Knodiasにおけるデータマイニング方式

2W－6

三石 彰純，山崎 高日子，小幡 康，田中 秀俊，白石 將†
†三菱電機（株）情報技術総合研究所

1 はじめに

我々が現在開発を進めているデータマイニングシステム Knodias では、バスケット分析を行う複数のマイニングエンジンを実装している。これらのエンジンでは、マイニングの指標として最小χ^2値を用い、χ^2適合度検定を行うことでルール抽出を行っている。この指標を用いることによって、抽出する相関ルールの質を高めるとともに、負の相関ルールの抽出を可能にした。

本稿では、χ^2適合度検定を用いたルール抽出の方法とその意味、および負の相関ルール抽出について述べる。

2 マイニング方式

Knodias のマイニングエンジンには下記の特長がある。

(1) χ^2値をマイニングの指標としている。
(2) 負の相関ルールを抽出できる。
(3) 相関ルールの結論部／条件部に出現できるアイテムを指定できる[1]。
(4) 排反事象を考慮に入れたマイニングを行う[1]。
(5) 限られたメモリ容量で大規模なマイニングを行うことができる[2]。
(6) 相関ルールの結論部は１アイテムに限定している(処理の簡素化)。

ここでは、(1)(2)について詳しく述べる。

2．1 マイニングの指標

マイニングの指標（制約）として、従来から最小支持度(minimum support)と最小確信度(minimum confidence)の組み合わせが広く知られている[3]。しかし、この指標を用いた場合には統計的に見て意味のない相関ルールも同時に抽出されてしまうことも知られている[4]。すなわち、A→Bという相関ルールにおいて、A、B、A∩Bの 支持度(出現度数)をそれぞれ S_A、S_B、S_{AB}、全体のレコード数をnと表現すると、

$$S_{AB} \approx \frac{S_A S_B}{n}$$

が成立する場合に、この相関ルールは統計的には意味がないと言える。

また、この指標に基づいて負の相関ルールを抽出することは、候補アイテムセットが膨大となるため困難であった。

我々はこの問題を解決するために、マイニングの指標として最小支持度と最小χ^2値を採用した。すなわち、最小支持度を満足するA→B という相関ルール候補に対して、AとBの間の自由度１のχ^2値を算出し、そのχ^2値が指定された最小χ^2値以上のルールを出力する。これにより出力される相関ルールは統計的な意味があるもののみ出力される。ただし、統計的なχ^2適合度検定においては、最小χ^2値として一般には3.8(有意水準 5%)や6.6(有意水準 1%)が使用されるが、我々は相関ルールの絞り込みを行うためにそれよりもかなり大きい数値を用いている（具体的な数値はマイニング対象データに依存する）。

これは、相関ルールのχ^2値は、事象A，Bを２値（例えば０と１）の数値と仮定した場合の相関係数と下記の関係が存在するためであり、χ^2値によるマイニングは相関係数によるマイニングと等価と言える。

$$\chi^2 = n(\text{相関係数})^2$$

なお、本稿では便宜上相関ルールをA→Bと表現しているが、χ^2値を用いてマイニングした相関ルールには方向性がなく、A→BはB→Aと等価であって、「AとBには相関関係が存在する」ことを示している。

Mining Strategy on Data Mining System KNODIAS.
Akitoshi MITSUISHI, Takahiko YAMAZAKI, Yasushi OBATA, Hidetoshi TANAKA, Masashi SHIRAISHI †
† Mitsubishi Electric Corporation

2．2 負の相関ルールマイニング

A→Bという負の相関ルールが最小χ^2値αを満足する条件は下記の通りである。

$$S_{AB} \leq \frac{S_A S_B}{n} - \frac{1}{n}\sqrt{\frac{\alpha S_{AB}(n-S_A)(n-S_B)}{n}}$$

ここに、$0 \leqq S_{AB}$ なる条件を加え式を変形すると、

$$S_B \geq \frac{\alpha n(n-S_A)}{nS_A + \alpha(n-S_A)} \quad \cdots ①$$

が成り立つ。この式の一例を図１に示す。図において網掛けを施した領域外には負の相関ルールが存在しない。

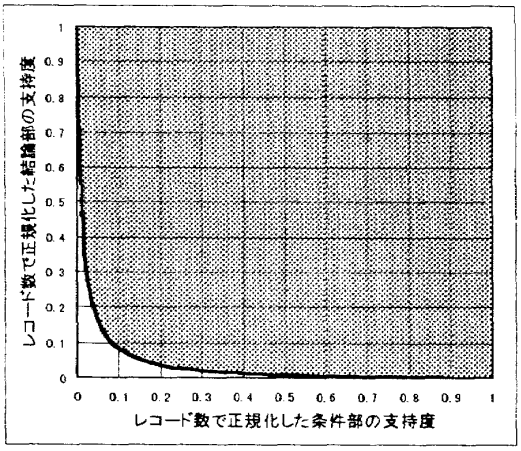

図1 負の相関ルールが存在する領域

図において、相関ルールが存在しない領域は狭く見えるが、条件部を構成するアイテムセットの大多数がこの領域に入るため、探索すべき候補アイテムセットは比較的小さくなり、負の相関ルールマイニングが可能となっている。なお、負のマイニングにおいては、最小支持度を①式によって内部的に算出している。

3 Knodiasにおけるエンジンの種類

Knodias ではχ^2適合度検定に基づく下記のエンジンを実装している。

① 正の相関ルール抽出エンジン
② 正の相関ルール抽出エンジン（有限メモリ版）
③ 負の相関ルール抽出エンジン
④ ルール検定エンジン

このうち④はユーザの仮説を検証するものであり、例えば、２つの相関ルールA，B→CとA，B→Dが得られた時にA，B→C，Dという仮説を検証するために用いる。

4 評価

図２は5770名分の健康診断データから正／負の相関ルール抽出を行った[5]時に得られたルール数を示したものである。図において“正のマイニング１”は最小支持度を１とした場合、“正のマイニング２”は最小支持度は負のマイニングエンジンが内部計算した最小支持度を適用した場合である。

“正のマイニング１”、“負のマイニング”は各々指定された最小χ^2値を満足する相関ルールを全て抽出している。“正のマイニング２”は指定された最小χ^2値を満足する相関ルールのうち、最小支持度値が小さいルールを出力してない。

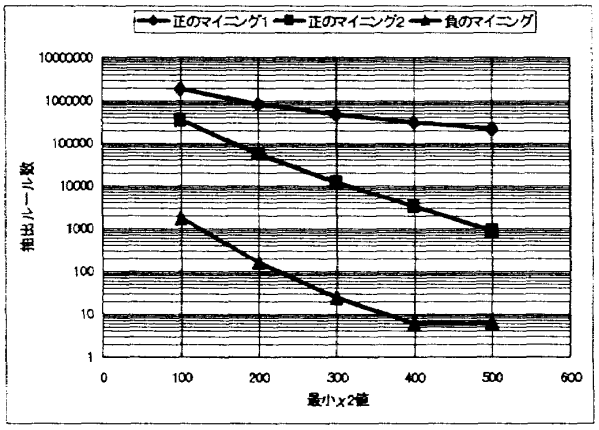

図2 マイニングによって得られた相関ルール数

負の相関ルール数は正に比べて格段に少ないが、得られたルールには意外で有用なものが多く含まれ、対象データを裏側から分析する強力な手段となっている。

5 まとめ

χ^2値（相関係数）による相関ルール抽出により、統計的に意味のあるルールのみを抽出できるようになった。また、負の相関ルール抽出をも可能とした。

今後は、各マイニングエンジンの更なる高速化を行っていく予定である。

参考文献

[1]山崎,他：Knodiasにおけるデータの性質に着目した相関ルール抽出の効率化,第56回情処全国大会1998.
[2]小幡,他：有限メモリ空間で相関ルールを抽出するマイニングアルゴリズム,第56回情処全国大会1998.
[3]Agrawal, R., Srikant, R. : “Fast Algorithm for Mining Association Rules” , Proc. VLDB '94.
[4]福田,他：相関ルールの可視化について,信学技報DE95-6 1995.
[5]加藤,他：データマイニングの手法を用いた検診データの解析，第37回近畿産業衛生学会1997.