1999年度日本オペレーションズ・リサーチ学会
秋季研究発表会

金融業界におけるデータマイニングの応用 ～応用事例：「保険解約の防止分析」～

ニッセイ基礎研究所 金融研究部門　小野　潔 *

最近の金融業界は、新しいマーケティングとリスク管理のツールとして、データマイニングを注目している。本報告は、金融業界におけるデータマイニングの位置づけとその分析手法を紹介し、適用例として保険解約の防止分析を説明する。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「浅い探索」と「深い探索」の定義と業務目的との関係について考察する。

1. 金融業界におけるデータマイニング

1.1. 金融業界の One to One マーケティング

21 世紀の日本の金融業界は、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。金融業界の One to One マーケティングは、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が気づく前に提案し、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、クロスセールスが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も含む。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と金融商品の寿命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。また金融商品には、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界には、顧客に関する多く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。

1.2. 金融業界の適用分野

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。日本の金融業界でも、最近、成功例が発表されてきた。1997 年、大手のクレジット・カード会社７社が、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。外資系損保会社の数社と一部の中小損保会社では、自動車種別や顧客属性と事故率の関係を分析し、リスク細分型自動車保険の開発した。また、ある銀行ではディシジョンツリーを利用した融資審査モデルを開発した。最近では、一部の生命

* 〒100-0006　東京都千代田区有楽町 1-1-1　　E-mail: ono@nli-research.co.jp

保険会社が、保険の解約防止のためにディシジョンツリー分析を始めた。

表 1 金融業界で既にデータマイニングが適用された分野（開発中も含む）

業種	適用分野	目的	分析手法
銀行 生保	融資審査の推定	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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ
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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ 順位分析
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、顧客ﾆｰｽﾞの発生ﾀｲﾐﾝｸﾞに合せて、商品を提供する。	ｱｿｼｴｰｼｮﾝﾙｰﾙ 順位分析
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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ｸﾗｽﾀﾘﾝｸﾞ
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、車の形状から、若者でも割安になる自動車保険を開発する。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ
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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ
証券 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。	ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ
ｶｰﾄﾞ	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。	ｱｿｼｴｰｼｮﾝﾙｰﾙ ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ
ｶｰﾄﾞ 銀行	消費者ローンの与信審査分析	顧客属性、勤務先、キャシング履歴から、与信審査する。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ
ｶｰﾄﾞ	クレジット・カードの不正請求防止	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。	ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ

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。	ｱｿｼｴｰｼｮﾝﾙｰﾙ 順位分析
生保 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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ｸﾗｽﾀﾘﾝｸﾞ
生保 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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ
生保	介護保険法における介護ﾆｰｽﾞの分析	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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ｸﾗｽﾀﾘﾝｸﾞ
銀行 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。	ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ
銀行 生保	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。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ
ｶｰﾄﾞ	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、販売に役たせる。	ｱｿｼｴｰｼｮﾝﾙｰﾙ 順位分析
全業種	ｺﾐｭﾆｹｰｼｮﾝ･ﾁｬﾈﾙの選択	多様な情報通信メディアの中から、顧客にあったコミュニケーション・チャネルを選択する。	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ ｸﾗｽﾀﾘﾝｸﾞ

2. データマイニングの概論

2.1. 分析手法のタイプ

データマイニングの分析手法をタイプ別にすると、①データ同士の相関関係を発見す

る「相関関係分析（Association)」、②データを意味あるクラスに分類する「クラス分類分析（Classification)」、③データを性質の似た複数のグループに分類する「クラスタリング分析（Clustering)」、④複数の出来事が時系列的に関係をもつ「順位分析（Sequential Patterns)」の４タイプに分けられる。

表 3 一般分析手法と適用例

分析タイプ	分析手法	適用分野
相関関係分析	ｱｿｼｴｰｼｮﾝﾙｰﾙ、ﾊﾞｽｹｯﾄ分析	商品の組合せパターンの発見
クラス分類分析	ﾃﾞｨｼｼﾞｮﾝﾂﾘｰ、ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ	顧客のセグメント化
ｸﾗｽﾀﾘﾝｸﾞ分析	ｸﾗｽﾀｰ分析、ｺﾎｰﾈﾝﾈｯﾄﾜｰｸ	マーケット分類
順位分析	時系列順位分析	時系列商品購入パターン
伝統型統計分析	判別分析、主成分分析、共分散分析	統計モデル
検証モデル	t 検定、F 検定、カイ 2 乗検定、AIC	モデルの検定

2.2. RFM取引属性

金融業界の業務データは、顧客データと取引データから構成されるが、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。従来の分析者は経験に基づいて、取引データから試行錯誤して前述の属性を作るが、本研究では金融商品ごとの RFM 取引属性を機械的に作成することを提案する。初心者でも、「全商品」と「商品毎」と「類似商品毎」の RFM 取引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を開発できる。

もともと RFM 属性は、データベースマーケティングに用いられた。その分析手法では、マーケティングの顧客の優先順位を、RFM 取引属性値に重みを掛けた合計値で求められるが、顧客属性を考慮できない点や重みの決定方法に問題があった。データマイニングではその問題点を解決し、さらに細かい顧客分類が可能になった。

表 4 金融商品の RFM 取引属性

RFM属性名	元の定義	金融商品のRFM属性の例
R(recency)	最後の購買日	最終購入・契約日から分析日までの期間（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

2.3. 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商品の RFM 取引属性から分析し、最終ゴールは業務の専門家が満足するシナリオである。「深い探索」は、経験と技術が必要とされるが、顧客のピンポイント・マーケティングの実現や新しいルールの発見ができる。

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」では、『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・融資審査を「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。業務データの分析では、「分析結果の精度」や「探索するレベル」を業務目的から決定するべきあり、分析技術や結果から決定すべきでない。

2.4. 「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、複数分類モデルによるコミュティー学習が用いられる。ハイブリッドモデルには、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、並列に組み合わせるキャタラクトモデルがある。

(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、最初にディシジョンツリー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とが多い。この手法を用いると、より顧客を絞り込むことができるため、融資審査や与信審査に使われる。

(2)キャタラクトモデル(Cataract model)

キャタラクトモデルは、最初に１つの分析手法でパターンを発見する。次にそのパターンのデータを分析データから外し、他の分析手法やパラメータを変更したモデルで先ほどと同じようにパターンを探索する。最終的には得られたパターンを集めて結論とする。

2.5. KDD プロセスとマイニング・ツール

データベースを用いた知識発見は、複数のプロセスから構成される。これを KDD（Knowledge Discovery in Databases、データベースからの知識発見）プロセスと言う。KDD プロセスは、①選択 (selection) 、②前処理 (preprocessing) 、③変形 (trance formation)、④データマイニング (data mining)、⑤ 解釈・評価 (interpretation / evaluation)、⑥ルール生成の６プロセスから構成される。これらプロセスを何度も繰り返すことによって、精度の高い知識が得られる。

マイニング・ツールは、KDD プロセスをモデルに実装したものである。マイニング・ツールは、手間のかかる②前処理、③変形、④データマイニング、⑤解釈・評価の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。しかし現状のマイニング・ツールでは、意志決定機能が未成熟であり、「深い探索」は分析者の能力に依存する。

2.6. データクリーニング

データウェアハウスを直接探索するとデータベース量が大きすぎるため、データマイニングでは、データウェアハウスから切り出したデータマート（情報検索用の小規模データベース）を探索する。このデータマートは、正事例と負事例の共通項目のデータを結合し、さらに商品ごとの RFM 取引属性（後述）を追加したデータベースである。

金融機関の業務データベースは、負事例と正事例が別々のデータベースになっている

ため、データベース間には、異音同義・同音異義・有意コード・単位相違・欠損値等の問題が存在する。そのため、正事例と負事例のデータベースを結合するときに、データの不整合が起こりやすく、データクリーニングが必要になる。データクリーニングは、作業時間の70%以上を占めることも珍しくない。

2.7. 金融機関のデータの問題点

(1) 負事例のデータ不足問題

データマイニングには、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。例えば融資審査では、正事例が「実際に契約に至ったデータ集合」であり、負事例が「契約できなかったデータ集合」である。金融機関では、契約に至らなかった負事例データをデータベースに蓄積しないことが多い。そして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すべて使えないことがある。

(2) リアルの顧客属性データの不足問題

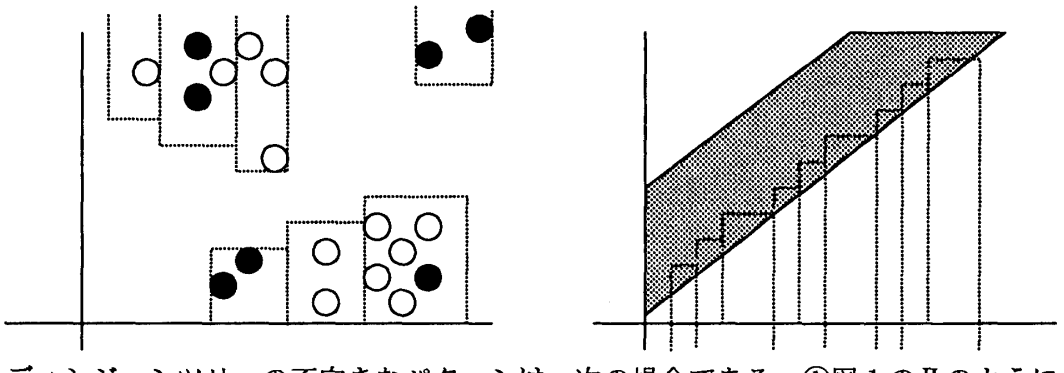
金融機関は、契約時や購入時の顧客属性（年収、勤務会社、病気等）データを多く保有するが、その後更新しないためにリアルの顧客属性データが存在しない。そのため分析で使える顧客属性が「性別」と「年齢」しかないことが多い。

3. ディシジョンツリー分析

3.1. ディシジョンツリーの特徴

ディシジョンツリー分析は、データのもつ属性からツリーの分岐の属性順序や閾値を自動的に算出して、グループ分類をする手法である。ディシジョンツリーが有効なデータ分布は回帰分析と比較するとわかりやすい。回帰分析では直線のみで判別分類するが、ディシジョンツリーは「ギロチンカット」と言う縦軸と横軸の囲みを用いて分類する。

図 1 ディシジョンツリーの有効なパターンと無効なパターン
Ⅰ.ギロチンカットによる分類　　Ⅱ.無数のギロチンカットが発生する場合



ディシジョンツリーの不向きなパターンは、次の場合である。①図１のⅡのように目標属性の境界がちょうど直線ように並ぶと、ディシジョンツリーでは無数のギロチンカットが発生する。②データ中のフラクタルやカオスを発見できない。③属性値の種類が多いほど、利得基準値（後述）が大きくなり優先的に選択されやすい。④ツリーは階層

構造を持つため、上位の説明属性が分割を複雑にさせやすい。⑤不明の属性があったり、新しいデータ内の値の組み合わせが若干異なると、うまく働かないことが多い。

3.2. 利得基準値

ディシジョンツリーでは、属性の利得基準値に基いて、分割属性の優先順位が決まる。利得基準値は、ルールが目的属性値の分布与える影響度合いを数値化したものである。利得基準値が大きくほど影響力が大きく、ツリーの最初の分割属性になる。よく使われるアルゴリズムとその利得基準値を表5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 個番目の値をもつデータがそれぞれ Xi(L)個（ i =1,.., 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 $Pi(L_1)=Xi(L_1)/|L_1|$とすると、各分割基準は下記のように定義できる。GINI 関数は情報エントロピー関数の対数部分をテイラー展開し1次近似すると求まるため、CART と C4.5 は類似のツリーを作りやすい。

情報エントロピー関数による利得基準値

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

GINI 関数による利得基準値

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

カイ2乗関数による利得基準値

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

表 5　ディシジョンツリーのアルゴリズムの種類

アルゴリズム名	元になる利得基準値	対象データ
C4.5、ID3	情報エントロピー関数	質的および量的データ
CART(Classification And Regulation Tree)	GINI 関数	質的および量的データ
CHAID(Chi-squared-AID)	カイ2乗関数	質的データ

3.3. 探索アルゴリズム

探索アルゴリズムは、まず利得基準値が最大となる説明属性を発見し、高い反応属性順の顧客セグメントのツリーを作成する。その属性値を頂点とするサブツリーを同じ手順で作成する。これを繰り返せば、ディシジョンツリーが完成する。

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、分割後の利得基準値を計算する。

③ 最大の利得基準値を有する説明属性により、分割を実行する。

④ 分割後の集合に対して、②、③を繰り返す。

⑤ アルゴリズムの終了は、利得基準値やカテゴリに含まれるサンプル数で判断する。

4. 応用事例：保険解約の防止分析

4.1. 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 6 にまとめる。契約変更属性は契約変更や配当金引出しの情報であり、営業職員属性は契約者を担当している営業職員の属性である。データは、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）を用いた。完成したデータマートは 8 万レコード、属性数は 200 個以上となった。訓練・検証データの配分は 1 ： 1 にして、ランダム・サンプリングを採用した。後にクロスバリデーションにより、モデルの安定性を検証した。

モデルによる分析前に、情報エントロピー値を用いて不要な属性を調査し、説明属性の絞り込みをする。本分析では、この作業を通じて 200 属性から 35 属性まで絞り込めた。

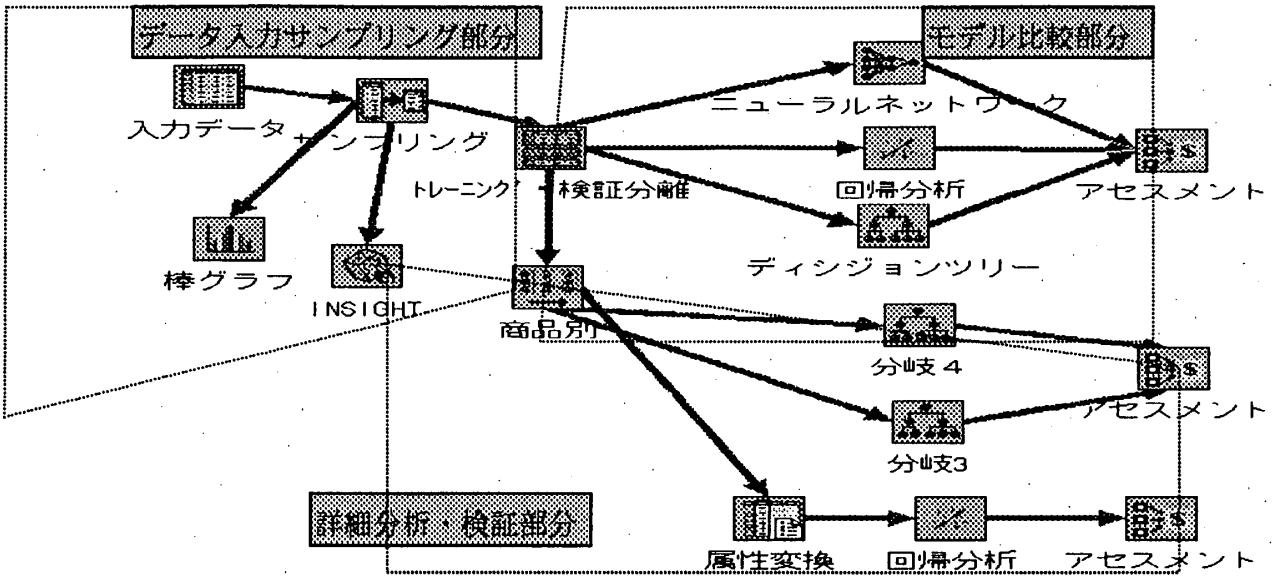
表 6　保険解約防止の属性

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

4.2. プロセス・フロー・ダイアグラム

SAS／Enterprise Miner によるデータ属性や手法・パラメータの選択やデータフローは、図 2 の「プロセス・フロー・ダイアグラム」で制御する。四角箱がノードと言い、サブルーチンの役目を果たす。分析作業は、その役割により 3 分割できる（図 2 の点線）。左上部は、データ入力・サンプリングの部分である。右上部は分析手法を比較調査する部分である。下部は商品別に詳細な分析を行い、当時に結果を検証する部分である。

図 2 解約防止分析のプロセス・フロー・ダイアグラム図



左端の［入力データ・ノード］では目標属性の指定し、［サンプリング・ノード］では入力データを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］のノードは、分布を調べ異常値や特殊パターンの探索に使用する。［アセスメント・ノード］では、複数モデルの分析結果をグラフ表示で比較し、この分析に適したモデルを決定する。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。ディシジョンツリーのパラメータ（分岐数と深さ等）を変化させて、導出ルールの変化を調べる。

4.3. 分析結果

4.3.1. 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、ディシジョンツリーの判別効率を正反応補足曲線で比較する（図 3）。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 4 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。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必要がある。図 3 からディシジョンツリー分析が最適であり、図 4 から商品 2 の方が商品 1 より予想しやすいことがわかる。

図 3 分析手法の比較

曲線の種類 （上から）
理想曲線、ディシジョンツリー、ランダム曲線
と回帰とニューロは、ほぼ同じ曲線上ある

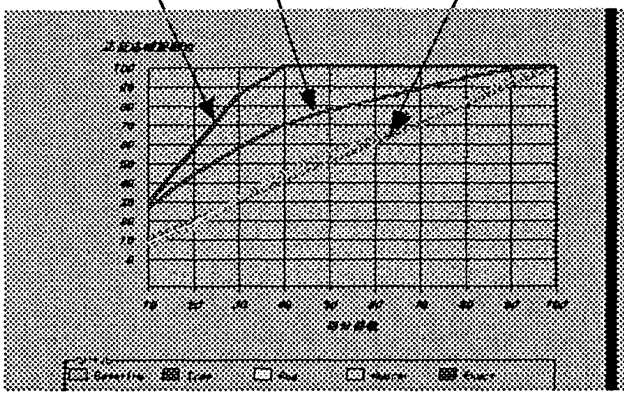
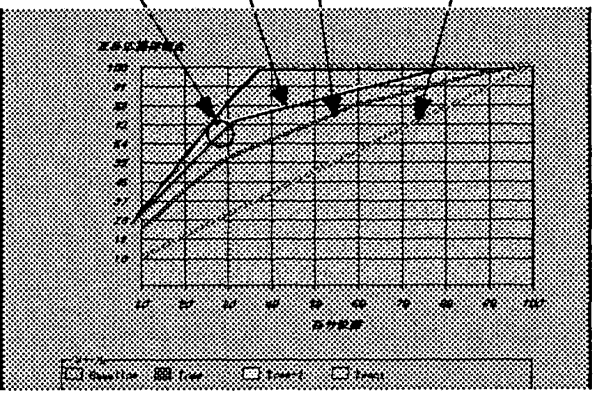


図 4 商品別のディシジョンツリー

曲線の種類（上から）
理想曲線、商品 2、商品 1、ランダム曲線



4.3.2. 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。ディシジョンツリーのパラメータには、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 個以下が人間にとって理解しやすい。図 5 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。結果は、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、導出ルールも大きく変わる。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。

図 5 パラメータによる影響

曲線の種類 （上から）
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

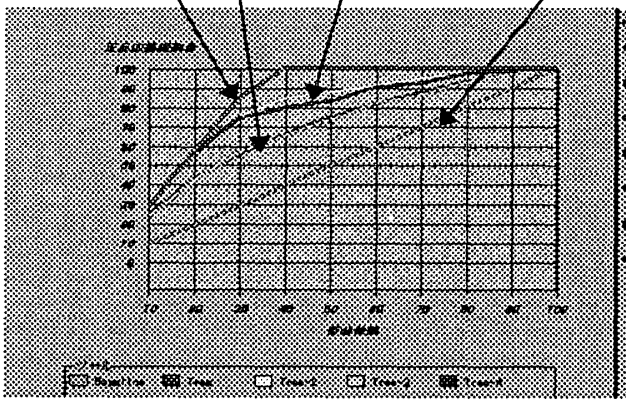
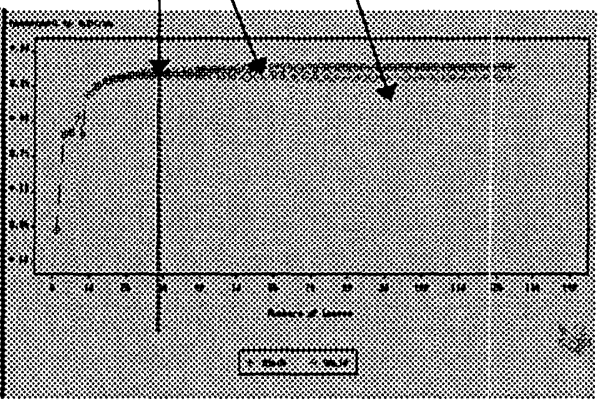


図 6 収束の葉数

曲線の種類（上から）
検証データ、訓練データ
収束線



4.3.3. 収束

ディシジョンツリーは、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、データの変動に対しても影響を受けづらくなる。

図 6 では、訓練・検証データに適用したモデルが共に、28 枚で収束したことを示している。この葉数ならば、ツリーの階層は 3 層から 4 層となり、人間にとって理解しやすい。訓練・検証データは、共に同じように収束しており、安定的なモデルと言える。

4.3.4. 検証結果

モデルの推測結果と実際の保有解約を比較する。商品 2 の解約モデルの正解率は、86％（＝62％＋24％）となり、業務的に満足できる結果が得られた。ただ今回の分析データ数は、全契約 0.3％にすぎない。

表 7　商品 2 の正解率（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)

4.3.5. ルール導出

ツリーの構造から IF－THEN－ELSE 形式のルールを導出できる。導出したルールは、ツリーの枝を狩りながら汎用化を試みる。汎用化したルールは、少ない説明属性で目標属性を説明でき、しかも予測精度が高い。保険解約分析では、顧客のプロフィールを導くことができる。この作業をプログラムで機械的に行うこともできるが、不要なルールも導出されことも多いため、今回は専門家と共同で分析した。

保険解約の顧客プロフィール：
経過年数が 1 年未満であり、
　　契約時の担当職員が退社するような契約の解約失効率は高い。
経過年数が 1 年から 2.5 年間で、
　　死亡保険額が低いほど解約失効率は高い。
経過年数が 2.5 年から 4.5 年間で、
　　振替貸付があれば解約失効率は高い。
経過年数が 4.5 年から 7.5 年間で、
　　保障倍率が 15.5 未満ならば解約失効率は高い。……（以下省略）

4.4. 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対して有効な対策は、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考えられる。どのアプローチが、どのような顧客に有効でかつ、コスト的に優れているかを計測するため、地区を限定して実験調査をすることが行われる。効果は「アプローチを試みたグループ」と「何も顧客へアプローチをしないグルー

プ」の差異分析から測定する。

5. まとめ

金融業界においてデータマイニングは、「One to One マーケティング」と「クレジットリスク」などが主な応用事例としてあげられる。データマイニングの成功の鍵は、「負事例」と「リアルの顧客属性」のデータ収集と「業務への反映方法」である。

最近のデータマイニングは、分析手法だけでなく KDD プロセスを指すことが多い。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」の分析データベースには「商品の RFM 取引属性」が欠かせない。分析結果は、業務の専門家にわかりやすく、安定であることが重要である。本報告では、応用事例として実務データに基いた「保険解約の防止分析」を取上げ、デシジョンツリーによる分析が有効であることを示した。分析結果を業務へ反映する場合は、一部の顧客へ実験的に対策を実施し、顧客ごとの有効なアプローチを特定することが望ましい。

日本のデータマイニング技術は、欧米に遜色がなくなりつつあるが、研究対象がアルゴリズムに偏る傾向が強い。これからは、実務分析で問題となるデータクリーニングや KDD プロセスの研究に期待したい。

6. 参考文献

- Fayyad, U.M., Piaetsky-Shapiro, G. and Smyth, P.: From Data Mining to Knowlede Discovery: An Overview, Advances in Knowlege Discovery and Data Mining, pp.1-34, AAAI/MIT Press, 1996.
- Joseph P.Bigus, "Data Mining with Neural Networks" , The McGraw-Hill Companies, 1996.
- 河野浩之, "データベースからの知識発見の現状と動向"、人工知能学会, vol.12 No.4 ,pp.497-504,1997.
- 寺野隆雄、"KDD ツールの動向と課題"、人工知能学会, vol.12 No.4 ,pp.521-521,1997.
- 安武史、"データウェアハウスとデータマイニング"、Financial Research, pp.15-31, NEC 総研、1998.
- J.R.キンラン、"AI によるデータ解析"、トッパン、1995.
- 丹後俊郎、山岡和枝、高木晴良、"ロジスティク回帰分析"、朝倉書店、1996.