ブール代数型データマイニングツール

4K－10

新上 和正　下川 信祐
(株)ATR 環境適応通信研究所

収集した色々なデータの中に隠れた相関関係を、任意のブール論理型の計算式を入力することで探るツールを作りました。これは0と1を要素とする行列で表現される任意の入力データに適用出来ます。このツールの特徴は、(1) 多くのツールがデータに予め相関構造を仮定するのに対して、それを行なわない反面、(2) 予め決められた相関構造を利用者に提示するのではなく、利用者自ずからが相関構造を模索する点にあります。

§1. 始めに

いろいろな社会調査やアンケート調査では、得られたデータからデータの表面には現れない相関関係を見出すことが主な関心事となります。多くのデータマイニングのツールは、データに、例えば、χ^2 相関構造を予め仮定して、相関関係を計算しています。これだと、仮定した相関構造に関しての相関関係を得ることが出来ますが、逆にこの相関関係以外のデータに潜む相関構造を探し出すことが出来ません。

ここでは、入力するデータは0と1を要素とする行列で表されるという前提を設けますが、データに潜む相関構造を仮定しない、任意のブール代数式を計算することで相関構造を探るツール[1]を紹介します。

§2. 入力データ

（入力データのファイルの中身）

	1	2	3	4	5	6	7	8	9	10	.....	M（列番号）
1	0	1	0	0	.	.	.	.	.	.		.
2	0	1	0	1	.	.	.	.	.	.		.
3	.	.	.	.	.	.	.	.	.	.		.
4	.	.	.	.	.	.	.	.	.	.		.
5	.	.	.	.	.	.	.	.	.	.		.
.												
.												
.												
N	.	.	.	.	.	.	.	.	.	.		.

（行番号）

入力データは0か1です。例えば、アンケート調査で行番号を持つ人は列番号を持つ質問項目にYesと答えれば1を、Noと答えれば0を割り振ることにします。

§3. ブール代数式の種類

入力出来る代数式の種類は、(a)-(f) に分類出来ます。

(a)　3, 4, 85, 100, 25　　→ 各々の質問番号にYesと答えた人の総数が表示される
(b)　((3+4) * 55+22), 55
(c)　33 | 45　　→ "|" の導入で質問番号33と45の集合関係[2]が計算される
(d)　33, 34 | 50, 51
(e)　345: 46, 65|78　　→ "345:" の導入で以下の入力論理式に制限が課される
(f)　33, 34 | 50, 51 | 150, 123 | 78, 48 | 230, 340[3]

＊の演算記号の意味：22*34は22と34を同時にYesと回答した人を取る
＋の演算記号の意味: 22 + 34は22か34のどちらか(両方を含める)をYesと回答した人を取る
(演算順序は、算数の演算と同様に＊を最優先でその後に＋の演算を行う)

上で入力は(c)の33を33*(45+23*100+34)+23*(21+11)*32... などのように、(、)、＋、＊の記号を使って幾らでも複雑に拡張することができます。

入力の簡略化：#： 2#5 := 2,3,4,5, %： 2%5 := 2+3+4+5, &： 2&5 := 2*3*4*5

Boolean Algebraic Data Mining Tool,
SHINJO Kazumasa, SHIMOGAWA Shinsuke
ATR Adaptive Communications Research Labs.,
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, JAPAN.
質問やコメントなどはelectronic address"infor@acr.atr.co.jp" へ.

実際に私たちが5校を対象に行なったインターネットに関するアンケート調査データに対する (d) の出力結果の例 (一部) を与えます。

```
          A=33 パソコン使いたくない理由: 使うのが面倒そう,B=50 Inet 利用場所: 自宅
( 1):33 <=> 50 ****
        (A*B):   0/112(1),  0/152(2),  0/112(3),  0/118(4),  0/108(5),  0/602(総計)
          A  :   2/112(1),  2/152(2),  0/112(3),  4/118(4),  0/108(5),  8/602(総計)
          B  :   9/112(1),  6/152(2),  3/112(3),  7/118(4),  0/108(5), 25/602(総計)
        相関1:   -.999(1),  -.998(2),   .000(3), -1.000(4),   .000(5),  -.998(総計)
        相関2:   -.040(1),  -.023(2),   .000(3),  -.047(4),   .000(5),  -.024(総計)
         A(%):    .00%(1),   .00%(2),   .00%(3),   .00%(4),   .00%(5),   .00%(総計)
         B(%):    .00%(1),   .00%(2),   .00%(3),   .00%(4),   .00%(5),   .00%(総計)
             :
```

これにより、最も基礎的な知識である、質問AとB及びAとBの両方にYesと回答した人の数、通常の相関値 (1) と特に小数の人が回答した場合に意味がある相関値 (2) などが計算されます。この基礎的な統計量をAやBを任意に組み合わせた式に対して出力するのがこのツールの基本的な特徴です。一般的に通常の相関値は上記のAとBと (A*B) を組み合わせることで計算されます。逆は、可能ではありません。

§4. ツールの (これ迄の) 利用範囲

上の例で分かるようにデータの相関構造について、強い制限を設けない反面、隠れた相関関係を利用者がトライアンドエラーをしながら探す必要があります。また、プログラム言語の素養があれば、ソースを修正することでいろいろな集計を取ることが可能です。それでは、このツールからデータのどのような性質を知ることができるのでしょうか? リストします。

(1) ある質問と他のある質問との関係: 多くの人が両方の質問にYesと答える。または、両方に同時にYesと答える人はいない など

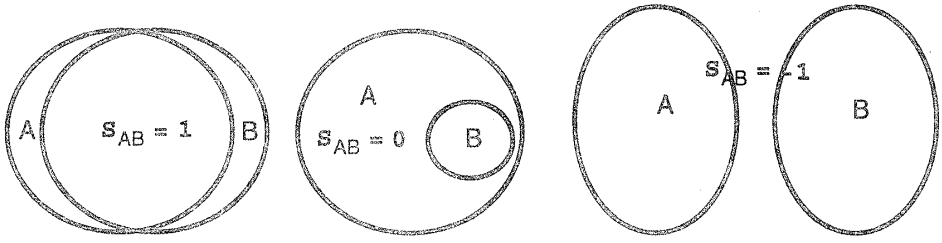
(2) ある質問と相関値の高い他の全ての質問がリストできる: 他のツールで主にやっている

(3) 全体の質問の中で特徴的な質問が分かる: 例えば、質問をグルーピングできる、また、グルーピングの核となる質問が分かる

(4) (1)~(3) は同様に質問の ((b) のような) 任意の組み合わせに対しても行なえる

(5) ある質問の集合と他の質問の集合の間の類似関係が分かる

(6) 通常の相関値が低くても、質問の間に強い相関がある場合がある：例えば、ある質問にYesと答えている全ての人は、他のある質問にもYesと答えている場合で、下図の相関値の低い真ん中の関係も見い出します



(7) マクロ集計が行なえる

などです。

謝辞：日頃、資料収集やデータ整理でお世話になっている林泰子 さんに感謝致します。

参考文献

1) 現在 公開されています。http://www.acr.atr.co.jp/~ shiomi/uniss/world2.HTML (このアドレスが近い内に変更されますが、公開されています) このツールは、現在Windowsやunixマシン上で動きます。また、このツールは購入することも可能です。(E-mail: infor@acr.atr.co.hp までお知らせ下さい)

2) 集合関係とは、(c) の例で言えば、質問33と45の両方にYesと回答した人数、各質問33と45にYesと回答した人数、両方にYesと回答した人数の割合についての知識を与えることを指します。((d) の場合の出力例を上で与えています)

3) この入力のタイプは、現在拡張中です。