論文・事例研究

データ圧縮とワイルドカードを利用した未払い履歴データに対するパターン分析

米田　知弘，森田　裕之

1. はじめに

順調な成長を続けているように見えるクレジットカード業界ではあるが，その内容を見ると，利用金額の増加状況と比してカード発行枚数は頭打ちになり，収益源のコアはキャッシング利用に依存している現状にあることがわかる（図1）．

カード会社にとって，キャッシング利用は手数料収入としては魅力的である．しかしリスクを詳細にみると，図2より貸倒償却率[1]は年々増加の一途を辿っており，デフォルトの危険が増大していることがわかる．そのためクレジット会社にとって，貸倒債権が事前に識別可能であれば，これらの損失を未然に防ぐことができ，利益率を改善することが期待できる．しかし，潜在的危険債権者を見極めることは，容易ではない．

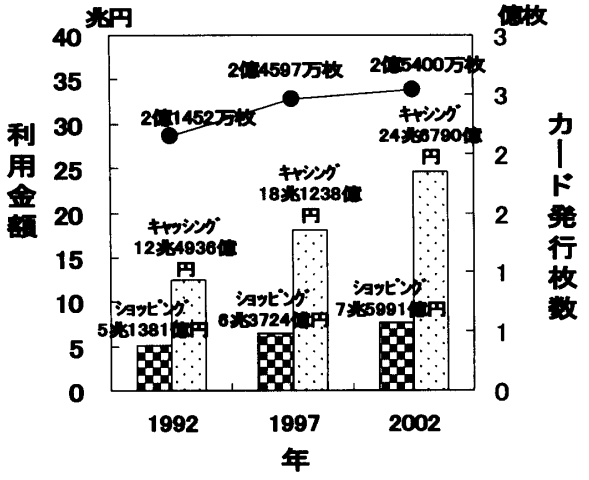


図1　クレジットカード業界概要

出所：㈳日本クレジットカード産業協会「日本の消費者信用統計」http://www.gpnet.ne.jp/card_06.html

よねだ　ともひろ
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

正常顧客を潜在的危険債権者と見誤れば，期待利益を失う危険性があり，逆に，潜在的危険顧客を正常顧客と見誤れば，貸倒損失を蒙る．

本稿では，顧客属性データと，顧客の債権に対する未払い履歴のパターンを利用して，より正確な貸倒債権を事前に判別するモデルを提案する．未払い履歴パターンの抽出については，データ圧縮で利用されるランレングス法を利用してデータ圧縮を行うと共に，ワイルドカード[2]を利用して，多少の差異を許容したパターン抽出方法を提案する．

最初に既存の関連研究について述べた後，提案方法を説明する．適用例として，クレジットカード会社の取引データを用いてパターンを利用した決定木モデルを作成し，それ以外の説明要因だけを用いるよりもモデル精度が向上したことを示す．

2. 関連する既存研究

データマイニング技術は，1990年代前半に基礎技術研究が本格化した．その急速な普及の背景として，データのデジタル化やデータベース化による豊富な対象データの蓄積，および既存の統計解析的手法と新しいデータマイニング手法の融合，そして統合化マイニ

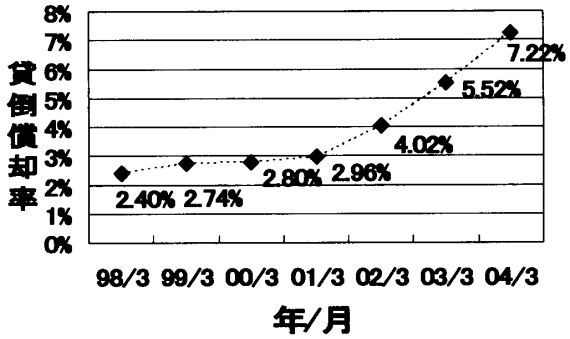


図2　貸倒償却率の推移

出所：TAPALS白書2004 消費者金融連絡会
http://www.tapals.com/archive/pdf/haku 2004_2.pdf

[1] 貸倒償却率＝貸倒償却額/期末営業貸付金残高

[2] 本稿では1文字以上の任意の文字を表すものとする．

ングツールの市販などが挙げられる[1]．このうちシーケンシャルパターンマイニングの研究分野については，ゲノム解析で用いられたBONSAIというツールが存在していた[2]．これをビジネスデータに適用するため，羽室ほか[3]は，改良を加えたE-BONSAI (Extended BONSAI) を提案し，ブランドスイッチに関する分析例などを示している．この手法の優れた特徴は，カテゴリの時系列データをモデル内で取り扱うため，より高い予測精度を持ったモデルを構築することができる点と，抽出されたルールの可読性（解釈可能性）が高く実用性に優れていることである．

これらの分析における課題は，データとしてのパターン長の増大に起因するパターンを抽出する計算コストの増大である．BONSAIでは，探索による計算コストを抑えるために，2つの工夫を行っている．1つは，元データを任意のより少ない種類の文字（インデックス）に変換を行う工夫であり，もう1つは，変換後の目的グループの識別に最適なインデックスを，ローカルサーチを利用して求める工夫である[3]．このようにして抽出されたパターンを利用して決定木を構築している．これは1つの優れた方法であるが，文字の変換により，いくつかのコードが同一に識別される点と，ローカルサーチによって最適性は保障されていないという点は改善の余地があるように思われる．

本稿では，BONSAIのアプローチとは少し異なる観点から計算量を抑制する工夫を行う．まずデータ量の増大に対しては，ある行動が行われたかどうかというデータに限定するが，ランレングス法を利用してデータ圧縮を行うことで，非常に効率よくマイニングできる工夫を加える．有効な部分パターンを抽出するためには，全体のコード長とそのコードの総数に関係した膨大な計算量が必要であり，コード長が大きな実際のデータに対して，すべての長さの部分パターンを調べることは計算時間の観点から難しいといえる．したがって本来，全体を探索すれば，より良いパターンが見つかる可能性があるものの，それが計算時間の観点からは困難な状況にある．そのため元データのコード表現を保存したままコード長を圧縮することができれば，効率よく計算でき，より多くの，長い部分パターンをも探索できることにつながる．これによって，解の質を向上させることが期待される．また抽出するパターン長が長くなると，部分的なコードの相違を許容することは，解の質を向上することに重要であると考えられる．そこで本稿では，ワイルドカードを導入し，部分的な相違を許容することによって，より長い有効な部分パターンの発見に挑戦している．

[3] 変換方法とローカルサーチの適用方法については，文献[2]を参照されたい．

3. 提案する分析のポイント

3.1 ランレングス法を利用したデータ圧縮

時系列データは，時間の経過と共にデータ量が増大する．これをそのまま利用するには，本来データが保持している意味合いを失わずに，容量を圧縮することが有益である．データの形式はさまざまだが，中でも，例えば来店購買経験のような，ある行動の有無は，0と1だけで記録することが可能であり，一般的なデータ圧縮で用いられるランレングス法を適用できる．これを利用すれば，データ本来の意味を失うことなく，容量を圧縮することが可能となる．具体的な方法としては，0と1で構成される時系列データを，0と1が連続する部分をカウントし，さらにそれを連続数にしたがって1対1の記号列に対応させることで，圧縮記号列に変換する．図3は，0の連続数をアルファベットの小文字の順番に，また1の連続数をアルファベット大文字の順番に，それぞれ対応させ，変換した例である．もちろん，圧縮記号列から元データへ復元することは可能であるとともに，圧縮記号列での部分圧縮記号列も，元データに対応させることが可能である．これによって最大n個の0または1のデータを，1つの記号列に圧縮することが可能となる．

3.2 ワイルドカードによる部分一致の許容

発見されるパターンは，完全に一致していないと意味がないという場合もあるが，逆に大部分が一致して

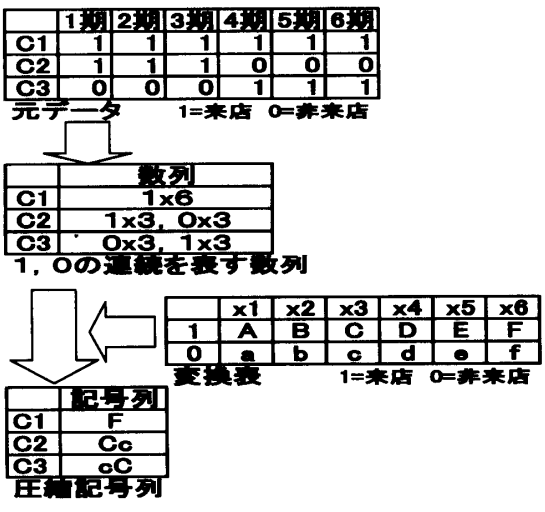


図3 記号列変換の流れ

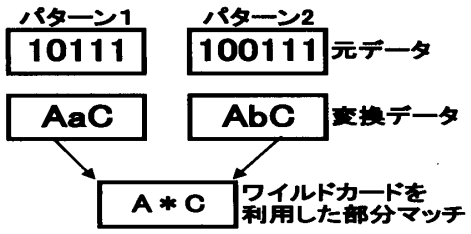


図4　ワイルドカードによる部分一致の例

いれば，多少の差異は問題ではないというケースもある．特にビジネスデータにおける顧客の行動に関係した分析である場合，全く同じ購買行動をとる顧客グループを見つけるということは困難であろう．今回分析対象とするデータは，その後者に該当するタイプである．図4は，2つのパターンを前述の方法でデータ圧縮した例である．このとき，パターン1とパターン2は，1の出現（何らかの購買行動）については全く同じ行動を示している（共通したAとCの出現）が，その間に出現した0のパターンが異なっている（aとb）．これが例えば，ある商品に対する購買の履歴であるとすると，1度購買し，1または2期間購買に間があき，その後3回連続して購買するパターンと解釈することができる．その場合，データによっては，購買と購買の間隔（この例ではaとb）に重要な意味があるという場合もあることは否定しないが，逆にaとbの違いは許容されるべき顧客行動の違いであり，この2つのパターンは同じと解釈されるべき場合もあると考える．実際，現実のデータを分析した場合，多くの顧客IDを含むデータになればなるほどデータは多様になり，それにしたがって完全に一致するパターンは，ごく限定的なものにならざるを得ない．

本稿では，そのような違いを許容して同じと解釈すべきパターンを識別するために，ワイルドカードを導入したパターン分析を提案する．図4の例では，ワイルドカードを導入することにより，“A＊C”が共通のパターンであると考えることができる．この例においては「＊」がアルファベット一文字（aとb）に一致しているが，本稿でのワイルドカード（＊）は，任意の1文字以上の文字と定義している．

4．適用データと基礎分析

適用データは，某クレジットカード会社の24ヶ月の期間にわたる，約55,000人分の日別利用データ，月別利用データ，月別残高データ，そして月別未払い履歴データである[4]．

表1　未払い有無と利用金額月別利用金額

未払回数	0回	1回以上	合計
人数	460,111 (84%)	8,953 (16%)	54,964
ショッピング利用金額	15,858 (82%)	3,523 (18%)	19,381
キャシング利用金額	2,793 (55%)	2,249 (45%)	5,042

＊利用金額単位：百万円
＊括弧内は構成割合を示す

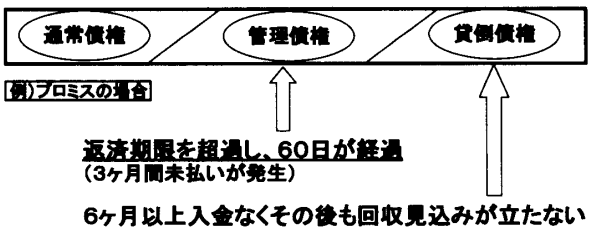


図5　債権の定義（出所[4]）

表1は全体に占める未払い経験者（1回でも未払いがある）の割合と，各グループのショッピングとキャッシングにおける利用金額の違いを表している．全体のうち約16%が未払いを経験しているが，これらの顧客の特にキャッシングの利用割合は全体の45%にも達しており，未払いの経験だけで簡単に貸倒危険顧客と考えることは，ビジネス上問題がある．したがって，より正確な分析によっての未払い経験者の中から，より危険度の大きな顧客を絞り込むことが必要である．

ここで債権の一般的な定義を確認する．図5は，クレジット業界における債権定義の1例である[4]．通常，返済期日までに入金が確認されると，「通常債権」として管理される．しかし返済期日を超過し，60日を経過すると「管理債権」と呼ばれる．更に返済期日から6ヶ月以上入金がなく，その後の回収見込みが立たないものは「貸倒債権」となり，償却処理されて損失となる．提供データからは返済期日を6ヶ月以上経過した後，回収の見込みの判断を行うことはできない．そこで本分析においては，返済期日を越えて6ヶ月以上支払を滞納した債権を貸倒債権として定義し，これを事前のデータから予測することが実際の分析目的となる．

[4] データは平成16年度データ解析コンペティションより提供された．

5. 提案手法によるパターン分析

5.1 候補パターンの抽出

利用するデータは，24か月分の各IDに対する未払いの履歴である．ある月に返済義務が存在し，当該月の返済期日に入金が完了するか，または返済義務が存在していなければ0が，逆に，返済義務が存在し，返済期日に支払が完了しなければ1が，記録される．したがって各IDは，必ず24個の0と1で構成されるデータを持つ（図6最上段）．前述の定義より，1の連続数が3回（変換後C）になれば管理債権，また連続数が6回（変換後F）になれば貸倒債権と判断できる．まず分析の最初のステップとして，未払い履歴データをランレングス法によって圧縮する．変換の様子は図6のようになり，最上段の元データが圧縮され，最下段のアルファベット列に変換される．

次に実際に判別対象とするデータについて説明する．全体の顧客データ約55,000人分から，未払い履歴が1度もつかない顧客は，絶対に貸倒にならないので除外する．また実用上，管理債権にもならない顧客に対して取引停止などの処置をすることは適当ではないので，未払い履歴回数が24ヶ月中累積で2回までの顧客は，判別対象から除外した．こうして4,567人分のデータが残った．このうち債権の定義より，以下の分析においては，記号列に「F」以上[5]を含むものを貸倒債権，含まないものを正常債権と呼ぶことにする．その結果，貸倒債権は1,812件，正常債権は2,755件と分類された．その後，属性値が不明な顧客を取り除き，件数が均等（1,391件ずつ）になるようにランダムサンプリングを行った後，決定木分析[6]を実行した．

パターンの抽出において，候補となるパターンは，純粋な組合せの観点からは膨大な数に達する．しかし，今回のデータについては，Fがつく前までの記号列の組合せ，すなわちA～Eの組合せが問題である[7]．実際の組合せを予め計算したところ，DやEが繰返し出現するケースは稀であった．

また出現する大文字アルファベットの順序を考慮し

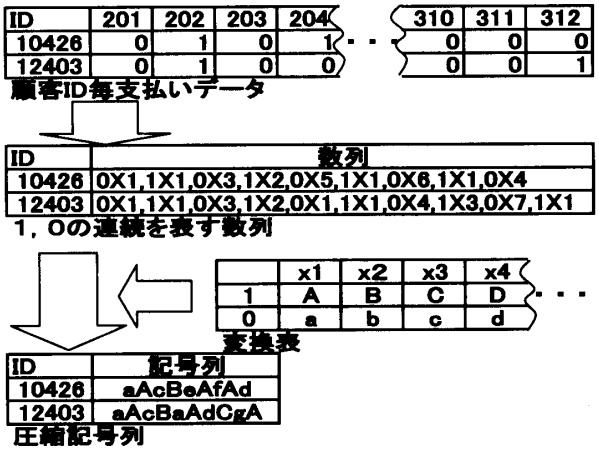


図6 顧客ID毎記号列変換の流れ

表2 候補パターン

未払行動パターン				支払行動パターン			
パターン	パターン内容	貸倒	正常	パターン	パターン内容	貸倒	正常
1	*A*A*A*	27.53%	77.64%	26	*a*	83.97%	100.00%
2	*A*B*	53.56%	72.32%	27	*b*	68.94%	100.00%
3	*C*	100.00%	42.27%	28	*c*	54.28%	99.35%
4	*A*A*A*A*	13.52%	47.81%	29	*d*	45.65%	95.69%
5	*A*A*B	27.53%	54.49%	30	*e*	37.96%	91.88%
6	*B*B*	33.57%	34.94%	31	*f*	31.56%	85.33%
7	*A*C*	53.56%	37.74%	32	*g*	26.74%	76.99%
8	*D*	100.00%	21.78%	33	*h*	22.21%	68.22%
9	*A*A*A*A*A*	4.31%	25.81%	34	*i*	18.40%	59.81%
10	*A*A*A*B*	13.52%	38.03%	35	*j*	15.82%	50.61%
11	*B*B*A*	21.21%	31.42%	36	*k*	12.29%	42.70%
12	*A*A*C*	27.53%	29.91%	37	*m*	10.50%	33.86%
13	*B*C*	33.57%	24.66%				
14	*A*D*	53.56%	19.55%				
15	*E*	100.00%	9.35%				
16	*A*A*A*A*A*A*	1.22%	12.44%				
17	*A*A*A*A*B*	4.31%	22.50%				
18	*B*B*B*	9.99%	13.16%				
19	*B*B*A*A*	11.43%	24.80%				
20	*B*D*	33.57%	14.09%				
21	*B*C*A*	21.21%	22.07%				
22	*C*C*	19.63%	10.50%				
23	*A*A*A*C*	13.52%	21.64%				
24	*A*A*D*	27.53%	16.03%				
25	*A*E*	53.56%	8.05%				

※アルファベット大文字=未払いの連続パターン
※アルファベット小文字=完済の連続パターン
※貸倒，正常は，それぞれ貸倒債権と正常債権の顧客に対する各パターンが一致した割合を表している

[5] データではG，H…も存在したが，F以上はFとした．

[6] データマイニングツールMUSASHIのコマンドxtclassifyで実行（http://musashi.sourceforge.jp）

[7] Fがついた時点で判別する目的関数と一致するため．

たパターンを当初は抽出する予定であったが，計算してみたところ一致件数が少なくなり，貸倒債権者，正常債権者両方の集合に対して一致率に差のあるような意味のあるパターンが抽出されなかった．逆に未払いの合計回数を説明変数として，事前に決定木分析も行ったが，単純な合計回数では貸倒になるかどうかを説明することは困難であった．

そこで本稿では，アルファベットの順序制約を緩和し，順不同としてパターン抽出を行うことにした．例えば累積3回の1がつく順不同のパターンは，“A＊A＊A”，“A＊B（B＊Aを含む）”，そして“C”の3つのパターンとなることがわかる．これと同様にして未払いの累積回数が1～6回となるような順不同の組合せパターンを，候補パターンとした（25パターン）．

また未払いではない0の出現パターンについては，それぞれの一致率を計算した結果，連続1年より長い期間のパターンを候補とする意味は少なかったので，ここでは1ヶ月から12ヶ月を表す，“＊a＊”～“＊m＊”の12のパターンを候補とすることにした．

これらの部分パターンが圧縮記号列と一致しているかどうかを判断する際は，ランレングス法を用いてデータを圧縮しているため，例えば圧縮記号列にdが存在する場合，部分パターンが“＊d＊”であればもちろん一致していると解釈するが，dはその前のa～cの過程を経過してdになっているため，部分パターンが“＊c＊”，“＊b＊”，“＊a＊”であっても同様に一致しているとする．これは大文字のアルファベットについても同様である．

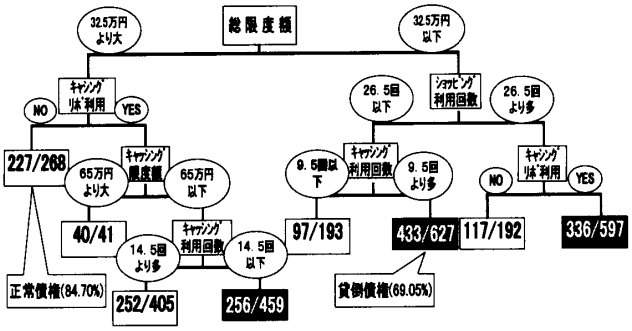
5.2 パターンを利用した判別モデル

図7は，抽出したパターンの効果を確認するために，比較対象として，顧客属性のみを説明変数とした判別モデルを表している．全体の判別精度は61.57%であり，50%ずつのランダムサンプルを出発点としているので，それほど判別精度はよくないことがわかる．図中のリーフ内の右側の数字は，それまでの条件に合致する人数を表す．また左側の数字は，黒字の場合，正常債権と判断された人数を，白字の場合は，貸倒債権と判断された人数を表す．これは図8，図9においても同様である．例えば，図7の一番左のリーフの数字“227/268”は，総限度額が32.5万円以下で，キャシングリボ利用のない人数が268人であり，そのうち正常債権であった人数が227人であったことを示している．分岐に現れたルールをみると，総限度額が一番強い説明変数として出現している．総限度額は，当該クレジット会社が設定したものであると考えられ，それは顧客の属性や取引状況から設定していると予想される．したがって，当然の結果であるとも言えるし，現状，一般的に考えられている説明変数からは，総限度額を妥当に設定しているともいえる．

次に管理債権経験までの未払いのパターンから，判別するための決定木分析を行う．その際，管理債権までの累積回数に出現文字を限定する必要があるため，表2のパターンのうちDとEを含んだパターン(8,14,15,20,24,25) は省くことにした．これら31パターンと顧客属性を説明変数として作成した判別モデルが図8である．

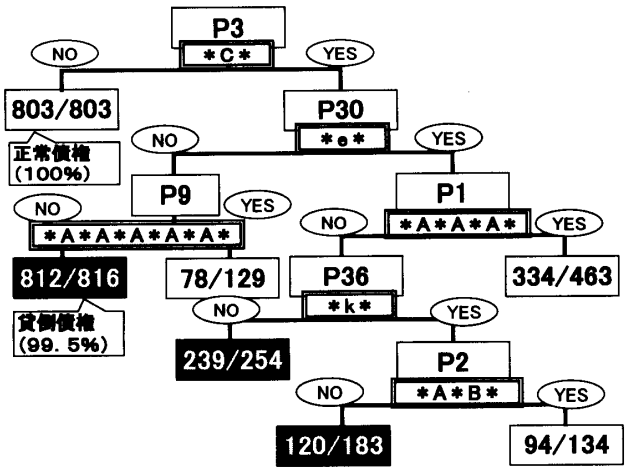
このモデルで出現したルールは，すべて抽出されたパターンであるので，属性データよりも説明力の強いパターンが存在しているといえる．またモデル精度もモデル1に比べて大きく向上している．

結果を詳細に見ると，正常債権となる顧客全体の半数以上は，3回連続未払いにはならない顧客である．一方，パターン3から右に分岐する貸倒債権になりや



＊精度61.57%（交差検証法10分割による）

図7 決定木モデル1（属性のみ説明変数）



＊精度86.62%（交差検証法10分割による）

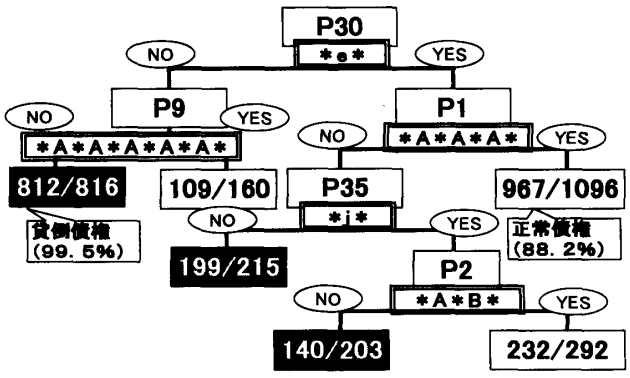
図8 決定木モデル2（未払い連続3回以下パターン）

すい顧客は，パターン30を含むかどうかがキーになっていることがわかる．つまり管理債権になっても，連続5回の支払い経験の有無が問題になり，更にパターン9に当てはまらない顧客の債権は，高い確率(99.5%)で貸倒債権になることが確認できる．さらに分岐をみると，パターン30とパターン1を含めば，ほとんど正常債権であり，逆にパターン30を含んでいてパターン36を含まないと，貸倒の危険性が高くなる．

これらの結果を総合的に解釈すると，管理債権になった時点で，この顧客が将来貸倒になるかどうかということを見極めるポイントは，それまでの連続した返済回数の長さと，未払いの状態を連続させない経験の回数であるといえる．すなわちこれはある程度の（例えば5ヶ月以上）月数，支払いを滞りなく実施しているという経験が，1つの経済的な安定性を示し，未払いを連続させないという経験の回数が，買い過ぎとその次の月の買い過ぎを調整する消費者の購買に対する健全性を示しているのではないかと推測できる．

逆にいえば，今まで未払いが発生すらしていなかったのに，管理債権になったような顧客は危険率が高いということである．これは様々な理由が考えられるが，極端な消費行動の変化や収入の減少など，顧客個人にとって，それ以前とは違う経済や消費の状況になった際に，出現するパターンではないかと推測される．

実際の現場においては，できるだけ未払い履歴が少ない回数で判断できたほうがよい．そこでモデル2での31の候補パターンから，さらに“C”がついている7つのパターン（3，7，12，13，21，22，23）をのぞいた24のパターンだけを説明変数として加え，属性データとともに判別モデルを作成したのが図9である．



＊精度88.10%（交差検証法10分割による）

図9　決定木モデル3（未払い連続2回以下パターン）

結論としては，モデル2と同様の結果を示していることがわかる．またモデル2での最初の分岐条件であったパターン3を省いたものの，モデルの精度自体にそれほど変化がない．つまりこれはパターン30とパターン3が，多くの一致する同様の顧客を含んでいたことがその理由であるといえる．その証拠に，モデル2とモデル3のパターン30の左側の分岐に該当している顧客の数に，それほど大きな変化は確認されない．このことから，顧客の未払い履歴を観察していれば，管理債権となる前に，すなわち未払い履歴が連続2回までの時点で，それまでの連続支払い月数と，不連続の未払い回数によって，顧客が将来貸倒になるかどうかを判断できることを示しているといえる．

6．まとめ

本稿では，ある顧客行動が0と1で表現されている時系列データの1つの例として，クレジット購買における未払い履歴データを用いて，有効なパターン分析の1つの例を示した．パターン抽出においては，ランレングス法を応用してデータ容量の圧縮を図ると共に，ワイルドカードを利用してパターンのあいまい性も考慮した抽出方法を提案した．抽出されたパターンと顧客の属性データから貸倒債権となるか，正常債権となるかを判別する決定木モデルを構築し，属性データのみを用いてモデルを構築する場合に比べて約25%モデル精度を向上する結果を示した．

分析から得られた結論として，ある顧客が将来貸倒債権となるかどうかは，それまでの支払いの経験が重要であることがわかった．すなわち，未払いが連続しても，それまでに連続して支払いを完了している月数が一定数以上存在していること，そしてもう1つは，未払いが発生していても，それを連続させず，必ず未払いが発生した次の月には支払いを完了していることが，貸倒債権とはならない健全性を示すバロメータであるといえる．

以上の結果は，ランレングス法を用いたデータ圧縮と，ワイルドカード導入によって比較的長い部分パターンを発見することによって得られた効果であると考えられる．今後の課題としては，今回の結果が分析事例の特殊性であるとも考えられるので，他のデータに対する適用も試み，よりその効果が一般的であることを確認したいと考えている．

参考文献

[1] 鷲尾隆, "グラフベースデータマイニングの基礎と現状", 情報処理学会誌, Vol. 46, No. 1, pp. 20-26, 2005.

[2] Shinichi Shimozono, Ayumi shinohara, Takeshi shinohara, Satoru Miyano, Satoru Kuhara and Setsuo Arikawa : "Knowledge Acquisition from Amino Acid Sequence by Machine Learning System BONSAI", *Trans. Information Processing Society of Japan*, Vol. 35, pp. 2009-2018, 1994.

[3] Yukinobu Hamuro, Naoki Katoh, Edward H. Ip, Stephane L. Cheung and Katsutoshi Yada : "Combining Information Fusion with String Pattern Analysis : A New Method for Predicting Future Purchase Behavior", *V. Torra (ed.), Information Fusion in Data Mining, Studies in Fuzziness and Soft Computing*, Vol. 123, Springer, pp. 161-187, 2003.

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