5U - 01

セグメント抽出とMBRを併用した予測システムの提案と評価

鮎川 江里香　　森田 豊久

(株) 日立製作所 システム開発研究所

1. はじめに

多くの分野・業種において様々な情報技術の活用が進むにつれ，そこで蓄積されるデータの規模も著しく増大し，大量データから有用な知識を発掘するデータマイニング技術にも注目が集まるようになってきた。

例えば，近年急速な勢いで成長を遂げつつある移動体通信業界では，加入者数の増大に伴って，それぞれの事業者内に蓄積される情報量も膨大なものになりつつある。また，同業界では，顧客獲得のための過剰なサービス競争が繰り広げられた結果，顧客の定着率が低下し，事業者によっては解約率の高さが収益に深刻な影響を及ぼすようになってきた。そこで，多くの事業者が「新規顧客の獲得」よりも「既存顧客の維持と優良化」に注力するようになり，その際の有力な支援手段として，大量の顧客データを活用した顧客の特性分析や行動予測に大きな期待が寄せられるようになった。

本稿では，ルール生成 [1] によるセグメント抽出とMBR（Memory-Based Reasoning：記憶に基づく推論）（例えば [2]）の実行時のパラメータ調節を特徴とする予測システムを提案し，移動体通信業界における解約予測への適用事例を通じて上記システムを評価した。

2. 概要

解約予測に用いるデータは，顧客の属性情報や通話データなどから顧客特性を抽出した顧客プロファイル・データである。顧客プロファイル・データは1顧客当たり数千項目から構成され，分析目的に応じて各項目を選択的に使用する。上記データには，顧客の解約実績（解約か否か）を表す「契約」項目も含まれるものとする。

解約予測では，過去に得られた既知データ（解約実績が既知）を用いて未知データ（解約実績が未知）を予測する。この時，セグメント抽出により事前に有効データを絞り込んでおき，また，テスト予測によりMBRパラメータを調節しておくことで，その後の実予測の精度向上を図る。

3. ルール生成を用いたセグメント抽出

セグメント抽出とは，MBRで使用する既知データ（未知データ）について，あらかじめ「解約率の高い（高そうな）セグメント」を抽出しておく処理である。この処理により，解約事例の濃度が高まり，予測精度の向上が期待できる。

解約率の高い（高そうな）セグメントの抽出方法は以下の通り。すなわち，既知データを用いて解約者の特性を表すルールを生成し，上記ルールを満たすレコード集合を解約率の高いセグメントと見なす。また，未知データについては，上記ルールを満たすレコード集合を解約率の高そうなセグメントと見なす。ここで，ルール生成とは，大量データ中に潜む規則性や因果関係を「もし～ならば…」というIf - Thenルール形式で抽出する手法である。ルール生成の概念図を図1に示す。

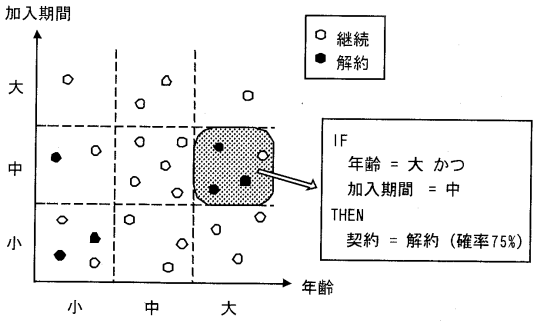

図1 ルール生成の概念図

4. MBRを用いた解約予測

解約予測の手法にはMBRを用いる。MBRとは，大量の既知データの中から未知データの類似事例を見つけ出

Predictive Modeling with Segmentation and MBR
Erika AYUKAWA, Toyohisa MORITA
Systems Development Laboratory, Hitachi, Ltd.

し，それらを元に未知データの値を推論する手法である。MBR の概念図を図 2 に示す。

例えば，図 2 の◎印で示された未知事例について「契約」項目（これを推論項目と呼ぶ）の値を推論する場合，まず，既知事例（○：契約 = 継続，●：契約 = 解約）及び未知事例の「年齢」項目と「加入期間」項目（これらを説明項目と呼ぶ）の値を参照し，未知事例から各々の既知事例までの距離を計算する。ここで，図中の点線で囲まれた部分を未知事例に関する近傍範囲とすると，その範囲内の既知事例がすべて類似事例となる。類似事例が確定したら，それらの「契約」項目の値を元に，未知事例の「契約」項目の値（予測値）を推論する。

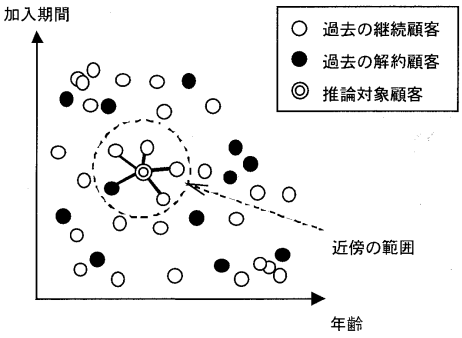

図 2　MBR の概念図

本稿で提案する予測システムでは，まずテスト予測により MBR パラメータを調節し，続く実予測にて上記パラメータを用いた MBR を実行する。

(1) テスト予測

まず，後述の実予測に先立って，実予測に最適な説明項目を決定するためのテスト予測を行う。テスト予測とは，既知データを二分し，一方をテスト予測用の既知データ，他方をテスト予測用の未知データと見なし，両者を用いて説明項目の組み合わせを網羅的に変更しながら MBR を実行する処理を指す。（ただし，説明項目の組み合わせ個数については適宜，上限値を設定する。例えば，「100 個の候補項目について最大 3 個までの組み合わせ（($_{100}C_1$ + $_{100}C_2$ + $_{100}C_3$) 通り）について MBR を実行する」など。）ここで得られたすべての推論結果について予測精度を評価し，その中で最も予測精度の高かった説明項目を実予測用の説明項目として選択する。

(2) 実予測

テスト予測により最適と見なされた説明項目を用いて実予測を行う。実予測とは，本来の既知データと未知データを使用した実際の予測を指す。

5. 評価

セグメント抽出により得られた 7 つのセグメントについて，テスト予測で選択された説明項目を用いて実予測を行った結果を表 1 に示す。表 1 の解約率とは予測後に判明したセグメント内の解約者の割合であり，正解率とはセグメント内で MBR による解約見込みスコアが上位 X%（= 全体の解約率と同数）までの顧客に含まれていた解約者の割合である。表 1 によると，全体の解約率に比べ，セグメント抽出後の解約率は 2.94～4.98 倍に伸びており，さらに MBR を用いた絞り込みによって，各セグメントの正解率は全体の解約率の 4.31～6.52 倍にまで高まっていた。以上から，予測により解約者を約 4.3 ～ 6.5 倍の濃さで効果的に絞り込めたことが明らかとなった。

表 1　予測結果

セグメント	解約率の伸び(倍)	正解率の伸び(倍)
全体	1.00	—
第1セグメント	2.94	4.31
第2セグメント	3.30	4.63
第3セグメント	3.42	5.75
第4セグメント	3.81	5.45
第5セグメント	3.75	4.92
第6セグメント	4.05	5.53
第7セグメント	4.98	6.52

6. おわりに

本稿では，セグメント抽出と MBR を併用した予測システムを提案し，移動体通信業界における解約予測への適用を通じてその評価を行った。その結果，上記システムにより，運用上，十分に有効な予測精度を得られることが確認できた。

参考文献

[1] 芦田仁史ほか：データマイニングにおける特徴的ルール生成方式，情報処理学会第 50 回全国大会，3－19（1995）

[2] Michael J. A. Berry and Gordon Linoff : Data Mining Techniques For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc. (1997)