# AICを用いた因果関係抽出手法の性能評価

2W－4

松本　一則　　橋本　和夫
国際電信電話株式会社

## 1. はじめに

筆者らは事象の履歴を記録したデータベースから因果関係を発見する問題について検討しており，これまでに単一事象が原因とみなせる場合の発見問題を定式化し，その具体的な解法を提案した[1]．同手法では，赤池の情報量基準(AIC基準)を用いて統計的に意味のないルールをフィルタリングしているが，同手法が大量の誤ったルール候補を全て棄却できたことが実験で確認されている[2]．

本稿では，$\chi^2$検定を用いた従来のフィルタリング手法とAICを用いた手法の性能を比較するために行った数値実験について報告する．

## 2. 因果関係発見問題とルールの棄却手法[1]

「事象 $e_x$ が発生してから，$\omega$ の時間が経過するまでの間に事象 $e_y$ が発生する」ことを意味する因果関係ルールは，　$e_x \overset{\omega}{\Longrightarrow} e_y$　$(e_x \neq e_y)$ のように表記される．因果関係発見問題は，与えられた事象の履歴から統計的に意味のあるルールを抽出する問題であり，ルールの棄却はを以下の様に行われる．

(1) 以下の事象の回数数を求める．

$n_{11}$ : $e_x \overset{\omega}{\Longrightarrow} e_y$ が成り立つ回数
$n_{12}$ : ($e_x$の出現回数) $- n_{11}$
$n_{21}$ : ($e_y$の出現回数) $- n_{11}$
$n_{22}$ : (各ルール候補が成り立つ回数の総和) $-(n_{11}+n_{12}+n_{21})$

この時，ルールが有意であるかどうかは，次の$2\times 2$分割表の検定問題となる．

| | $Y$ | $\neg Y$ |
|---|---|---|
| $X$ | $n_{11}$ | $n_{12}$ |
| $\neg X$ | $n_{21}$ | $n_{22}$ |

(2) $e_x$と $e_y$が独立に発生すると仮定するモデル(IM)と，依存関係があると仮定するモデル(DM)のAICを求める．

$$MLL_{IM} = (n_1+n_2)\,log\,(n_1+n_2)$$
$$+(n_1+n_3)\,log\,(n_1+n_3)$$
$$+(n_3+n_4)\,log\,(n_3+n_4)$$
$$+(n_2+n_4)\,log\,(n_2+n_4)$$
$$-2N\,log\,N$$

$$AIC_{IM} = -2\times MLL_{IM} + 2\times 2$$

$$MLL_{DM} = \sum_i n_i\,log\,n_i - N\,log\,N$$

$$AIC_{DM} = -2\times MILL_{DM} + 2\times 3$$

$$(N = n_1+n_2+n_3+n_4)$$

(3) $AIC_{IM} < AIC_{DM}$ ならばルールを棄却する．

## 3. AICを用いたフィルタの評価実験[2]

ポワソン過程に従う複数のシンボル発生源の出力記録から，シンボル間の因果関係を見つけるタスクに対し，AICを用いたフィルタの性能を測定し，以下の結果を得ている．

- 約$2.7\times 10^7$個の候補の中から約$2.5\times 10^3$個のルールを抽出しており，大量のルール候補を棄却できる．
- $\omega$や時系列の長さを変えて実験したが，誤検出率は常に0だった．
- 時系列が長くなるにつれ，見逃し率(正しいルールを発見できない確率)は単調に減少し，約27%まで下がったところまで確認した．

## 4. $\chi^2$検定を用いたフィルタとの比較実験

3.の実験でAICを用いたフィルタが実用的であることは示せた．しかし，「$\chi^2$を用いる従来の統計的なフィルタリング手法[3]に比べ，見逃し率は大きくないか」の点については不明のままである．そこで，$\chi^2$検定との比較評価進める上で，まず比較的実施しやすい数値実験を行った．

### 4.1 AICと$\chi^2$との抽出量の比較

表中の各セルの値が $1,2,3,\cdots,200$ となるような$2\times 2$表を全種類 ($200^4 = 1.6\times 10^9$種類) 生成し，棄却されずに残る表の数をそれぞれAICと$\chi^2$検定で求めた．得られた数は，$1.6\times 10^9$種類の事象の分布が与えられた時に棄却されずに残るルールの数でもある．

---

"Evaluation of Causality Detection Method Bsed on AIC", Kazunori MATSUMOTO, and Kazuo HASHIMOTO: Kokusai Denshin Denwa Co. Ltd.

表1に残ったルールを示す．AICの場合，その数は約1.3G個である．$\chi^2$の場合，有為水準$\alpha$によって数は異なり，$\alpha = 0.25$で約1.4G個，$\alpha = 0.1$で約1.3G個となっている．

$\alpha = 0.25$の場合，$\chi^2$検定で残るルールは，AICでも残るルールを全て含んでいる．

$\alpha = 0.1$の場合，$\chi^2$検定では残るがAICで棄却されるルールが10,864個あり，$\chi^2$検定で棄却されるがAICで残るルールが42,065,840個ある．

$\alpha \leq 0.05$では，$\chi^2$検定で残るルールは全てAICでも棄却されずに残る．

**表1: 棄却されないルールの数**

| ルール数 (AIC) |
|---|
| 1,341,761,348 |

| ルール数 ($\chi^2$) | | | ルール数 |
|---|---|---|---|
| $\alpha$ | $\chi^2$のみ | AICと共通 | (AICのみ) |
| 0.25 | 47,447,646 | 1,341,761,348 | 0 |
| | (計 1,389,208,994) | | |
| 0.1 | 10,864 | 1,299,695,508 | 42,065,840 |
| | (計 1,299,706,372) | | |
| 0.05 | 0 | 1,243,345,146 | 98,416,202 |
| | (計 1,243,345,146) | | |
| 0.025 | 0 | 1,193,414,560 | 148,346,788 |
| | (計 1,193,414,560) | | |

表1からすると，通常使用される有意水準の範囲($0.05 \leq \alpha \leq 0.1$)ならば，むしろ$\chi^2$検定のほうが見逃し率が高くなる可能性がある．

### 4.2 事象の分布特性を考慮した時の抽出量

図1, 図2は，$\alpha = 0.1$の時に，$\chi^2$検定でのみ抽出されるルールとAICでのみ抽出されるルールの例である．

| | $Y$ | $\neg Y$ |
|---|---|---|
| $X$ | 11 | 100 |
| $\neg X$ | 1 | 1 |

(a)

| | $Y$ | $\neg Y$ |
|---|---|---|
| $X$ | 11 | 1 |
| $\neg X$ | 100 | 1 |

(b)

| | $Y$ | $\neg Y$ |
|---|---|---|
| $X$ | 100 | 13 |
| $\neg X$ | 1 | 1 |

(c)

**図 1: $\chi^2$検定でのみ得られるルールの例**

| | $Y$ | $\neg Y$ |
|---|---|---|
| $X$ | 1 | 2 |
| $\neg X$ | 100 | 5 |

(a)

| | $Y$ | $\neg Y$ |
|---|---|---|
| $X$ | 87 | 79 |
| $\neg X$ | 100 | 66 |

(b)

| | $Y$ | $\neg Y$ |
|---|---|---|
| $X$ | 100 | 99 |
| $\neg X$ | 75 | 100 |

(c)

**図 2: AICでのみ得られるルールの例**

$\chi^2$検定やAICを用いる場合，以下の様なルール候補が与えられた時も，通常のルールと同じように棄却判定が行われる．

- $e_x \overset{\omega}{\Longrightarrow} \neg e_y$
- $\neg e_x \overset{\omega}{\Longrightarrow} e_y$

しかし，因果関係抽出の場合，上記のルールを候補に挙げないため，図1,2の(a),(b)のような事象分布を判定することはない．また，事象の組み合わせでルール候補が生成されるので，$2 \times 2$表の下段の値が上段の値に比べて小さい(c)のようなことも起こらない．そこで，事象の分布特性を反映した状況での抽出量を比較するため，以下の条件を満たす$2 \times 2$表で，棄却されないルールの数を調べたものが表2である($\alpha = 0.1$，各セルの値は$1, 2, \cdots, 100$)．

(A) $a_{11} > a_{12}$

(B) $\frac{a_{11}}{a_{11}+a_{12}} > \frac{a_{21}}{a_{21}+a_{22}}$

(C) $a_{11} + a_{12} < a_{21} + a_{22}$

**表2: 分布条件を適用した時の抽出量**

| 適用した条件 | $\chi^2$のみ | AICのみ |
|---|---|---|
| なし | 1,640 | 3,694,743 |
| A | 724 | 1,823,860 |
| A+B | 410 | 1,169,775 |
| A+B+C | 0 | 68,759 |

条件(A),(B),(C)を考慮した場合，$\chi^2$でのみ残るルールがなかったことから，因果関係抽出問題においては，$\chi^2$を用いた棄却方式の方が見逃し率がむしろ高くなる可能性があることが分かった．

## 5. おわりに

以前行ったシミュレーション実験では，既提案のAICを用いたルールの棄却手法が実用的であることは示したが，従来の$\chi^2$を用いた手法との優劣は不明だった．そこで，今回，数値実験を行った結果，因果関係を抽出目的の場合，AICで棄却する方が見逃し率が低くなるとの感触を得た．今後，両手法の得失を解析的もしくはシミュレーション実験によって明らかにしていく予定である．

## 参考文献


[1] 松本等: 因果関係発見手法の検討, 情処全大 53 回平成 8 年後期 1S-01

[2] 松本 一則, 橋本 和夫. 因果関係発見手法の性能評価実験, 情処全大 54 回平成 9 年前期 2R-06

[3] 福田等: 相関ルールの可視化について, 信学技報 DE95-6 1995