# データマイニングにおける特徴的ルール生成方式

7P－1　芦田仁史*　前田章*　高橋ヨリ**
*（株）日立製作所システム開発研究所　**（株）日立製作所ソフトウェア開発本部

## 1 はじめに

計算機の高速化，ディスクの大容量化に伴い，蓄積されるデータ量は増大しつつある．これらの大量データを有効に活用する技術として，データ中に埋もれた知識を発掘するデータマイニング技術が注目されている．我々は，大量データ中に埋もれた規則性・因果関係をルールの形で抽出する特徴的ルール生成技術を開発した．本稿では，特徴的ルールを用いた推論方式を提案し，金融指標データによる検証実験の結果を報告する．

## 2　特徴的ルール生成方式

大量データ中の規則性・因果関係を表現する手段として，我々は「もし～ならば・・・」というルール形式を用いた．この表現は，人間にとって理解が容易であると同時に，計算機処理にも適している．生成したルールは次の形式をとる．

もし　X1 がA1 かつ X2 が A2 かつ・・
　Xn が An ならば
　　YがB1である確率はα1
　　YがB2である確率はα2
　　・
　　YがBmである確率はαm

ここで，Xiはそれぞれ入力変数名称，Aiは変数Xiのカテゴリの名称である．同様にYは出力変数名称，Bjは変数Yのカテゴリ（出力カテゴリ）名称である．ここで，ルールの結論部は出力カテゴリの分布である．

開発したルール生成方式の要点は，考えられる条件節の組み合わせを生成し，ある評価尺度に従って，その条件部の優劣を評価，選択することである．

Aを条件部，Bi(i=1,...,m)を出力カテゴリとし，

$$\mu(A) = P(A)^{\beta}\sum_{i=1}^{m} P(Bi|A)\log\frac{P(Bi|A)}{P(Bi)}$$

を評価尺度とする．ここでP(A)は事例がAという条件を満たす確率，P(Bi)は出力カテゴリがBiである確率，P(Bi|A)はAという条件を満たすという前提で，出力カテゴリがBiである確率を表す．βは0と1の間の実数値で，一般性パラメータと呼ぶ．

上式の第１因子は，カバー率のβ乗となっている．カバー率とは，ルールの条件節を満たす事例の割合である．βを大きくする程，カバー率（一般性）を重視した評価尺度となる．第２因子は，ルールの精度を評価し，Kullback情報量[1]に対応している．

条件節の組み合わせによりルールを生成する．生成ルール数と最大の条件節数を設定し，最大条件節数に至る全ての条件節の組み合わせを生成し，その中で評価尺度の高いものから生成ルール数分のルールを選択する．

## 3　特徴的ルールを用いた推論方式

### 3．1　特徴的ルールによる推論方式の概略

図１に提案方式の概略図を示す．図１において，特徴的ルールとは，前節で述べた「もし～ならば...」の形式のルールの集まりである．カテゴリ化情報とは，ルール生成時に作成されるもので，各カテゴリの範疇を示す．特徴的ルールは全て，カテゴリにより表記されるので，実数値に対する推論をおこなうには，この情報が必要である．

推論エンジンは，特徴的ルールを推論に利用するための前処理と，推論アルゴリズムにより構成される．

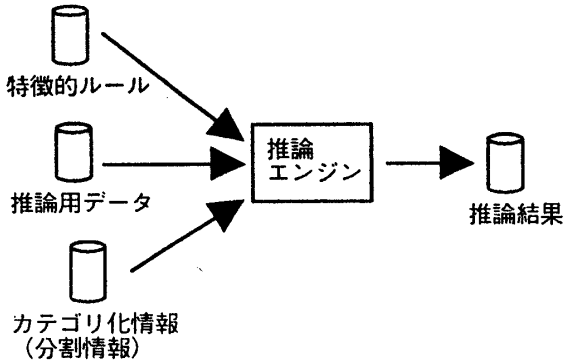


図１　特徴的ルールを利用した推論方式の概略図

A Characteristic Rule Extraction Algorithm for Data Mining
Hitoshi ASHIDA*, Akira MAEDA*, Yori TAKAHASHI**
*Systems Development Laboratory, Hitachi Ltd.
**Software Development Center, Hitachi Ltd.

### 3．2　特徴的ルールによる推論方式

特徴的ルールを推論に利用するために以下の２つの前処理を施す．

i）各出力カテゴリの代表値の決定

ルール結論部の属性値はカテゴリで表記されているので，数値計算に利用するため，各カテゴリに数値を割り当てる．ルール生成時に教師データから算出したカテゴリ化情報を利用する．

具体的には，カテゴリ化情報から与えられる各カテゴリの両端点（最大値と最小値）の平均値を代表値とする．

ii）各ルールの出力値の決定

各ルールにおいて，i）で求めた各出力カテゴリの代表値をそれぞれの生起確率で重み付け平均したものを，出力値とする．

各出力カテゴリＢi(i=1,...,m)の代表値をＲi(i=1,...,m)とすると，そのルールの出力値は，

$$\frac{\sum_{i=1}^{m} R_i \cdot \alpha_i}{\sum_{i=1}^{m} \alpha_i}$$

となる．

以上の前処理を施した後，以下の２stepにより，各入力データに対して推論をおこなう．

**step1: 適合ルールの選択**

**推論用データの条件部が適合するルールを選択する．ルール条件部は全て，カテゴリで表記しているが，入力データは数値データも含んでいる．各ルールは数値データに対しては，前節でも触れた分割情報を参照し，入力変数の値が各カテゴリの範疇に含まれるかどうか判定する．**

**step2: step1で選択したルールの出力値の単純平均を算出し，推論結果とする．**

## 4　金融指標データを用いた検証実験

金融指標データを用いて，提案したルール生成方式と推論方式の性能を評価した．実験には1988年2月から約１年間のデータを２等分し，前半の　155データをモデル生成用，後半の157データで検証用とした．終値，乖離率3日，乖離率10日，移動平均値変化3日，前日との終値変化額，RSI(Relative Strength Index)5日の６項目を入力変数，終値変化額を出力変数とした．例えば，終値のカテゴリと範疇は，

小(12670未満)，中（12670以上13130未満），

大（13130以上）

とした．生成したルールは例えば，

IF　乖離率3日 = 大．
　　移動平均値変化3日 = やや大
　　ＲＳＩ５日 = 大
THEN　終値変化額 = やや小 <45%> [5/11]
　　終値変化額 = 中 <27%> [3/11]
　　終値変化額 = 大 <27%> [3/11]

のような形式である．このようなルールを２章で述べた評価尺度に基づき３０個選択し，ルールモデルとした．ルール生成時のパラメータは，一般性レベル：1，生成ルール数：30，最大条件節数： 3とした．比較のためルールモデルと同じ入出力変数を用いてニューロモデルも生成した．ニューロモデルの学習パラメータは，学習係数：0.7，モーメント係数：0.1，学習回数：30000，中間ノード数：5とした．

モデル生成に利用したデータと検証用データに対して推論を実行した．結果は表１の通りである．

表１の各値はデータ数を表す．例えば，４カラム目の４行目の30という値は，検証用データにおいて，ルールモデルにより指標が下がると判断した（変化額<0）データの中で，実際に指標が下がった事例が30件あったことを示す．

**表１：ルールモデルとニューロモデルの検証結果**

| | 特徴的ルールモデル | | | | ニューロモデル | | | |
|---|---|---|---|---|---|---|---|---|
| | モデル生成用データ | | 検証用データ | | モデル生成用データ | | 検証用データ | |
| 実測／予測 | 下 | 上 | 下 | 上 | 下 | 上 | 下 | 上 |
| 下 | 21 | 5 | 30 | 26 | 43 | 13 | 41 | 43 |
| 上 | 16 | 20 | 11 | 13 | 35 | 64 | 32 | 41 |

## 5　おわりに

特徴的ルールの生成方式と，特徴的ルールを用いた推論方式を提案した．特徴的ルールモデルは，与えられたデータ中の，特徴的な部分空間のみ抽出するため，全ての空間をカバーするわけではない．実験データにおいて，モデル生成用データに対して４割，検証用データに対して約５割のデータのみ説明している．しかし，モデルの精度はニューロモデルと同等である．またルールモデルは，可読性が高いため，ユーザが問題の特性を理解，確認するのに役立つ．ユーザが専門的な知識を持っていれば，容易に追加し，モデルを改良することも可能である．

## 参考文献

[1] 稲垣：数理統計学（裳華房　1990.11）