論文・事例研究

ターゲット顧客を識別するためのクレジット購買履歴データを用いたパターン分析

中原　孝信，森田　裕之

1. はじめに

好調な業績をあげていたクレジット業界も近年，収益性の悪化や，デフォルト[1]の増加に伴い厳しい経営環境にさらされつつある．クレジット業界の収益源となる3つの柱は，年会費収入，店舗手数料，そして会員手数料である[1]．このうち年会費収入と店舗手数料は，業界内の競争の激化から減少傾向にある．会員手数料は，大きく分けてキャッシング利用に伴う手数料と，ショッピングのリボルビング払い（分割払いを含む）による手数料の2つに分かれる．キャッシング利用からの手数料収入は増加しているが，デフォルト率も近年急増しており，貸倒れの危険率も増大している．また，ショッピングのリボルビング払いによる手数料は魅力的な収入源であるが，リボルビング払いは日本で解禁されて間もないこともあり，いまだ主要な収入源となりえない現状にある．一方，米国のクレジット業界では，日本で10%程度であるリボルビング払いの取扱高は70%を占め，安定的な収入源となっている[2]．

以上のような現状に鑑みて，ショッピングのリボルビング払いへのプロモーションと，キャッシングにおけるデフォルトの削減はクレジット業界において必要不可欠な経営課題といえる．

本稿では，ショッピングのリボルビング払いを効率良くプロモートするために，ターゲット顧客となりうる顧客を特定するための判別モデルを作成する．判別モデルを作成するための説明変数は，顧客属性と，顧客の購買額パターンを用いる．購買額パターンを抽出するために，各顧客の時間を考慮した購買額を，グラフ構造の1つである木構造を用いて表現し，遺伝的アルゴリズムを応用した方法で部分パターンを抽出する．最終的に，抽出した部分パターンを説明変数として用いることで，顧客属性だけを説明変数として利用する場合に比べて，モデルの判別精度が向上することを示す．

2. 関連する既存研究

データマイニング研究の中でも，グラフ構造で表現されたデータから特徴的な部分グラフを抽出する方法は，グラフマイニングと呼ばれ，1990年前後を端緒とし，高速なアルゴリズムの開発に主眼が置かれ発展してきている．

代表的なアルゴリズムには，AGM[3]やFSG[4]，gSpan[5]などが挙げられる．AGMやFSGは，隣接行列でグラフを表現し，最小サポートを制約条件として利用しながら解を探索する方法である．gSpanは，最小サポート制約に加えて，深さ優先探索木と最右拡張を用いる手法である．これらのアプローチは，一種の部分列挙法であり，グラフの大きさやグラフノードの種類が限定的であるときに有用な手法であるといえる．したがって，報告されている適用例の多くは，有機物質の化学構造式から有効な部分グラフを抽出する事例である．最近，ID付きPOSデータにAGMを適用し，特定商品グループに対する同時関連購買分析が報告された[6]が，ビジネス領域での適用例はまだまだ少ない．後述の本稿で対象としている問題も，形式的にはこれら既存の手法が適用可能である．このうちAGMは，利用可能なツールが公開[2]されているので，本稿で想定している問題に対して部分グラフの抽出を試みた．しかし，現在AGMが想定しているグラフの大きさの範囲外であったため，残念ながら部分グラフを抽出することができなかった．

なかはら　たかのぶ
大阪府立大学 大学院経済学研究科
〒599-8531 堺市学園町1-1
もりた　ひろゆき
大阪府立大学 経済学部
〒599-8531 堺市学園町1-1
受付05.7.25　採択05.11.14

[1] 債務が期日までに返済されないこと．

[2] データマイニングツールMUSASHIのコマンドの1つとして実装されている．http://musashi.sourceforge.jp/

一方，グラフを木構造に限定した研究としては，FREQT[7]やTreeMiner[8]，UNOT[9]などが挙げられる．これらが対象とする問題は，木に限定したデータ構造であるため，大規模なデータから部分木を抽出するという点では，本稿と共通している．しかし対象としている問題のタイプは，異なる木を識別して，それらに共通する部分木を抽出するものではない．実際に適用されている問題例としては，Webのブラウジングにおける共通した部分木の抽出などである．そのためこれらの手法では，例えば，ブラウジングした状態を大きな1つの木で表現し，その中に部分木として出現するある共通したブラウジングパターンの頻度を評価することが，その目的となる．本稿で対象としている問題は，顧客毎の購買履歴をそれぞれ独立した木として表現し，それらの木に共通する部分木として表現される共通の購買パターンを識別しようとするものである．そのためこれらの既存研究とは，適用対象としている問題のタイプが異なっているといえる．

以下ではまず，提供されたデータに対する基礎分析について説明した後，ショッピングの一括とリボルビング払いを併用するようになる顧客グループが持つ，特徴的な購買パターンを部分グラフとして抽出する分析例について述べる．

3. 分析対象データと基礎分析

分析に用いたデータは，あるクレジットカード会社の2年間のID付き購買履歴データである[3]．データのレコード数は約400万件で，顧客人数は約4万3千人である．ショッピングとキャッシングという2種類の用途，そして手数料の必要な支払い方法である分割・リボルビング払い（以下，リボと略す）と，手数料のかからないボーナス払い・一括払い（以下，一括と略す）という支払方法が存在する．それぞれの利用金額は特定できるが，購買商品についての情報はなく，利用場所や利用店舗は，1割程度しか特定できない．また，月々の返済に対する遅延情報や，性別，年齢，年収，限度額などの顧客属性が利用可能である．

表1は，年別にショッピングの支払い方法による取扱高の割合を示したものである．ショッピングを利用する顧客の主要な支払い方法は一括払いであり，全体の約90%を占めている．手数料の必要なリボ払いや分割払いは，取扱高が小さく増加していないことが確認できる．

表1　ショッピング支払い別取扱高割合

支払い方法	2002年	2003年
ショッピング一括	89.66%	91.03%
ショッピングボーナス	2.47%	1.80%
ショッピングリボ	4.68%	3.97%
ショッピング分割	3.19%	3.20%
合計金額	¥9,837,644,453	¥9,765,451,109

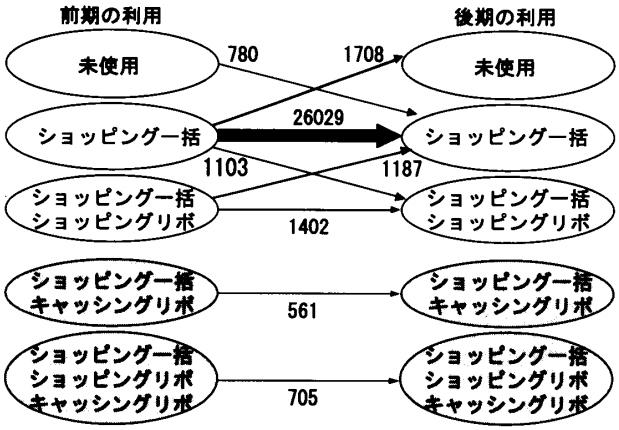

図1　年別の利用形態に対する移動人数

図1は，2002年（以下，前期と呼ぶ）から2003年（以下，後期と呼ぶ）への利用形態の移動を図にしたものである[4]．白い楕円はショッピングのみ，または未使用を，グレーの楕円はキャッシングを併用している利用形態を表している．利用形態の多くは固定的なものであり，年による利用形態の変化は少ないことがわかる．特にキャッシングの利用については，顧客の移動が1%以上起こる利用形態の変化は確認できない．

ショッピングの利用に焦点を当てると，ショッピング一括の利用が大半で，固定的であることが確認できる．注目すべき点は，ショッピング一括から後期にはリボを併用する顧客が約1,000人確認されるが，逆にリボ併用からショッピング一括のみに移動する顧客も同数程度確認できる点である．つまり，現状ではショッピングリボを利用する顧客人数は，ほとんど増加していないことがわかる．また，ショッピングリボのみという利用形態は，現状ではほとんど存在しない．したがって，ショッピングリボの利用を増やすには，ショッピング一括・リボ併用顧客を増加させる必要があり，そのためには，併用顧客の固定化と，ショッピング一括からの併用化が重要であることがわかる．

[3] 平成16年度データ解析コンペティションで提供されたデータを用いている．

[4] ただし全顧客42,787人の1%以上の移動が確認できた利用形態だけを図示している．

4. グラフマイニングを応用した提案手法と判別モデル

基礎分析より，将来リボ払いを併用する顧客を増加させることと，現在リボ払いを併用している顧客が，将来減少してしまうことが問題であると確認できた．そこでまず，併用顧客の減少の問題から考える．

前期にショッピング一括・リボを併用していた顧客が，翌年にはリボの利用を休止する顧客1,187人を「休止」，2年間継続して一括・リボを利用する顧客1,402人を「継続」とし，決定木分析を行った．説明変数は，年齢，年収などの各種顧客属性データと，前期だけの購買に関するデータを用いた．分析の結果約70%の判別が可能であり，出現したルールは，リボ利用回数が2回以上なら後期も継続して併用し，1回なら併用しなくなるというものである．このことから，継続してリボを利用してもらうには，顧客に2回以上リボの利用を体験してもらい，リボの利便性を感じてもらうことが重要ではないかと考えられる．

次に，リボ未利用顧客へのプロモーションを効率良く行えるように，リボを併用するようになる顧客の特徴を識別する．上述のリボ併用顧客の減少に関する分析の結果から，定常的にリボを利用するようになるには，年間2回以上リボを利用してもらう必要があるということがわかった．そこで，後期のリボ利用回数が2回以上の顧客を対象に，前期は一括だけを利用し，後期にリボを併用する顧客380人を「リボ併用顧客」と呼び，前期後期ともに，一括だけを利用する顧客23,939人を「一括選好顧客」と呼ぶことにする[5]．

以下の分析では，リボ併用顧客と一括選好顧客を公平に評価するため，両方のグループが一括払いのみを利用している前期の購買データだけを利用して，分析を行うことに注意されたい．また分析対象となる各顧客集合の人数は，リボ併用顧客380人と一括選好顧客23,939人である．

表2は，リボ併用顧客と一括選好顧客に違いの見られた属性を示している．この表から，リボは若年層で低所得者に利用されやすいのではないかと考えられる．しかし，部分的な違いは見られるものの，これらの値だけでは，決定木分析による精度は56%であり，

表2 属性による違い

顧客属性	一括選好	リボ併用
年齢:20代	5.88%	16.05%
年収:200万未満	8.42%	16.05%
総限度額:20万	2.47%	6.05%
勤務区分:パートアルバイト	7.89%	16.57%

全体のモデルとして，リボ併用顧客と一括選好顧客を識別できるほどの結果を得ることができなかった．

そこで他の説明変数として，両者の購買額のパターンによって何か特徴的な違いが現れるのではないかと考え，一括選好顧客とリボ併用顧客を識別する有効なパターンの抽出を行う．

4.1 ヒストリカル・ツリーを用いたコード化

購買履歴データは，ビジネス領域で様々な分析における説明変数の1つとして用いられる．購買履歴データを用いる際，購買に関する時間単位としては，1回の購買，1日，1週間など様々な時間単位で集計して利用することが可能であるが，どの時間単位で分析に用いるのが適切かを判断することは，それほど簡単ではない．

本稿では，集計期間の時間的単位を特に意識することなく，特徴的な購買額のパターンを抽出する方法を提案する．提案する方法では，まず購買額を1日（必要ならば時間，分，秒）単位から平日（月～金），休日（土日），週，月，季節，そして年といったおよそ意味を持つと思われる時間単位で購買額を集計し，これを時間的な関係が壊れないように，木構造で表現する（以下，ヒストリカル・ツリーと呼ぶことにする）．そして，各顧客のヒストリカル・ツリーから特定の顧客グループに多く共通する部分木を抽出すれば，それは，グループを識別するための説明変数として利用できる．以下では，まずヒストリカル・ツリーの作成方法について説明する．

ヒストリカル・ツリーでは，前述のように複数の時間単位で購買額を集計して木構造に表現するが，1円単位での購買額の違いにそれ程意味があるとは考えられない．また，後述のパターン抽出の観点から，あまり値が多様になることは望ましいとはいえない．そこで，各期間の購買額の合計値を少数のコードに写像することが有用であると考えられる．ただ，同じ購買額といっても，当然のことながら個人の購買能力やカード依存度によってその意味は異なる．つまり，ある日に同じ1万円を利用した場合でも，個人的には多い額を利用したと思う顧客と，それほどでもないという顧

[5] ただし後期のリボ併用顧客をリボ利用2回以上，一括利用1回以上の計3回以上としているので，一括選好顧客，リボ併用顧客ともに前期の一括利用回数が3回以上の顧客を分析の対象とする．

客がいると考えられる．そこで購買額をコード化する際に，個人の他の購買額と比較してその購買がどの程度かという個人コードと，他の顧客全体と比較してどの程度かという全体コードの2種類で，同じ購買額を以下のようにコード化する．

個人コード　期間毎に各個人の集計値を標準化し，$\mu \pm \sigma$を中心とする3つの区分コード

全体コード　期間毎に全顧客の集計値を3等分した区分コード

まず，個人コードについて説明する．上記で述べた各期間の持つ集計値を期間毎に平均$(\mu)=0$，標準偏差$(\sigma)=1$になるように標準化し，値が$(-\infty, \mu-\sigma)$のとき1，$[\mu-\sigma, \mu+\sigma]$のとき2，$(\mu+\sigma, \infty)$のとき3にコード化し，利用のない期間の値は0にする．

全体コードについては，各期間の全顧客の集計値を件数が均等になるように3分割して，値が最も小さい範囲内にあれば1，中央ならば2，大きければ3にコード化し，利用のない期間は0とする．

当初は，個人コードも全体コードと同じように件数均等配分によるコード化方法を試みた．しかしクレジット購買のためか，比較的同点の購買額が多く見受けられ，購買額の降順にソートしてコード化を行うと3が多く出現し，1の出現が少ないというコード数の偏りが確認された．個人コードに反映させたい点は，ある購買額が個人にとって通常利用する金額か，それよりも大きいのか小さいのかという点である．中心の利用価格帯を基準としたコード化の方法が，今回のデータに対してはより適切であると判断したため，個人コードと全体コードではそのコード化の方法が異なっている．

図2は上述のようにコード化したヒストリカル・ツリーのイメージである．各ノードは個人コードと全体コードを持ち，ノード数は季節レベルが4個，月レベルは季節毎に3個のノードを持つので合計12個，週レベルは各季節の最後の月が5個[6]のノードを持ち，それ以外は4個のノードを持つので合計52個のノードを持つ．また平日休日レベルは各週が平日，休日2つのノードを持つので合計104個あり，全体で172個のノードからヒストリカル・ツリーが構成されている．エッジで結ばれているノードは，それより下位のノードの購買金額を集計した値がコード化されており，この関係を保ったまま以下で示すようなコード列として利用できる．また，今回のデータに対する分析においては，1日単位で集計すると0が多すぎてデータとして冗長になるため，最小の集計単位を平日休日レベルとしている．

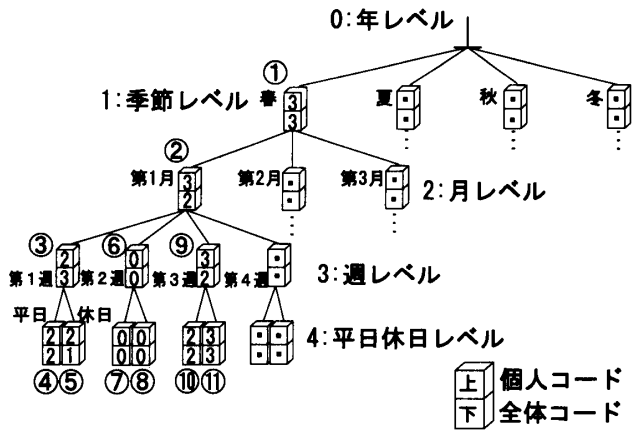

図2　ヒストリカル・ツリーのイメージ

ヒストリカル・ツリーのサイズが小さければ，既存の研究で提案されているグラフマイニングのアルゴリズムも適用可能であるが，今回のノード数は172個とかなり大きいため，特に大きな部分グラフの抽出は困難である．

そこでこれらの大規模な問題に対しても適用可能な方法として，メタ戦略の1つである遺伝的アルゴリズム（GA）を応用した方法を用いて部分パターンを抽出する．まずGAを適用するために，ヒストリカル・ツリーを遺伝子列に変換する．変換方法はルートノードから深さ優先探索順[7]にノードを辿り，その順番で個人コードと全体コードをそれぞれ1次元配列に変換する．その際，抽出した部分パターンが元のグラフのどの位置に対応するかを特定するために，位置コードとして，季節レベル(1)～平日休日レベル(4)のコードをつけておくことにする．これによって，抽出された部分パターンが元のヒストリカル・ツリーのどの部分であるか特定できる．このようにして図2のヒストリカル・ツリーを変換した例が図3となる．

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	…
個人コード	3	3	2	2	2	0	0	0	3	2	3	…
全体コード	3	2	3	2	1	0	0	0	2	2	3	…
位置コード	1	2	3	4	4	3	4	4	3	4	4	…

図3　ヒストリカル・ツリーから変換された遺伝子列

[6] 暦と各期間の関係をできる限り一致させるため，各季節の最後の月を5週として表現している．

[7] 幅優先探索順の場合，各レベルが連続して遺伝子列に変換されるため解釈が可能な部分パターンの抽出は難しく，今回のデータに対しては深さ優先探索順による結果のほうが良かった．

ここで長さ $l(l=1,\cdots,172)$ の1つの部分遺伝子列を g_l とする．そしてある顧客集合 S に対するサポートを，以下のように定義する．

$$SUP(S,g_l)=\frac{g_l\text{を含む}S\text{の要素数}}{|S|} \quad (1)$$

ここで $|\cdot|$ は集合の要素数を表すものとする．

このとき一括選好顧客集合を L，リボ併用顧客集合を LR とすると，$SUP(L,g_l)$ と $SUP(LR,g_l)$ の値が顕著に異なるような g_l が，有力な説明変数の候補となることがわかる．そこで本稿では，Max $SUP(L,g_l)$ かつ Min $SUP(LR,g_l)$ と，Min $SUP(L,g_l)$ かつ Max $SUP(LR,g_l)$ のそれぞれ2目的を同時に最適化する g_l を求める問題と考え部分パターンを抽出する[8]．

また，g_l の抽出にあたっては，完全に一致する g_l に限定すると l が大きくなるにつれ部分パターンは出現しなくなる．これらの部分パターンは，完全に一致するというよりは，その大部分が一致するものを見つけることが顧客行動の多様性を許容する観点からは重要であると考える．そこで，g_l の各コードについてワイルドカード[9]の出現を許容し，その総数は最大で l という制約条件を与え部分パターンを抽出する．

4.2 GAを用いた部分パターンの抽出

前述の2つの2目的最適化問題について多目的GAの1つの手法である[10]を応用してパレート近似解集合を抽出する．その際，データについては，一括選好顧客とリボ併用顧客それぞれ380人[10]ずつ選択し，半分のデータでパターン抽出・モデル生成を行い，残りの半分を検証用データとして用いる．[10]の特徴は，初期解集合に既存の手法を利用して有望な解を発見して採り入れ，交叉，突然変異，そして局所探索法などの解を改善するオペレータを用いて，パレート最適解の導出を目指す方法である．その際，事前にパレート最適解の数を把握することができないため，世代間で継承するエリート解[11]集合の規模を可変とし，解の探索を行う工夫を加えている．本稿で行う解の探索は，初期解はランダムに発生したものを利用するが，その基本的な考え方は[10]と同じであり，アルゴリズムの流れは図4のようになる．

部分パターンの長さは $l=1,\cdots,172$ まで考えられるが，$l=1$ の部分パターンでは，L と LR の両方のサポートが大きくなってしまい，$|SUP(L,g_l)-SUP(LR,g_l)|$ は小さくなる傾向にある．逆に l が大きくなると，L と LR の両方のサポートが小さくなり，これもまた同様の傾向を示す．本稿においては，ワイルドカードを許容しているため，全体としては，両方の集合に対してサポートは増大する傾向にあるものの，予備実験の結果から $l\geq 13$ では L と LR の両方のサポートは0に近くなり，パレート近似解は，ほとんど出現しなかった．そこで今回の分析については，両方の問題に対して $l=2,\cdots,12$ と設定し，各 l に対して同じ方法でパレート近似解を抽出する．

各部分パターンについては，ワイルドカードの数が多くなりすぎると，両方の集合におけるサポートは増大する．しかし一方で，ある程度の数を許容しなければ，l が大きくなるにつれ，急速に両方の集合に対するサポートが減少する．予備実験の結果から，ワイルドカードの上限数は l に関係のない定数で定めるより，l に比例した数を設定するほうが良いようであった．部分パターンには，個人コードと全体コードをあわせると $2l$ 個のコードが存在するため，その半数の l をワイルドカードの上限値とし，それより多くのワイルドカードを含む解は致死遺伝子として扱うことにした．

初期解は，個人コードと全体コードに関しては，＊，0～3をそれぞれ一様ランダムに各位置に発生させる．位置コードについては172の一連の位置コード

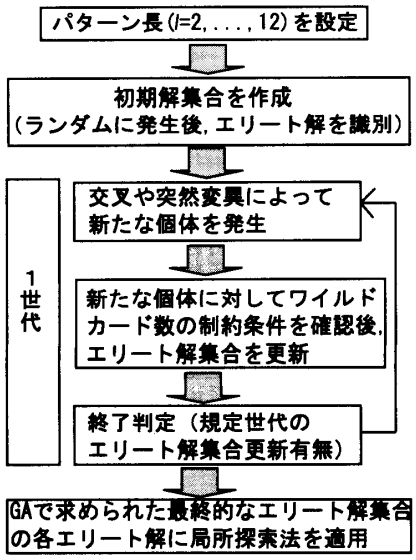

図4 GAによる部分パターン抽出の流れ

[8] $|SUP(L,g_l)-SUP(LR,g_l)|$ を最大化する単一目的問題と考えることも可能だが，部分パターンの候補は複数抽出することが望ましいため，これらの問題を解くことにした．

[9] 任意の1コード（0～3）を表すコードとする．

[10] リボ併用顧客の人数と一致させるため，一括選好顧客の中からランダムサンプリングにより選択した．

[11] 各世代における近似パレート解を意味している．

の開始位置をランダムに決定し，そこから l の長さのコードを採用することによって作成する．このようにして200個の解を作成し，そこからエリート解を識別して初期解集合とする．

交叉では，通常利用される1行の遺伝子に対する操作をそのまま適用することは難しい．本稿では，まず2つの異なる親個体をエリート解集合からランダムに選択する．その後図5に示すように，両方の親個体からまず個人コードを取り出し，1点交叉を適用することで，2種類の個人コードの候補遺伝子を作成する．全体コードについても個人コードと同様の操作によって，2種類の全体コードの候補遺伝子を作成する．位置コードは，一般の交叉操作を適用すると位置コードの順番を壊すことが多く，それによって致死遺伝子となる可能性が高まる．そのため，ここでは親個体から位置コードをそのまま複製して，2種類の位置コードを候補遺伝子とすることにした．以上のようにすると，個人コード，全体コード，そして位置コードのそれぞれに2種類の候補遺伝子を用意することができ，これらを組合せることによって，2^3 個の子個体を作成することが可能になる．

突然変異は，個人コードと全体コードについては，各コードをそれぞれそれ以外の別のコードにランダムに入れ替えることで実現される．また位置コードについては，初期解の発生時と同様の方法で新たに作成して，入れ替えることにする．以上のようにして作成された新たな解は，ワイルドカードの上限数を超えていれば致死遺伝子として廃棄され，そうでなければエリート解であるかどうかのチェックが行われる．そして新たなエリート解であれば，エリート解集合が更新される．

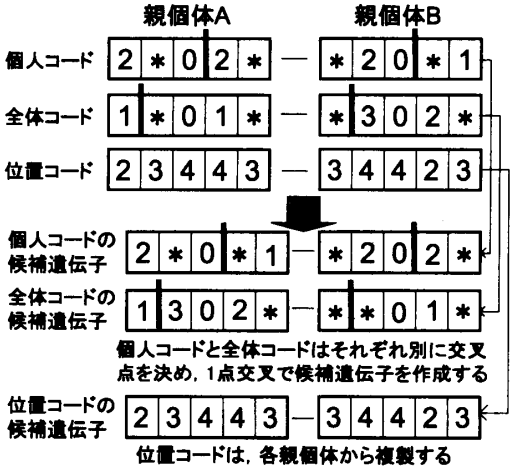

図5　交叉方法の説明

GA部分は，各世代の終了時に当該世代にエリート解集合が更新されたかどうかをチェックして，連続した100世代で更新が行われなければ終了される．GA部分のアルゴリズム終了後，各エリート解については，個人コードと全体コードについて2点交換による近傍探索が実施される．その際，いずれかの近傍解が新たなエリート解であれば，エリート解集合が更新され，そうでなければ単に廃棄される．

図6は，GAによって抽出された $l=7$ の場合のパレート近似解の例である．散布図の各点は抽出された部分パターンの1つを表し，下側に出現している点は，リボ併用顧客のサポートが一括選好顧客よりも高い部分パターンであり，上側はその逆である．このような部分パターンの集合が各 l について発見され，合計492個の部分パターンが抽出された．

4.3　ターゲット顧客識別のための判別モデル

リボ併用顧客の識別を目的変数とし，表3に示す説

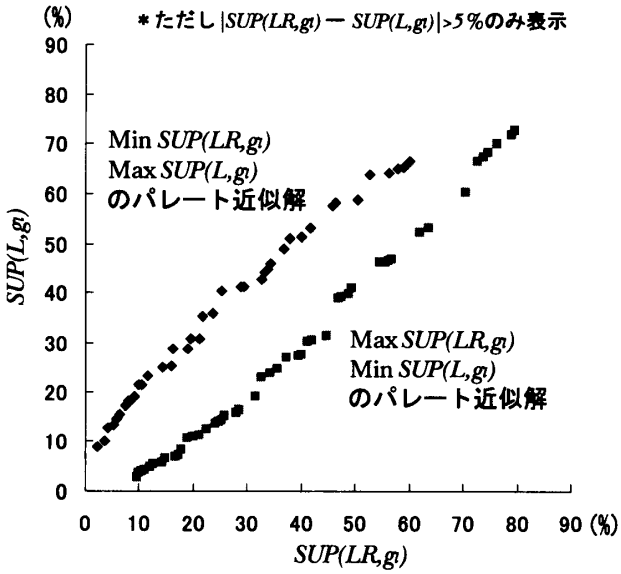

図6　GAによるパレート近似解

表3　利用した説明変数

抽出パターン	1=あり，0=なし
配偶者	1=あり，0=なし
性別	男性，女性
年齢	18～90歳
キャッシング限度額	10～80万(5万円刻み)
総限度額	10～200万(5万円刻み)
年代	10代～80代
子供人数	1～5人
勤務区分	自営業，会社員，公務員，学生，パートアルバイト，年金，その他
年収	200万未満，200，300，400万以上，500，700，1000万以上，年金
職種	事務職，技術職，営業職内勤，営業職外勤，経営者，その他

明変数を利用する．抽出パターンについては，GAで抽出した492個の部分パターンをすべて説明変数として利用しており，その部分パターンを各顧客が含むか含まないかを1，0で表している．他の属性については，表3に示す通りである．

図7は決定木により生成されたモデルである．モデルを生成するために利用したツールは，CART[11]と同様に枝の分岐基準としてGini Indexを用いて二進木を生成する．また，枝刈りはC4.5[12]と同様のError-based pruningを採用しており，図7は信頼係数CFの値を0.25に設定した場合のモデルである．精度はテストサンプル法で66%である．抽出した部分パターンを含めず，顧客属性だけでモデルを生成した場合は56%であり，抽出した部分パターンを含めることで精度が向上している．

図7の3つのパターンは，目的変数を分類する上で，他の出現していない部分パターンに比べ相対的に説明力の高い部分パターンである．抽出した部分パターンに対してモデルに出現する部分パターンの数が少ないように思えるが，これは，抽出された部分パターンが類似している[12]からだと考えられる．

まずパターン1を見ると，位置コードの3番目に月があることから，月の次のコードの週はある月の最初の週を表していることがわかる．そして月と第1週の平日の個人コードが2で，第1週の全体コードが3であり，かつそれに続く2週目の位置コードは0であることがわかる．これを解釈すると，月初の平日に個人的に通常程度の金額を購買するが，その翌週は利用を控えているような購買行動を表している．同様にパターン2を解釈すると，月末の休日に週単位で個人的に通常程度の購買を行い[13]，翌月の第一週目には購買がないパターンである．またパターン3は，月末の利用はなく，月初の平日に個人的に週および平日単位で通常程度の購買をしていることがわかる．

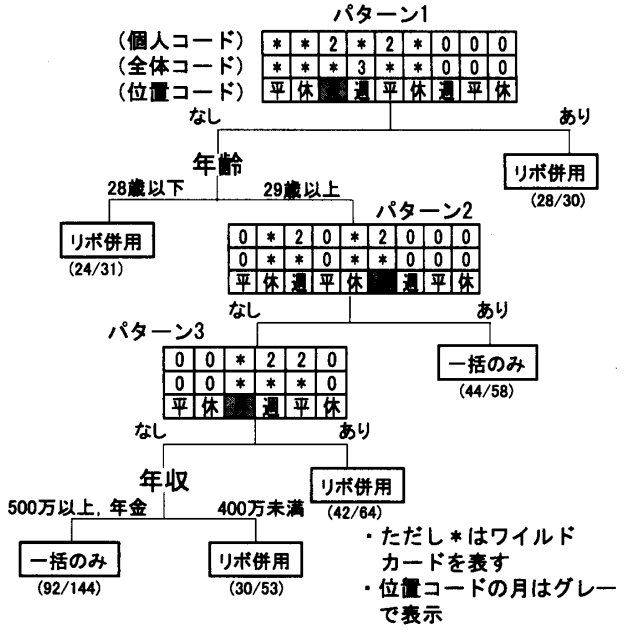

図7　決定木分析の結果

これらのパターンの意味を踏まえて決定木を解釈すると，将来リボを併用することが期待される顧客の特徴は，次のようにまとめることができる．

- 月初平日を表すパターン1や3を持つ顧客
- パターン1を持たない28歳以下の顧客
- 3つのパターンを持たない29歳以上で，年収が400万未満の顧客

逆に将来リボを併用することが期待できない顧客の特徴は，

- 月末の休日に利用し月初平日に利用しない（パターン2を持つ）29歳以上の顧客
- 3つのパターンを持たない29歳以上で，年収が500万以上または年金の顧客

である．

リボを併用することが期待される顧客の特徴として，月初平日に利用するというパターンが出現している．このパターンが出現する理由は，当該クレジットカード会社の締め日が月末であることが考えられる．つまり月初に利用すると，返済までの期間が長期になるからである．そして，平日の利用であることを考えると，生活に必要な購買をしているのではないかと予想される．また，比較的若年層で低所得者という特徴もあわせて考慮すると，リボ併用顧客は，あまり経済的な余裕があるわけではないが，購買意欲は強い顧客ではないかと解釈できる．このような顧客の特徴を把握することで，ショッピングの一括払いだけを利用している顧客の購買履歴から将来これらの特徴が確認できれば，その顧客をターゲットとして，リボルビング払いや分割払いの利用をより効率良くプロモーションできるであろう．

[12] モデルに出現しなかった部分パターンのいくつかを確認したところ，出現している部分パターンと1文字だけ異なる部分パターンが多数発見された．

[13] 月に隣接する左側の平日は，個人および全体コードは0で，その週の個人コードは2なので，その週の休日に1以上の購買があることがわかる．

5. おわりに

本稿では，購買履歴データからターゲット顧客を識別するためのパターンを発見する方法として，購買データをヒストリカル・ツリーで表現する方法を提案し，有効な部分パターンを抽出する方法を提案した．

分析例として，クレジット購買データを用いて，リボ併用顧客を識別するための判別モデルを生成し，抽出した部分パターンを利用することによって，モデル精度を約10%向上することを示した．

本稿で示した適用例は，グラフマイニングにおける新たな1つの適用例であるといえるとともに，提案手法は，クレジット購買履歴データだけでなく，他のID付きPOSデータにも十分に適用可能であると考える．今回のデータについては，利用場所や購入商品を特定することが困難だったため，全購買額についてヒストリカル・ツリーを生成した．しかし商品が識別可能であれば，特定の商品，または商品群の購買額に限定したヒストリカル・ツリーを同様に生成して，部分パターンを抽出することも可能である．また判別する顧客も特殊である必要はなく，定義可能なものであれば特に限定されない．しかしながら，完全にモデルを一般化するには，まだいくつかの自由度も含んでおり，適用例を増やしながら，一般化に向けた研究を進めたいと考えている．

参考文献

[1] 岩田昭男：「クレジット＆ローン業界ハンドブック」，東洋経済新報社，2003.

[2] 佐伯隆博：“クレジットカード市場におけるリボルビング拡大モデルの構築”，消費者金融サービス懸賞論文，消費者金融サービス研究振興協会，2004.

[3] A. Inokuchi, T. Washio and H. Motoda: “An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data,” *Proc. of PKDD 2000, LNAI 1910*, Springer-verlag, pp. 13-23, 2000.

[4] M. Kuramochi and G. Karypis: “Frequent Subgraph Discovery,” *Proc. of IEEE ICDM '01*, pp. 313-320, 2001.

[5] X. Yan and J. Han: “gSpan: Graph-Based Substructure Pattern Mining,” *Proc. of IEEE ICDM '02*, pp. 721-724, 2002.

[6] M. Kuroda, K. Yada, H. Motoda and T. Washio: “Knowledge Discovery from Consumer Behavior in an Alcohol Market by Using Graph Mining Technique,” *Proc. of Joint Workshop of Vietnamese Society of AI, SIGKBS-JSAI, ICSIPSJ and IEICE-SIGAI*, pp. 111-116, 2004.

[7] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto and S. Arikawa: “Efficient Substructure Discovery from Large Semi-structured Data,” *Proc. of SDM 2002*, pp. 158-174, 2002.

[8] M. J. Zaki: “Efficiently mining frequent trees in a forest,” *Proc. of SIGKDD '02, ACM*, pp. 71-80, 2002.

[9] S. Nakano and T. Uno: “Efficient Generation of Rooted Trees,” *NII Technical Report NII-2003-005E*, ISSN 1346-5597, National Institute of Informatics, July 2003.

[10] X. Gandibleux, H. Morita and N. Katoh: “The Supprted Solutions Used as a Genetic Information in a Population Heuristic,” in E. Zitzler et al. (editors), *1 st International Conference on EMCO, Springer-Verlag. Lecture Notes in Computer Science* No. 1993, pp. 429-442, 2001.

[11] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone: *Classification and Regression Trees*, Chapman & Hall, 1984.

[12] J. R. Quinlan: *C 4.5 Programs for Machine Learning*, Morgan Kaufmann, 1999.