5を超えており、かつどのデータセットに対しても安定したパフォーマンスを示していることがわかる。
- topK=100に着目すると、3つの評価平均値全てにおいてCP&CSPの方がEP&ESPよりも上回っている。この原因としては、CP&CSPの未分類数は1件であるのに対して、EP&ESPの未分類数は176件もあることが影響しているだろう。つまり、増加率の高い100個（同点含め362個）のパタンでは少なすぎることが確認できた。
- topK=300の場合では、EP&ESPの未分類数は大幅に減少し、評価値も上昇している。逆にCP&CSPでは評価値が下がっており、パタン数が多すぎて過剰適合を起こしてしまっている可能性がある。

計算結果 分類問題2

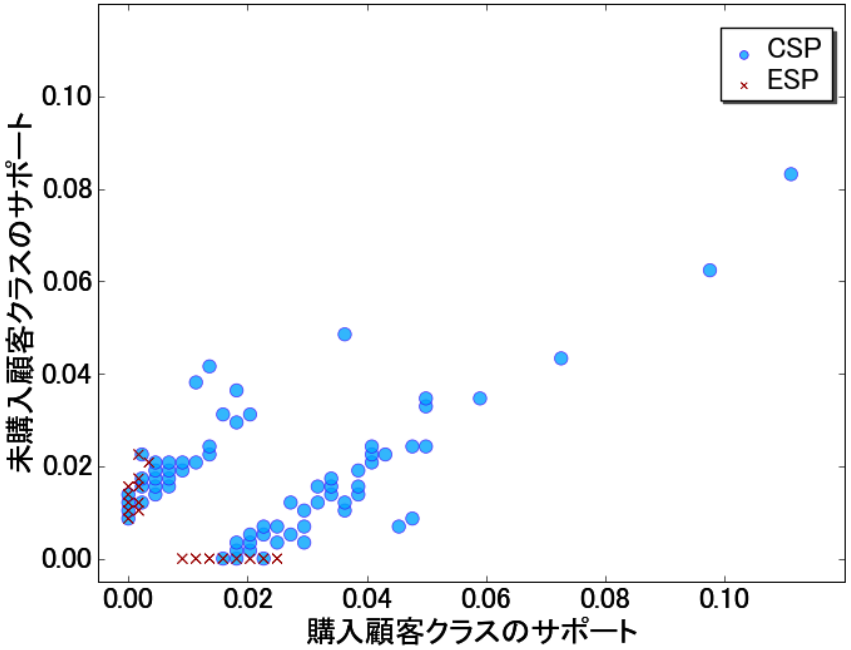
topK	100		300	
パタン種類	CP&CSP	EP&ESP	CP&CSP	EP&ESP
Accuracy_平均	0.578	0.473	0.573	0.601
2回以上購入クラスf1_平均	0.667	0.603	0.665	0.691
1回購入クラスf1_平均	0.416	0.408	0.380	0.440
Accuracy_最大	0.602	0.515	0.589	0.610
Accuracy_最小	0.544	0.449	0.545	0.590
2回以上購入クラスf1_最大	0.702	0.653	0.705	0.696
2回以上購入クラスf1_最小	0.368	0.541	0.591	0.682
1回購入クラスf1_最大	0.464	0.462	0.492	0.465
1回購入クラスf1_最小	0.604	0.331	0.289	0.421
未分類数	1	64	1	1
計算時間(sec.)	2	105	2	124
使用パタン数	114	274	171	303

- 分類問題2に関しても、分類問題1の時とほぼ同様の結果であるが、全体的にパフォーマンスは上がっている。
- 1回購入クラスのf1値が低いことが気になるが、これは、2回以上購入クラスの方のConfidenceが全体的に大きいため、2回以上購入クラスの予測の圧力が大きくなっている結果であると考えられる。しかし、この分類問題において予測したい方のクラスは2回以上購入クラス側であり、2回以上購入クラスのf1値は高いため、予測モデルとしては十分であるといえる。

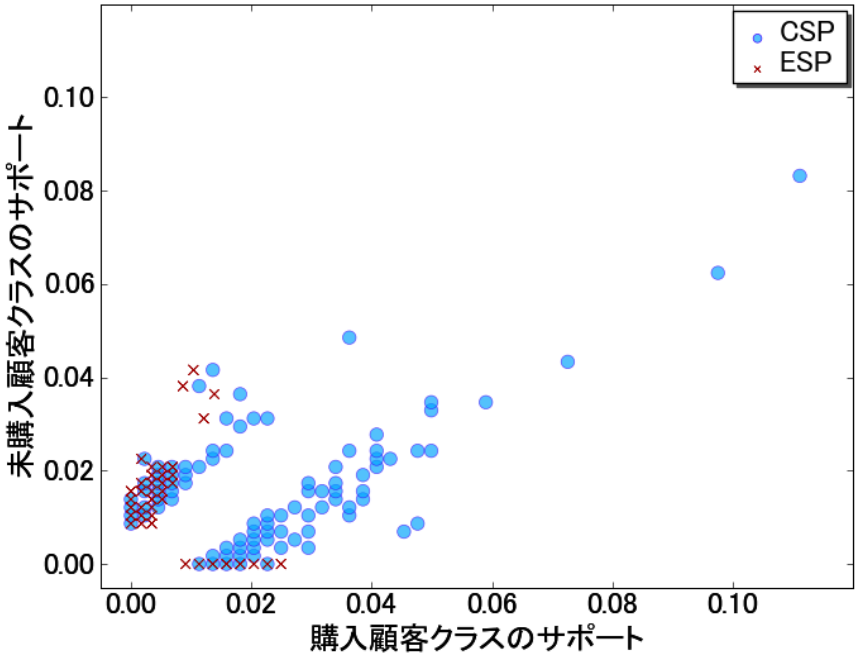
計算時間の問題などで、多くのパタンを列挙できず、ある程度の予測精度を得たい場合は、CP&CSPを用いた方が良く、逆に十分に時間をかけて予測モデルを構築できるのであれば、EP&ESPを用いた方が良い精度を得ることが出来るだろう

各パタンのサポート分布

先ほどの予測結果を踏まえ、各パタンとtopKの関係を確認する。
以下の2つの図は、分類問題1のデータセット1において出現したESPとCSPを各クラスのサポートを軸とした空間に布置したもので、左図がtopK=100の時、右図がtopK=300の時の出現パタン分布である。



topK=100の場合のESP, CSPのサポート分布



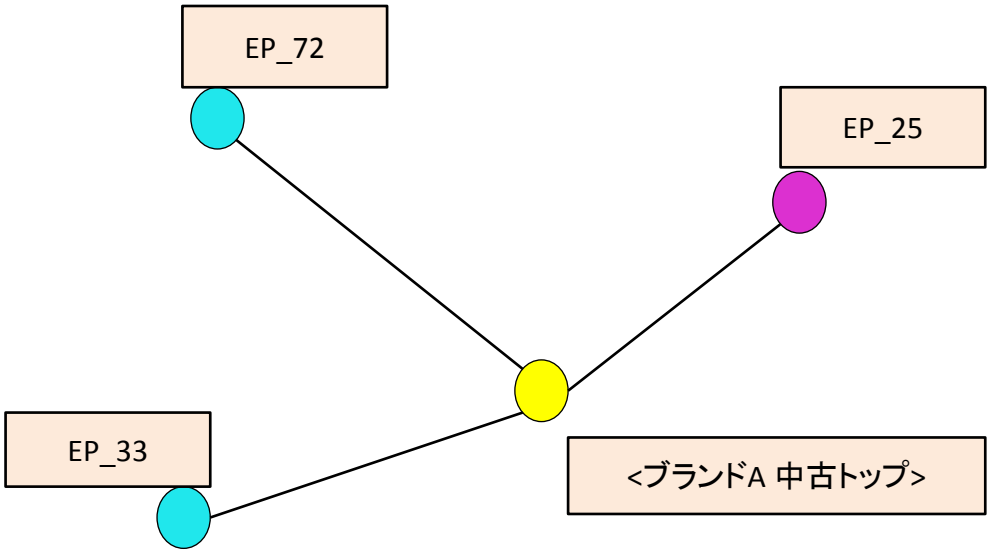
topK=300の場合のESP, CSPのサポート分布

- topK=100の時、ESPは、一方のクラスにしか出現していない強力なパタンばかりが出現しているが、サポートが比較的小さいため、予測不能ID数が多く出現してしまうことがわかる。CSPはサポートが比較的高く、かつ十分に一方のクラスに顕著なパタンが出現している。
- topK=300の時、CSPでは、サポート差が比較的小さいパタンも出現しており、このパタンが過剰適合を引き起こし、誤分類の原因となってしまったと思われる。

代表的なルールの視覚化

Visual Mining Studioのネットワークグラフ機能を利用し、得られたルールの視覚化を行う

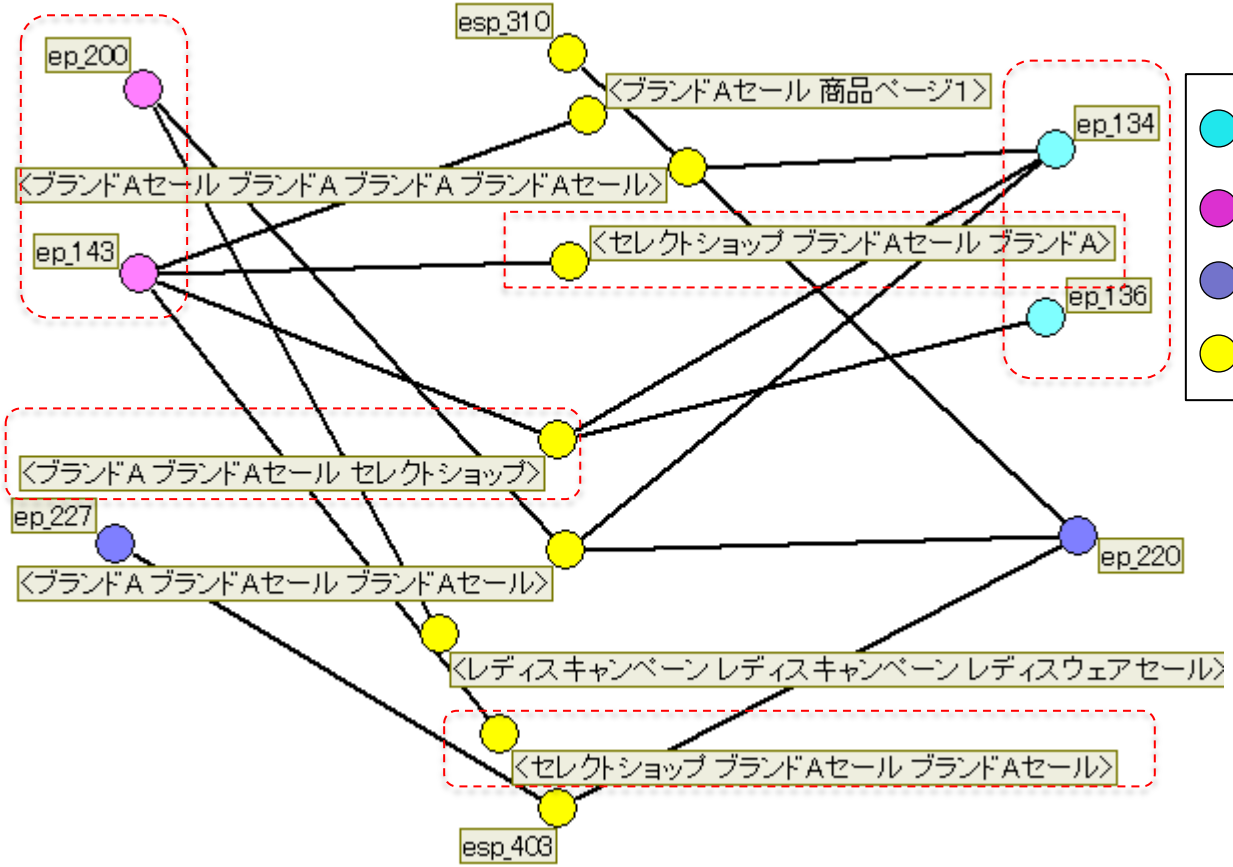
- 各クラスで列挙されたルールの中から、長さ2のルールに限定し、Confidenceが高い上位30ルールを抽出し、それぞれのネットワークグラフを作成する。
- 分類問題1のグラフにはEP（ESP）を、分類問題2のグラフにはCP（CSP）を用いたルールを使用している

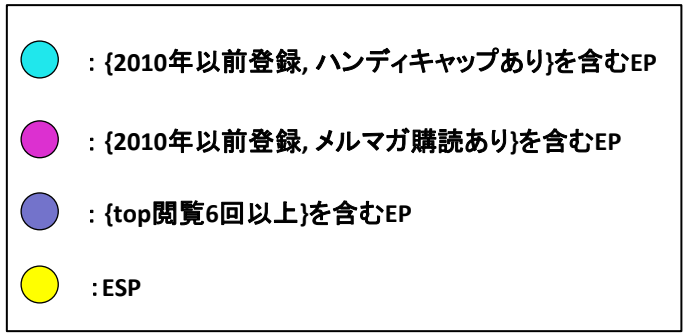


ネットワークグラフの例

1ノード	1パタン
ノードのラベル	パタンもしくはパタンID。例えば、EP_72はEPのID72のパタンを表し、<ブランドA 中古トップ>は、ブランドAのページを閲覧し、中古トップのページを閲覧したという意味を表す
ノードの色	EP（またはCP）では、ある興味深い共通の部分パタンを持つパタンを同じ色に分類し、ESP（またはCSP）は同じ色として統一
エッジ	1つのエッジで結ばれたパタンが、1つのルールを表す
エッジの太さ	Confidenceの大きさに比例

購入顧客クラスの特徴的ルール





＜EPの特徴＞

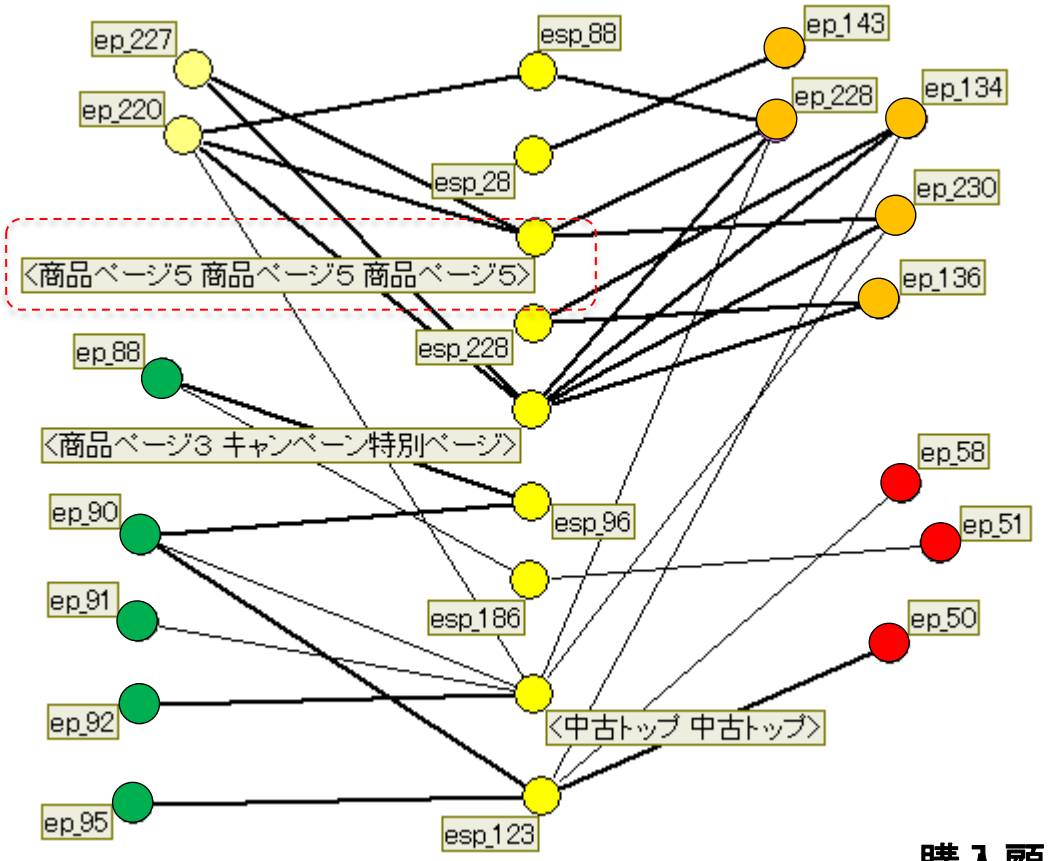
1年以上前に登録をしていて、ハンディキャップを持っていたり、メルマガを購読している顧客セグメントが出現している。これら顧客は、熟練者、またはレベルに関係なく情報収集を積極的に行っている関心度の高い顧客である。

＜ESPの特徴＞

セレクトショップや、その中の特定のブランドのページを複数回閲覧していることが分かる。

このような関心をもった顧客が関与に踏み出そうとするタイミングが、1つの重要なプロモーション機会であることが分かる。また、レディス関連のページを複数回閲覧しているパタンも出現していることから、レディス用品を目当てにサイトを訪問した顧客は、比較的購買アクションを起こしやすいのではないかと考える。

未購入顧客クラスの特徴的ルール



＜EPの特徴＞

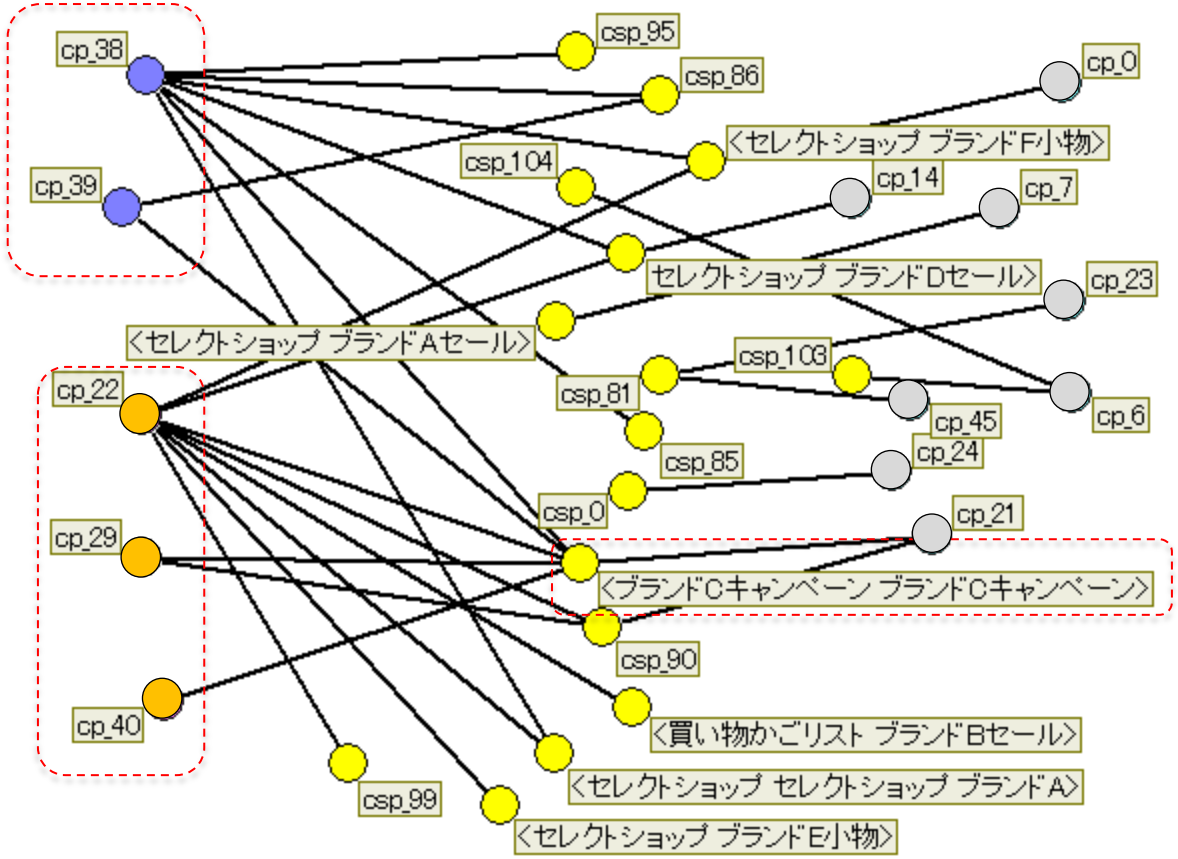
1年以内に登録した新規顧客や、メルマガ購読を行っていない関心度が比較的弱いセグメントが出現している。

＜ESPの特徴＞

商品ページ3や5、中古トップを複数回閲覧している閲覧行動が確認できる。商品ページの内容は不明であるが、このようなページへのアクセスは、何か特殊な購買行動である可能性がある。

購入顧客の特徴と比較すると、顧客の関心度、および収集しようとする情報が異なっていることがわかる。最初の一歩は、関心度の高い顧客が、関与に踏み込もうとするタイミングでの取りこぼしがないような施策が、またより深い掘り起こしとしては、関心度の弱い顧客の関心度を高め、その上で関与につなげる施策が重要となる。

2回以上購入顧客クラスのルール



＜EPの特徴＞

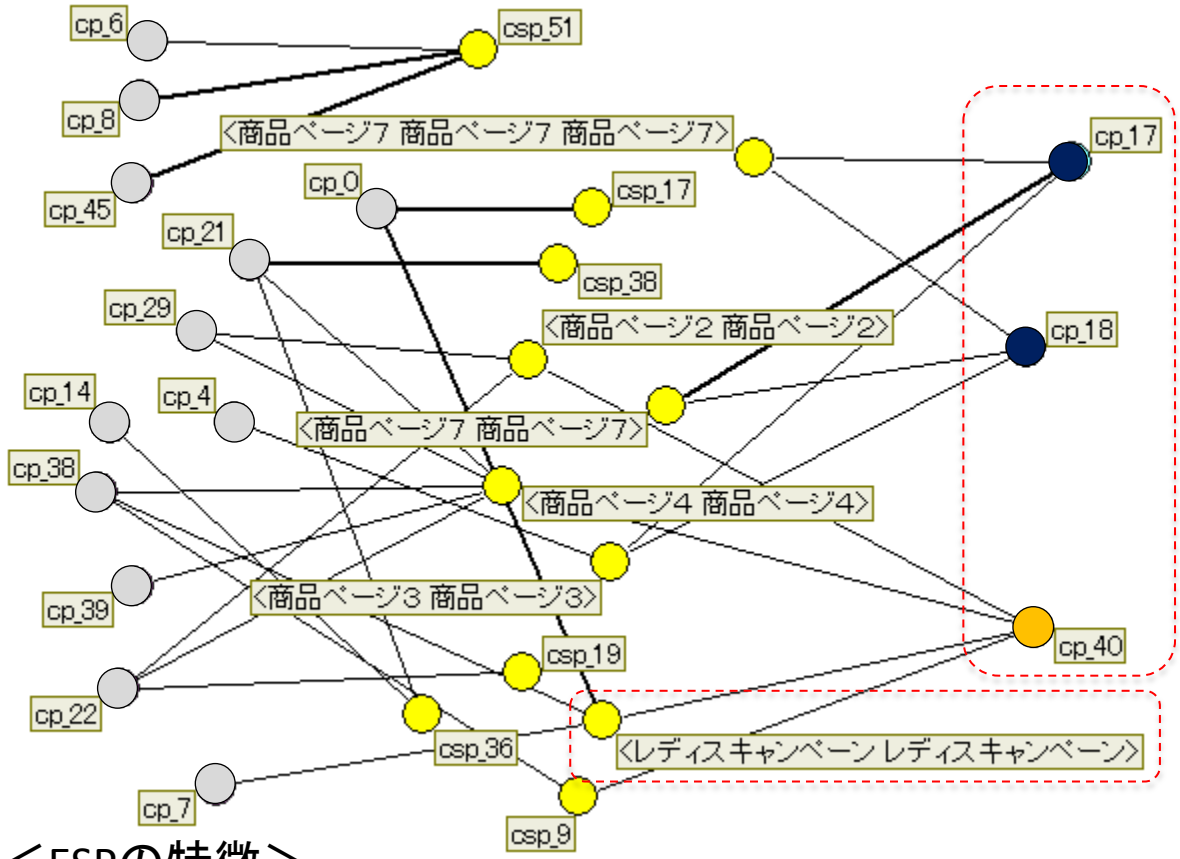
購入顧客クラスと同様、１年以上前に登録している顧客や、メルマガ購読をしている関心度の強い顧客が確認できる。

＜ESPの特徴＞

興味深い新たな行動パタンとしては、ブランドCのキャンペーンページの複数回閲覧が出現している。

ゴルフの熟練者、または情報収集も行っているセグメントでは、ゴルフグッズへのこだわりを示しており、キャンペーンに反応する行動は、関与度を更に強めようとしているサインの１つであり、複数購買が期待できるポイントである。

１回購入顧客クラスのルール



＜ESPの特徴＞

レディスキャンペーンページへの複数回アクセスは、単なるバーゲンハンターなのかもしれない。もちろん今回のデータ期間外での長期のタイミングでは複数購買がありえるかもしれないし、関与度自体は高まるのかもしれない。

：{ハンディキャップなし, メルマガ購読あり}を含むCP
：{2010年以前登録}を含むCP
：その他のCP
：CSP

＜EPの特徴＞

２回以上購入顧客クラスと同様、関心度の強い顧客が確認できるが、一方でそれ以外のノードが多数出現していることがポイント。

キャンペーンは重要な施策要素であるが、短期的な効率を期待するなら、関心度の高い顧客セグメントに限定して、実施することが望ましい

まとめ及び今後の課題

- **学術的意義**

本研究では、実際のオンラインショップのデータを利用して、属性データとログデータが混在するデータに対する予測モデルを構築し、2つの分類問題に対する計算機実験の結果から、良好な予測性能を確認した。またCPやEPを利用する場合、および幾つかのパラメータの感度分析も実施した。

- **応用可能性の確認**

クラス予測に寄与したパタンをもとに、各クラスのプロファイリングを行い、それぞれのクラスに対するマーケティングアプローチを行う際の着目すべき点について議論した。

- **分析結果から予想されるアプリケーション**

発見された示唆として、今回のデータからはゴルフ用品をオンライン購買しようとするには、ゴルフに対する関心度の高さ、および一定期間のゴルフ経験が重要であると思われる。またそれらの関心度の高い顧客が、ゴルフ用品の中でもウェアに対するアクセスをしている際に、プロモーション施策のチャンスが有ることがわかった。単純なキャンペーンでは、意図しないバーゲンハンターを誘引する可能性もあるので、限定した関心の高い顧客セグメントに対する、比較的高級ブランドに的を絞ったキャンペーンが、顧客関与度を高める観点からも有効であるように予想される。

- **今後の課題**

今回の予測モデルでは、パタン間でのルールベース予測の基準であるConfidenceの分布の偏りを十分考慮できていない。例えば、全体的にCP（またはEP）のConfidenceが、CSP（またはESP）のConfidenceよりもはるかに大きければ、CSPが予測モデルの結果に及ぼす影響は弱くなり、パタンを統合的に使用する意味がないという状況が発生する。これはパタン出現の状況に応じて、対処しなければならない課題であると予想される。またCPとEPの利用方法として、排他的に利用するか、または補完的な利用方法を考えるか今後検討を行い、より汎用性の高い予測モデルとしての改善を行いたい。

主要参考文献一覧

1. R, Agrawal., R, Srikant., "Mining Sequential Patterns." *Proceeding of the 11th International Conference on Data engineering*, Taipei, Taiwan, 1995
2. 矢野経済研究所, “http://www.toyokeizai.net/business/industorial/detail/AC/eb43fcef65a9ae6a0717107407e0e2b1/page/1/”
3. 経済産業省, “電子商取引実態調査”, 平成22年度, “http://www.meti.go.jp/policy/it_policy/statistics/outlook/ie_outlook.htm”
4. 羽室行信, 中西正雄, 山本昭二, “統合化顕在パターン判別モデルによるWebアクセスログデータの分析”, *オペレーションズ・リサーチ：経営の科学53(2)*, pp75-84, 2008
5. 国立情報学研究所, 宇野毅明先生のホームページ “http://research.nii.ac.jp/~uno/index-j.html”
6. G, Dong., X, Zhang., L, Wong. and J, Li.: "CAEP: Classification by Aggregating Emerging Patterns", *Proceeding of the second International Conference on Discovery Science* , 1999
7. R. Agrawal and R. srikant: “Fast algorithms for mining association rules”, *Proceeding of the 20th International Conference on VLDB*, pp.487-499, 1994