

6 A A - 3

データマイニングにおけるルール関連構造の抽出と DB マーケティングへの適用

鮎川 江里香 牧 秀行 森田 豊久 前田 章

(株)日立製作所 システム開発研究所

1. はじめに

計算機の高速度化、ディスクの大容量化などにより、多くの分野・業種において大規模なデータの蓄積が可能となりつつある。これらのデータを有効に活用するための技術として、大量データ中に埋もれた知識を発掘するデータマイニングが注目を集めている。

本稿では、特徴的ルール生成方式 [1] を用いて生成されたルールに対し、それらの関連構造を抽出する方法を提案する。また、流通業における DB マーケティングへの、上記抽出方法の適用結果について報告する。

2. 特徴的ルール生成方式

特徴的ルール生成方式では、大量データ中の規則性・因果関係を表現する手段として、「もし～ならば…である」というルール形式 (If-Then ルール) を用いる。各々のルールは、データ変数の属性値をあらかじめいくつかのカテゴリに変換 (カテゴリ化) したカテゴリ値によって表現される。生成されるルールの形式は以下の通り。

もし X_1 が x_1 かつ X_2 が x_2 かつ … X_n が x_n
ならば Y が y である

ここで、 X_i は入力 (説明) 変数名称、 x_i は変数 X_i のカテゴリ名称である。同様に Y は出力 (被説明) 変数名称、 y は変数 Y のカテゴリ名称である。「 X_i が x_i 」という組を条件節と呼ぶ。また、「もし A ならば B である」の A を条件部、 B を結論部と呼ぶ。

ルールには評価尺度と呼ぶ実数値を割り当て、評価尺度の大きいルールほどルールの価値が高いものとみな

す。特徴的ルール生成では、条件節の組み合わせの中から、評価尺度の大きいルールを選び出す処理を実行する。

3. ルール関連構造の抽出方法

特徴的ルール生成方式では、ある特定の結論部をできるだけうまく説明する条件節の組を探索する。そこで、まず上記の結論部にデータ中のあらゆる変数とそのカテゴリ値の組を指定してルール生成を繰り返す。その後、得られたルールの条件部と結論部に含まれる変数、及びそのカテゴリ値同士の相互関係を調べる。以下に、ルール関連構造抽出方法の詳細を示す。

(1) 属性値のカテゴリ化

各変数の属性値をあらかじめカテゴリ化しておく。カテゴリ化の例としては、変数の時間変動を、その値により「減少」、「一定」、「増加」に分類することなどが考えられる。

(2) ルール生成

ルールの結論部の出力変数にすべての変数を順に指定し、上記変数のカテゴリ値についても取りうるすべての値を順に指定して、ルール生成を繰り返す。

(3) ルールのグループ化

生成したルールの中から評価尺度がある閾値を上回るルールのみを選択する。上記ルールに含まれる変数とカテゴリ値の組を有向グラフにより連結してグループ化する。

(4) グループ間の関係抽出

(3) で求めた各グループについて、グループ内のそれぞれのルールを満たすレコードを調べ、グループ内の和集合、及びグループ間の共通集合を求める。

(5) カテゴリ化の調整

変数のカテゴリ化方法を変更することで、事象をより適切に表現するルールの獲得が期待できる。そこで、必

要に応じてカテゴリ化方法を変更し、再び(2)~(4)の処理を繰り返す。

4. DBマーケティングへのデータマイニングの適用

DBマーケティングへのルール関連構造抽出方法の適用実験として、百貨店販売データを用いた購買動向の抽出を行った。購買動向の中でも、特に商品間の購買変動の関係について詳しく調べた。

4.1 データ

使用したデータは、百貨店におけるカード会員顧客の3年間の購買履歴トランザクションデータであり、商品1~商品22の買上金額についてまとめている。上記データを用いて、商品間の買上金額の変化量の関係を調べた。

4.2 分析結果

期間中に買上金額が増加した商品間の関係を調べるため、ルール「もしXが増加ならばYが増加である」のみを対象としてルールのグループ化を行った。その結果、図4.1に示すような3種類のグループが成立した。以上より、買上金額が増加した商品間には、局所的に密接な関係が成り立っていることがわかった。

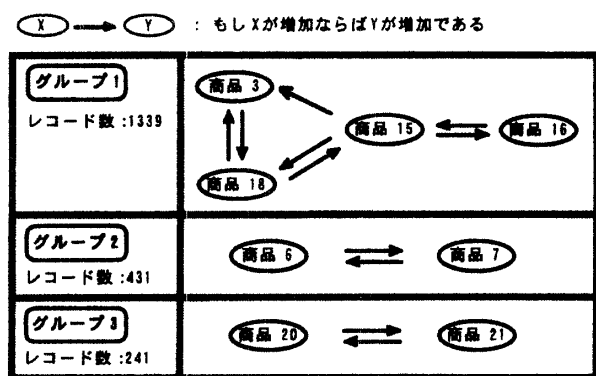


図4.1 ルール関連図

次に、グループ間の関係を抽出したところ、図4.2に示すような構成となっており、グループ間の顧客の重なりが比較的小さいことが明らかとなった。よって、上記の結果を用いた顧客のセグメンテーションも有効と考えられる。

続いて、カテゴリ化方法の調整を行った。成立するグ

ループ数と各グループに含まれる商品の種類は図4.1と同等であったが、グループ間の関係は図4.3のように変化した。すなわち、グループ内の各ルールを満たす顧客数はやや減少するものの、グループ間の顧客の重なりがより小さくなっていった。この結果より、カテゴリ化方法の調整による顧客のセグメンテーションの高精度化が可能と考えられる。

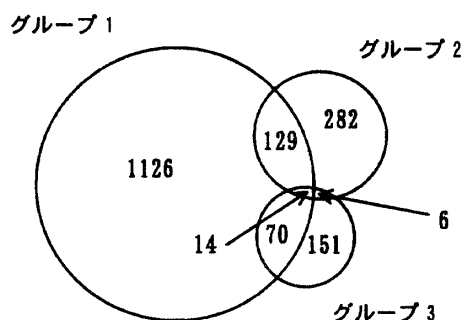


図4.2 グループ構成図

(図中の数値は、各領域のルールを満たす顧客数)

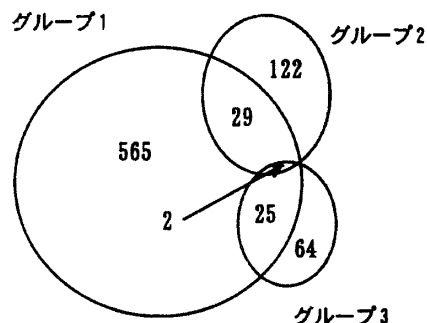


図4.3 カテゴリ化方法調整後のグループ構成図

5. おわりに

本稿では、特徴的ルール生成方式により生成されたルール間の関連構造を抽出する方法を提案し、DBマーケティングへの適用実験を行った。その結果、上記抽出方法が顧客のセグメンテーションに対し有効となることを確認した。

参考文献

[1] 芦田ほか：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)