# 金融業におけるデータマイニングの応用
## —保険解約の防止分析—

小野　潔

## 1. はじめに

金融業は，新しいマーケティングとリスク管理のツールとして，データマイニングに注目している．本報告は，金融業におけるデータマイニングの位置づけと分析手法を紹介する．また，分析精度の向上のために金融商品のRFM取引属性を提案し，さらに「探索のレベル」と「業務目的」との関係について考察する．適用事例は，生命保険の解約防止分析を説明する．

## 2. 金融業におけるデータマイニング

### 2.1. 金融業のOne to Oneマーケティング

21世紀の日本の金融業は，マス・マーケティングからOne to Oneマーケティングへと大きく戦略を変更しようとしている．金融業のOne to Oneマーケティングの目的は，顧客ごとのライフサイクルをとらえ，顧客の求める可能性の高い，あらゆる商品や情報を顧客に提供し，顧客の生涯価値（Lifetime Value）における自社の金融商品を高めることである．そのために総合金融機関は，クロスセールスを中心としたリテール戦略を採用し，顧客囲い込みを試みている．クロスセールスとは，同一顧客へ単一の金融商品だけでなく，グループ会社の複数の金融商品（ローン・保険・株式・カード・投資信託等）を販売することである．また，金融商品の寿命をとらえ，金融商品が陳腐化する前に，新しい商品を顧客へ勧めることを目指す．このコンセプトは古くから存在したが，ようやくコンピュータの発達とデータマイニング技術の進歩で顧客選別や商品寿命予測が実現可能になった．

One to Oneマーケティングを実現するには，顧客ごとにライフサイクルを把握する必要がある．ライフサイクルのどのカテゴリーにいるかによって顧客の住宅取得意欲や貯蓄に関する意識が異なるためである．また，金融商品は買い替えがつきものであるが，顧客ごとに買い替えるタイミングが異なるため，顧客ごとの商品寿命を推測する必要がある．幸いなことに金融業には，顧客に関する多くのデータが蓄積されているため，データマイニングで顧客行動を推測する試みが始まった．

最近，日本版401k型年金が話題になっているが，これはデータマイニングにとっても重要な分析対象である．401k型年金の導入は，金融機関に生涯にわたる個人情報と年金情報をもたらす．401k型年金のデータベースに収められる情報には，個人属性だけでなく，年金の運用方針の変更や転職等の情報がある．年金情報のみでは，有益な情報を得ることはできないが，各企業が保有する個人情報と合わせることで可能になる．

今後は，顧客セグメントと地理情報システムを結びつけることにより，エリア・マーケティングの利用が期待されている．

### 2.2. 金融業の適用分野

米国では，1994年頃から流通業や金融業でデータマイニングの成功例が報告されている．日本の金融業でも，最近，成功例が発表されてきた．1997～2000年にかけて日本のクレジット・カード会社約50社が，国際犯罪グループに対処するため，ニューラルネットワークを用いたカード不正使用探索ソフトを米国から導入した．外資系損保会社は，自動車種別や顧客属性と事故率の関係を決定木で分析し，リスク細分型自動車保険を開発した．銀行は，決定木を利用した住宅ローンモデルを開発した．昨年は，決定木併用型のスコアリングモデルによる中小企業向けの融資モデルが，注目を集めている．生命保険会社は，保険の解約防止のために決定木分析を行っている．また，介護保険制度の要介護認定に使われる1次判定ソフトにデータマイニングの技術が利用されている．表1と表2にデータマイニングの金融業の適用範囲をまとめる．

おの　きよし
株式会社ニッセイ基礎研究所 金融研究部門
〒100-0006 千代田区有楽町1-1-1


© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

表1　金融業で既にデータマイニングが適用された分野

| 業種 | 適用分野 | 目　的 | 分析手法 |
|---|---|---|---|
| 銀行<br>生保 | 個人ローンの与信審査の推定 | 個人の住宅ローンやアパートローンの融資の可否を顧客の属性から推定し、業務の効率化を計る。 | 決定木<br>ニューロ |
| 銀行 | 住宅ローンの見込み客の推定 | 住宅の購入見込み顧客や住宅ローンの借り替え見込み顧客を特定し、ダイレクトメールで勧誘する。 | 決定木<br>MBR |
| 銀行 | 銀行の商品組合せ | 定期預金・ローン・公共料金振込等の組合せを分析し、顧客ニーズの発生タイミングに合せて、商品を提供する。 | 相関ルール<br>クラスタリング |
| 生保 | 生命保険解約の防止 | 解約された契約の属性を把握し、解約予備群を発見し、解約による保有減少を抑制する。 | 決定木<br>クラスタリング |
| 損保 | リスク細分型の自動車保険 | 年齢、年間走行距離、地域、使用目的、車の形状等から、割安になる自動車保険を開発する。 | 決定木<br>MBR |
| 証券 | 顧客管理分析 | 証券顧客と営業マンとのトラブルを、証券の売買属性と顧客属性から事前に予測する。 | 決定木<br>ニューロ |
| 証券<br>生保 | 社債格付け推測 | 格付け機関の発表する格付けと、その企業の財務項目・指標の関係から、他の社債の格付けを推測する。 | ニューロ<br>MBR |
| カード | ダイレクトメール販売の活用 | 加盟店と顧客の行動を分析し、ダイレクトメールのヒット率を上げる。 | 相関ルール<br>決定木 |
| 消費<br>金融 | 消費者ローンの与信審査分析 | 顧客属性、勤務先、キャシング履歴、自動車保険、保険証から、与信審査する。 | 決定木<br>ニューロ |
| カード | クレジット・カードの不正利用防止 | 購入金額、日付、店の業種からカードの不正利用を発見する。 | ニューロ<br>決定木 |

表2　今後データマイニングの適用が見込まれる分野

| 業種 | 適用分野 | 目　的 | 手　法 |
|---|---|---|---|
| 全業種 | セット商品開発の活用 | 顧客情報に基づいて生命保険・損害保険・投信等の商品セットをして、新たな契約獲得に結びつける。 | 相関ルール |
| 生保<br>損保 | 優良顧客の選択 | 継続率の良い保険契約を分析して、優良顧客の獲得に重点をおく。 | 決定木<br>クラスタリング |
| 生保<br>損保 | 保険金・給付金不正請求チェック | 過去の不正請求パターンと属性の関係から、不正請求の被害削減・不正請求の調査費用を削減する。 | 決定木<br>ニューロ |
| 生保 | 介護保険法の介護ニーズの分析 | 介護のアセスメント項目の分析し、介護ニーズにあわせたケアプランを選択する。 | 決定木<br>MBR |
| 銀行<br>生保 | 企業倒産リスクの定量モデル | 企業倒産を分析し、財務項目から予測モデルを開発する。 | ニューロ |
| 不動産 | マンション賃貸料の推定 | マンションの賃貸価格を推定し、将来の資産価値を計算する。 | 決定木<br>ニューロ |
| カード | カード業界の商品分析 | 利用店の業種の組合せ分析により、顧客の趣向性を推測し、販売に役立たせる。 | 相関ルール<br>クラスタリング |
| 銀行 | マネー・ロータリングの発見 | 不審な口座取引から、犯罪と関係するマネー・ロータリングを早期に発見する。 | 決定木<br>MBR ニューロ |

## 3.　データマイニング概論

### 3.1.　分析手法の種類

よく使われるデータマイニングの分析手法は，決定木，ニューラルネットワーク，記憶ベース推論(Memory-Based Reasoning: MBR)，相関ルール，クラスタ分析等がある．

タイプ別では，クラス分類分析（Classification）が，顧客のセグメントに利用できる．手法には，決定木分析とニューラルネットワークがある．クラスタリング分析（Clustering）は，データを性質の似た複数のグループに分類できる．手法は，クラスタ分析K-means法・Ward法とコホーネン・ネットワークがある．相関関係分析（Association）は，購入商品の組み合わせパターンを発見する．手法は，相関ルールと時系列関係も探索する時系列順位分析（Sequential Patterns）がある．類似パターン分析は，記憶ベース推論が使われる．

### 3.2.　RFM属性

金融業の業務データは，顧客データと取引データか


© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

表3　金融商品のRFM取引属性

| RFM属性 | 元の定義 | 金融商品のRFM属性の例 |
|---|---|---|
| R(recency) | 最後の購買日から現在までの日数 | 最終購入日や契約日から現在までの期間(年数&月数&日数) |
| F(frequency) | 累計購入回数 | 保険の加入総数、投信の購入回数 等 |
| M(monetary) | 累計購入金額 | 契約金額の合計、保険金額の合計、投信金額の合計 等 |

ら構成されるが，このデータのみでは高い判別効率が得られない．精度を上げるには，単独取引の属性だけでなく，複数取引を把握する属性が必要である．従来は経験に基づいて，取引データから新しい属性を作るが，本研究では，金融商品ごとのRFM属性を作成することを提案する．「商品毎」と「全商品」と「類似商品毎」のRFM属性を基本属性に追加すれば，分析用の基本データベースが機械的に開発できる．表3にRFM属性の定義を示す．

RFM属性は，もともとデータベースマーケティングに用いられる．そこでは，マーケティングの顧客の優先順位を，RFM属性値に重みを掛けた合計値で求める．しかし，顧客属性を考慮できない点や経験に基づく重みの決定方法に問題がある．データマイニングはそれらの問題点を解決し，細かい顧客分類が可能である．

### 3.3.　探索の深さと業務目的

データマイニングの探索には，「浅い探索」と「深い探索」がある．「浅い探索」は，顧客属性と取引属性と金融商品のRFM属性から分析し，最終ゴールは業務の専門家が満足するシナリオである．それに対して「深い探索」は，新しい属性の作成や，ハイブリッドモデルを取り扱うために，経験と技術が必要とされるが，マイクロ・マーケティングや新しいルールの発見が期待できる．

分析結果を業務へ反映する場合，「深い探索」の分析結果が必ずしも「浅い探索」より優れているとは限らない．例えば，「保険解約の防止」や「カード会員へのダイレクトメール」は，『お客とのコミュニケーション』を目的とする．このようなときは，極端に顧客を限定する「深い探索」より，やや広範囲の顧客を対象にした方が業務の目的にかなっている．反対に与信審査を「浅い探索」で分析しただけでは，審査が甘くなるため，大きな損害が発生する．要するに業務分析では，「分析結果の精度」や「探索レベル」を業務目的から決定するべきであり，分析技術や結果から決定すべきでない．

マイニング・ツールは，手間のかかるKDD (Knowledge Discovery in Databases) プロセスを簡単に扱えるため，作業効率は3倍～5倍以上になる．そのため，ツールは金融機関の業務モデル開発に必要不可欠である．しかし，現状のマイニング・ツールでは，意思決定機能が未成熟であり，未だ「深い探索」は分析者の能力に依存している．

### 3.4.　「深い探索」の分析手法

分析モデルのパラメータをいくら調整しても，1種類の分析手法の精度向上には限界がある．そこで「深い探索」の分析では，データ分布の変更・クロスバリデーション・ハイブリッドモデルが用いられる．ハイブリッドモデルは，複数の分析手法を「直列に組み合わせるカスケードモデル」と「連続に組み合わせるキャタラクトモデル」と「並列に組み合わせるコミィティモデル」がある．

①カスケードモデル (Cascade Model)

カスケードモデルは，分析手法を逐次的に適用し，多段階でデータマイニングを行う．実務では，最初に決定木やクラスタ分析やコホーネンネットワークで分類し，次に分類したクラスごとにロジスティック回帰やニューラルネットワークを適用する．このモデルは，精度をあげるのに適しているが，クラスごとに違ったモデルを作成するため，モデル数が多くなる．このモデルは，高い精度を必要とされる与信審査モデルに使われる．

②キャタラクトモデル (Cataract Model)

キャタラクトモデルは，連続に複数の違った分析手法を適用する．1種類の分析手法では，判別可能顧客だけでなく，多くの判別不能顧客も発生する．このモデルは，判別できない灰色顧客に他の分析手法を適用して，判別不能顧客を少なくする方法である．最後は，判別可能の顧客を集めて結論とする．

③コミィティモデル (Committee Model)

コミィティモデルは，同じデータに対して複数の分析手法を同時に試みる．分析手法ごとに，顧客の分類や優先順位に相違が発生するが，このモデルは多数決・線形結合・平均値・最大値・最小値を用いて，1個の解に収束させる．中小企業の融資モデルで使われる決定木併用型のスコアリングモデルは，コミィティモデルの一つである．


© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

### 3.5. 金融業のデータ不足問題

①負事例のデータ不足問題

データマイニングは，正事例（positive set）と負事例（negative set）の共通属性を対比するため，負事例が存在しなければ分析できない．しかし，金融機関は負事例データを保持しないことが多い．例えば与信審査では，正事例が「契約に至った顧客」であり，負事例が「契約をお断りした顧客」である．金融機関は，契約に至らなかった負事例データをわざわざデータベースに蓄積しないことが多い．また負事例データが存在しても，正事例に比べて属性数が少ないため，正事例の属性データを有効に使えない問題が発生する．

②リアルの顧客属性データの不足問題

銀行・生命保険会社は，口座開設時や契約時の詳細な顧客属性データ（年収，勤務会社，病気等）を保有する．しかし，データの更新機会が少ないために，分析で使えるリアルの顧客属性が「性別」と「年齢」しか存在しないことがある．

それに対して，損害保険や消費者金融会社は，契約更新のたびに顧客データを更新するため，多くのリアルの顧客属性データを使ってデータマイニングができる．

## 4. 決定木分析

### 4.1. 決定木の利得基準

決定木は属性の利得基準に基いて，分割属性の優先順位が決まる．利得基準は，ルールが目的属性値の分布に与える影響度合いを数値化したものである．よく使われるアルゴリズムとその利得基準を表4にまとめる．

データ集合 $L$ に，$j$ 個のカテゴリー値をもつ目標属性が存在し，集合 $L$ 内に $i$ 番目の値をもつデータがそれぞれ $X_i(L)$ 個 $(i=1,\cdots,j)$ あると仮定する．ルールRで $L_1$ と $L_2$ に2分割し，部分集合 $L_1$ 内の $i$ 番目の値の分布比率を $p_i(L_1)=X_i(L_1)/|L_1|$ とすると，各分割基準は次のように定義できる．なお，情報エントロピーの利得基準は，多数値を取る属性が偏重されるため，C4.5以降は，利得比基準を用いる．

表4　決定木の種類

| アルゴリズム名 | 利得基準 |
|---|---|
| C5.0　C4.5　ID3 | 情報エントロピー関数 |
| CART | GINI 関数 |
| CHAID | カイ2乗関数 |

①情報エントロピー関数による利得基準

$$Ent(R)=-\sum_{i=1}^{k}p_i(L)\log p_i(L)$$
$$-\left(-\frac{|L_1|}{|L|}\sum_{i=1}^{k}p_i(L_1)\log p_i(L_1)\right.$$
$$\left.-\frac{|L_2|}{|L|}\sum_{i=1}^{k}p_i(L_2)\log p_i(L_2)\right)$$

② GINI 関数による利得基準

$$Gini(R)=\left(1-\sum_{i=1}^{k}p_i(L)^2\right)$$
$$-\frac{|L_1|}{|L|}\left(1-\sum_{i=1}^{k}p_i(L_1)^2)\right)$$
$$-\frac{|L_2|}{|L|}\left(1-\sum_{i=1}^{k}p_i(L_2)^2\right)$$

③カイ2乗関数による利得基準

$$Chi(R)$$
$$=\sum_{i=1}^{k}\frac{|L_1|(p_i(L_1)-p_i(L))^2+|L_2|(p_i(L_2)-p_i(L))^2}{p_i(L)}$$

### 4.2. 探索アルゴリズム

①判別対象の目標属性（ターゲット）を決定する．

②すべての説明属性について，分割後の利得基準を計算する．

③最大の利得基準を有する説明属性により，分割を実行する．

④分割後の集合（サブツリー）に，②・③を繰り返す．

⑤アルゴリズムの終了は，利得基準やカテゴリに含まれるサンプル数で判断する．

## 5. 応用事例：保険解約の防止分析

### 5.1. データ条件

マイニング・ツールのSAS/Enterprise Minerを使用して，保険解約の防止分析を試みた．防止分析の目的は，保険解約者を判断力の高い属性の組み合わせで，保険解約予備軍の顧客パターンをグループ分類し，業務へ反映することである．

分析する属性を表5にまとめる．契約変更属性は契約変更や配当金引き出しの情報である．営業職員属性は契約者の担当営業職員の属性であり，営業職員とその顧客の因果関係を調べるために，分析対象に加えた．

分析データベースは，3営業店の保険契約の契約データ（6万件，98年9月末）と解約データ（2万件，5年分）を用いて，8万レコード，属性数は200個以上を開発した．分析前に情報エントロピー値を用いて不要な説明属性の絞り込みを行った結果，35属性に絞り込めた．


© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

表5　保険解約の属性

| 属　性 | 基本データ | 内　容 |
|---|---|---|
| 目標属性 | 加工データ | 契約継続、解約 |
| 顧客属性 | 元データ | 性別、生年月日　等 |
| 契約（取引）属性 | 元データ | 保険種類、保険金額、保険料、契約年齢　等 |
| 上記のRFM属性 | 加工データ | 契約期間、契約回数、保険金額合計　等 |
| 契約変更属性 | 元データ | 契約変更の内容、配当金額 等 |
| 上記のRFM属性 | 加工データ | 変更からの日数、月数、年数　等 |
| 営業職員属性 | 元データ | 担当営業職員の役職、勤務年数　等 |

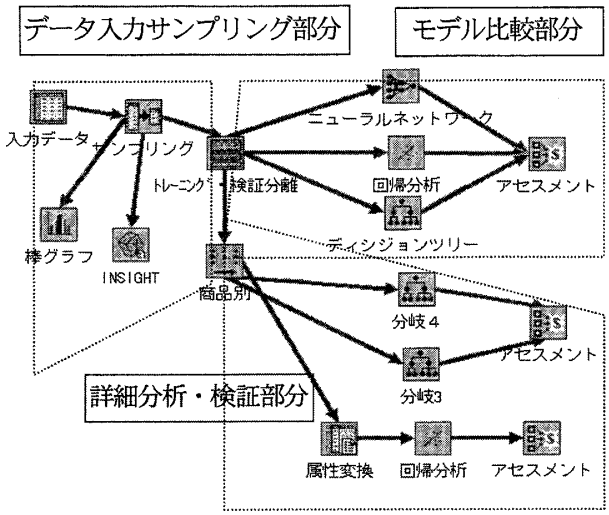


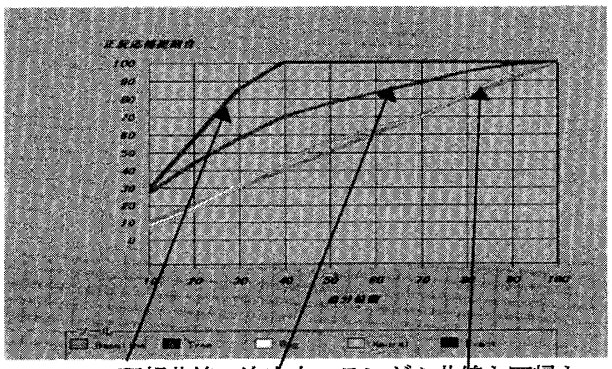

図1　解約防止分析のプロセス・フロー・ダイアグラム

訓練・検証データは，ランダム・サンプリングで採取し，1：1の配分にした．後にクロスバリデーションを用いて，モデルの安定性を検証した．

### 5.2.　プロセス・フロー・ダイアグラム

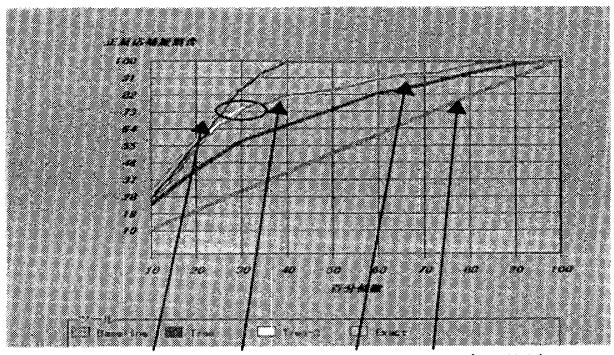
SAS/Enterprise Minerは，データ属性・分析手法選択・パラメータ選択・データフロー等を，図1のプロセス・フロー・ダイアグラムで制御する．四角箱をノードと呼び，サブルーチンの役目を果たす．左上部は，データ入力・サンプリングである．右上部は複数の分析手法を比較調査する．下部は商品別に詳細な分析を行い，同時に結果の検証も行う．

左端の[入力データ]ノードは目標属性を指定し，[サンプリング]ノードでは，入力データとしてサンプリング方法とデータ数を決める．[(多次元) 棒グラフ]と[INSIGHT]ノードでは，分布を調べ異常値や特殊パターンを探索する．[アセスメント]ノードでは，複数モデルの正解率をグラフ表示で比較できる．詳細な分析は，[グループ別]ノードを用いて，商品毎に分析する．



理想曲線、決定木、ランダム曲線と回帰とニューロは、ほぼ同じ曲線上ある

図2　分析手法の比較



理想曲線、商品2、商品1、ランダム曲線

図3　商品別の決定木

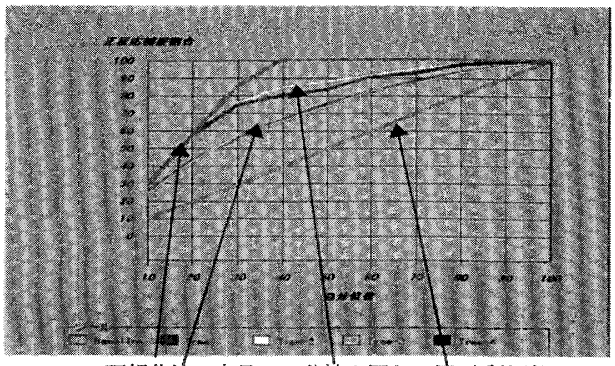
最後に決定木のパラメータ（分岐数と深さ等）を変化させる．

### 5.3.　分析手法の選択

ニューラルネットワーク，ロジスティク回帰，決定木の判別効率を正反応補足曲線で比較する（図2）．縦軸が正反応補足率であり，横軸が顧客の占有率である．

商品2の正反応補足率は，横軸30%のとき縦軸の正反応補足割合が75%である（図3の丸印）．この意味は，全契約者の30%にダイレクトメールを発送するだけで，解約予備群の75%へ郵送できることを示している．理想曲線（＝正解率100%）は，原点とターゲット最大比率[(実際の解約数)÷(全データ数)＝40%（横軸)]を結んだ直線である．ランダム曲線は0%と100%を結ぶほぼ直線となる．そのため，開発モデルの正反応補足曲線は，理想曲線とランダム曲線の間に存在する．図2から決定木分析が最適であり，図3から商品2が商品1より予想しやすいことがわかる．


© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．



理想曲線、商品2の分岐3個と8個（重複線）、
商品1の分岐3個と8個（重複線）、ランダム曲線

図4　パラメータによる影響

### 5.4.　パラメータ調整

次に，モデルのパラメータを調整して最適値を求める．決定木のパラメータには，3種類の利得基準（情報エントロピー値，GINI基準値，カイ2乗値），最大分割数，木の深さ等がある．サブツリーの分割数やツリーの深さを増やすと，グループが細分化され，ルールがわかりづらくなる．経験則では，最大分割数は7個以下が人間に理解しやすい．

図4は，商品1と2に対して分岐数の最大値を8個から3個に変化させた場合の結果を示したものである．商品ごとに曲線が重なり，分岐の最大値の変化による精度の差がほとんどなかった．しかし分岐数を変化させると，ツリーの構造が大きく変化し，導出ルールが大きく変わる．そこで業務の専門家がルールを検証し，最適なツリーを決定した．

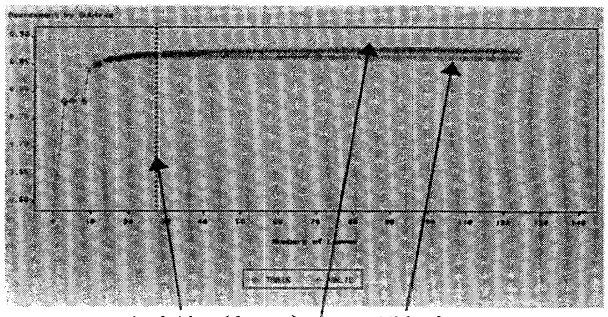
### 5.5.　収束条件

決定木は，葉数（分岐数）が増えるにつれて，分類が細かくなるので，正解率は向上するが，ある葉数以上では正解率が収束する．収束枚数が少ないほど，モデルの構造は柔軟であり，データの変動にも影響を受けづらくなる．

図5では，訓練・検証データに適用したモデルが共に，28枚で収束している．この葉数ならばツリーの階層は，3～4層となり，人間に理解しやすい．また両者は同じように収束しており，安定的なモデルである．

### 5.6.　検証結果

モデルの推測結果と実際の解約数を表6に示す．モデルの完成度は，次の指標を用いる．通用度は，モデルが解約顧客をどの程度予測するかを示す．精度は，モデルがどの程度正しいかを示す．正解度は，全体に占めるモデルの正解率を示す．



収束線、検証データ、訓練データ

図5　収束の葉数

表6　商品2の検証結果（括弧内の数字は人数）

| | 実際継続 | 実際解約 | 合　計 |
|---|---|---|---|
| モデル継続 | 62%(8744) | 2%(270) | 64%(9014) |
| モデル解約 | 12%(1602) | 24%(3293) | 36%(4895) |
| 合計 | 74%(10346) | 26%(3563) | 100%(13909) |

通用度＝3293/3563＝92%

精　度＝3293/4895＝67%

正解度＝62%＋24%＝86%

以上から，業務的に満足できる結果が得られたと言える．

### 5.7.　業務への反映

データマイニングの目的は，精度の良い分析を行うことはもちろんであるが，分析結果を業務へ反映させて業績を向上させることも重要である．保険の解約予備群団に有効な対策には，①営業職員の面接対応，②ダイレクトメール，③顧客電話サービス，④社内キャンペーンがある．地区を限定して実地調査し，その結果をさらにデータマイニングで分析すれば，顧客ごとの有効なアプローチを特定できる．

## 6.　おわりに

金融業におけるデータマイニングの応用の一例として，保険解約の防止を紹介した．現在の金融業は，ビジネスに役立つ情報を引き出すツールとして，データマイニングに期待している．今後は，金融業のIT革命が進みさらにデータが蓄積されるため，データマイニングの需要がますます増加する傾向にある．

### 参考文献


Joseph P. Bigus, “Data Mining with Neural Networks”, The McGraw-Hill Companies, 1996.

J. R. キンラン, “AIによるデータ解析”, トッパン, 1995.

大滝厚, 堀江宥治, “応用2進木解析法”, 日科技連, 1998.



© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．