5M－5

属性間相関ルール抽出の試み

三石 彰純[†]
[†]三菱電機（株）情報技術総合研究所

1 はじめに

大量のデータから相関ルールを抽出するデータマイニング技術[1]を用いると大量の相関ルールが得られることがあるが、その中から真に有用なルールを見出すのは困難な作業である。この問題に対して、我々は無意味なルールを自動的に排除する方法として、数理的フィルタによる方法を試みた[2]。この方法によって、注目すべき相関ルールを絞り込むことができたが、それでもなお多くの相関ルールを人手をかけて検討せざるを得ない状況である。

相関ルールが大量に得られる理由の１つとして、従来の相関ルールは属性の特定の値（属性値）に着目していることがあげられる。このため、同一属性に関するルールが複数得られることが多い。属性に着目した相関ルール（以下、属性間相関ルール）を用いれば、相関ルールの数を減少させることができ、人手による検討作業を軽減することが期待できる。更に、複数のルールの連鎖によって生じる派生的な相関ルール（以下、派生ルール）を自動的に削除できる可能性がある。

本報告では、属性間相関ルールと派生ルールの定義、抽出方法、派生ルールの検出方法について提案し、簡単な試行実験を行った結果について報告する。

2 属性間相関ルールの抽出

2.1 目的 多値属性を持つデータに対して、従来の相関ルール抽出を行うと、例えば「身長：高い→体重：重い」のように、属性値に着目した相関ルールが得られる。この場合、同時に「身長：低い→体重：軽い」のように、同一属性に関するルールが複数得られることが多い。そこで、「身長→体重」のように属性のみに着目した属性間相関ルールとその指標を定義して、自動的に抽出することを考える。

2.2 従来の指標 通常の（属性値間）相関ルールでは、支持度と確信度がルールの指標として用いられることが多い。我々は、主として医療データや品質管理データから有用な相関ルールを抽出することを目的としているため、確信度よりも相関係数に興味がある。

２つの数値属性間の相関係数は、一般的に(1)式で与えられる。

$$r=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \cdots (1)$$

ここで、nは標本数、x_iはi番目の標本の属性xの値、x_iはi番目の標本の属性yの値、$\bar{x},\bar{y}$は属性x,yそれぞれの平均値を示す。

この場合、属性値は比例尺度や間隔尺度で表せる数値であることが必要であるが、我々が対象とするデータでは順序尺度や名義尺度（例：問診結果、性別、職場など）も扱わなければならない。また、相関ルールの各辺は属性集合も許すため、各属性は比例尺度であっても「身長と年齢の組合せ」などのような属性集合は比例尺度ではなくなる。

ところが、属性値が「有無」や「大小」のように２値である場合に限っては、相関係数の絶対値は属性値に依存しないため、順序尺度や名義尺度であっても相関係数を算出できる。

従来、我々はこの点に着目して、属性値間の相関係数を算出し、相関ルール抽出の指標としてきた[3]。以下では相関係数を指標としたルールを相関ルール(Correlation Rule)と呼ぶ。また、A→BとB→Aの相関係数は同一であるので、A⇔Bと書くことにする。

2.3 属性間相関係数 属性間相関ルールを考える場合、(1)式の意味で相関係数を計算することはできない。そこで、林の数量化第Ⅲ類[4]を援用し、図１の手順で相関係数を計算する。図中(b)(c)の処理は固有値問題に帰着できるため、実際には固有値を求めることで相関係数を計算する。

この手順で求めた相関係数が閾値以上となる場合、

An attempt to discover the correlation rules between attributes.
Akitoshi MITSUISHI[†]
[†]Mitsubishi Electric Corporation

(a) 属性(属性集合) x, y に関する分割表を作成する
(b) x, y の属性値を適当に与えて、(1)式で相関係数を計算する
(c) (b)求めた相関係数が最大になるように、x, y の属性値を変化させる
(d) 求めた最大の相関係数を属性間相関係数とする

図1 属性間相関係数算出手順

$x \Leftrightarrow y$ を属性間相関ルールと定義する。

2.4 抽出方法 理想的には、ユーザが与えた最小相関係数以上の相関係数を持つ属性間相関ルールを求めたい。ところが、この場合apriori[1]のような効率的アルゴリズムがなく、また全件探索も現実的ではない。我々は、属性間相関ルールが存在すれば、その属性に関する属性値間相関ルールも存在することが多い、という仮定のもとに、図2の方法でルールを抽出することを試みる。

① データベースから属性値間相関ルールを抽出する
② 属性値間相関ルールから値に関する情報を除去して属性間相関ルールの候補とする
③ 属性間相関ルール候補の相関係数を計算し、ユーザが与えた最小相関係数を満たすものをルールとして出力する。

図2 属性間相関ルール抽出手順

3 派生ルールの抽出

得られた相関ルールを調べると、A⇔B，B⇔Cに加えてC⇔Aなるルールが存在し、ルールが循環関係になる場合がある。この時、C⇔AはA⇔BとB⇔Cの連鎖によって生じる派生ルールであることが多い。AとCの間に直接の相関関係が存在する場合、もとの母集団を複数の部分母集団に分割しても、AとCの間には相関関係が認められるはずである。このことを利用して、下記の方法で派生ルールの抽出を試みる。

④-1 属性間相関ルールの集合から循環関係にあるルールの集合を抽出する
④-2 循環関係にあるルールの集合から、隣接する属性A，Bを取り出し、残った属性の全属性値の直積データベースを分割する
④-3 分割した全ての部分データベースについてA⇔Bを検証し、全部分データベースについて相関関係が認められなかった場合には、A⇔Bを派生ルールとする

図3 派生ルール抽出手順

4 試行実験結果

属性間相関ルール抽出、および派生ルール抽出の実現性を確認するため、5770 名分の 338 属性からなる健康診断データに対して、上記手順でルール抽出の試行実験を行った(Celeron500MHz 使用)。
各処理で得られたルール数と処理時間を表1に示す。

表1 試行実験で得られたルール数

Step	処理内容	ルール数	処理時間
①	属性値間相関ルールの抽出	1886	6.0秒
②	属性間相関ルール候補の作成	716	17.5秒
③	属性間相関ルールの抽出	96	
④	派生ルールの抽出	5	4.9秒

処理①では、最小支持度 0.2%、最小相関係数 0.3 で抽出を行った。処理②によって、ルール数が半減しており、同一属性に関するルールが平均 1886/716=2.6 個存在したことがわかる。③では属性間相関係数が 0.5 以上のものをルールとした結果、ルール数は格段に減少している。この結果から、属性間相関ルールを用いることによって、相関ルールの数を大幅に減少させることができ、有用なルールを発見するための作業が容易になることが期待できる。ただし、③で用いる閾値を大きくすると有用なルールまで削除してしまう危険があり、処理①で与える属性値間相関係数との兼ね合いが難しい。一方、④において派生ルールは少数しか抽出できないが、派生ルールは有用に見え易いため、これを除去することには意味があると考えている。

5 まとめ

属性間相関ルールと派生ルールを定義し、その抽出方法を示した。また、この方法に基づいて抽出実験を行い、実現性と有用性の見通しを得た。今後、実験例を増やすと共に、属性間相関係数およびその閾値の意味についての検討、派生ルール抽出方法の理論的裏付けを行うことが課題である。

参考文献

[1] Agrawal, R., Srikant, R. : "Fast Algorithm for Mining Association Rules", Proc. VLDB '94,1994.
[2] 川上他：相関係数による数理的フィルタリング機能検証，データ工学ワークショップ(DEWS99), 1999.
[3] 三石他： Knodias におけるデータマイニング方式, 第56回情処全国大会 2W-6, 1998.
[4] 田中豊他：多変量統計解析法、ISBN4-7687-0154-X、現代数学社.