1U－10

AICを用いたデフォルトルール生成法の拡張

三澤邦嗣　山口文彦　斎藤博昭　中西正和
慶應義塾大学大学院 理工学研究科 計算機科学専攻

1. はじめに

データベースから事象の間で因果関係を述べるルールを発見することは、ルールにおける原因が、ルールを発見した後で観測された場合、結果を予測できるため、重要であると思われる。

不完全な情報のもとで推論を行なうための論理の一つとして、デフォルト論理がある [1]。

デフォルト論理では、

$$\frac{\alpha : \mathbf{M}w}{w} \qquad (1)$$

という形で表される、デフォルトルールを用いて推論を行なう。α, w は論理式である。また $\mathbf{M}$ は、それが作用する命題が成り立つと仮定したとき矛盾が生じないならば、結論が成り立つ、ということを意味している。したがって式 (1) は、前提 α が成り立つとき、w が成り立つとして矛盾が生じないとき、w が成り立つ、ということを意味する。

松本、橋本らは、事象間の因果性を記述したデフォルトルールをデータから生成する手法を提案した [2]。

本研究の目的は、松本、橋本らの先行研究を、より相関性の高い、あるいはより記述力の高いデフォルトルールを、データから生成できるように、拡張することである。

2. 先行研究

松本、橋本らの先行研究では、まず Agrawal らの相関ルール発見手法 [3] を不確かなデータに適用できるよう拡張し、事象間に共起の可能性があるものを発見し、ルールを生成する手法を提案した。次に生成したルールに対し、因果性の有無を AIC を用いて検証している。

2.1 相関ルール発見手法

松本、橋本らの研究では、記号を以下のように定めている。

表 1: 観測された回数

	H	$\neg H$
B	n_{11}	n_{12}
$\neg B$	n_{21}	n_{22}

$c_1, \cdots, c_N$	観測されたデータの中のケース
$x_1, \cdots, x_K$	ケースにおける事象
$v_i(x_1), \cdots, v_i(x_K)$	ケース c_i において、x_j が観測される確率
i	ケースの添字 $i = 1, \cdots, N$
j	事象の添字 $j = 1, \cdots, K$

また、データにおける関係

$$\frac{p_1 \wedge p_2 \wedge \cdots \wedge p_m : \mathbf{M}q}{q} \qquad (2)$$

の一つを R としたとき R が持つ特性値 $Conf(R)$ を以下のように定めている。

$$Conf(R) = \frac{\sum_{c_i}(\prod_{j=1}^{m} v_i(p_j) \cdot v_i(q))}{\sum_{c_i}(\prod_{j=1}^{m} v_i(p_j))} \qquad (3)$$

ただし $p_j, q \in \{x_1, \neg x_1, \cdots, x_K, \neg x_K\}, (1 \leq j \leq m \leq K)$ $Conf(R)$ は、確信度のことで、0 から 1 の間の値を取る。

確信度が高いほど、相関関係が高いことを意味する。ここで、閾値を定めて、その値より低い関係を取り除く。

2.2 AIC を用いた因果性の有無の検証

ここで、B を状況、H を結果とする。n_{11} を B と H が同時に観測された回数とする。n_{ij} $i, j = 1, 2$ も表 1 に準じて定める。

2.1 で生成される関係は、相関性の低いものも多い。そこで、AIC を用いて相関性の低い関係を取り除く。ここで、AIC を用いるのは、経験による閾値の設定などを用いることをせずに、観測された事象の数だけから関係の有無を確かめることができるからである [4]。AIC を用いるに当たって、二つのモデルを考える。一つは B と H が独立ではないというモデルである。これを DM とよぶ。もう一つは B と H が独立

Expansion of Generation of Default Rules Based on AIC
Kunitugu MISAWA Fumihiko YAMAGUCHI Hiroaki SAITO Masakazu NAKANISHI
Department of Computer Science, Faculty of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223, Japan

であるというモデルである。これを IM とよぶ。この二つのモデルにおいて、計算されるそれぞれの AIC の値が小さい方がモデルとして適当である。AIC とは、この二つのモデルのどちらが現実に合っているかを確かめるための基準である。

B と H の間の関係を表す確率のパラメータは、DM において三つ、IM において二つである。

ここで、AIC は、次のように表される。

$$AIC = -2 \times MLL + 2 \times M \tag{4}$$

MLL は、以下で述べる。M は、独立なパラメータの数である。

$h = n_{11} + n_{12}, k = n_{11} + n_{21}$,
$N = \sum n_{ij}$ としたとき、$AIC(DM), AIC(IM)$ を、以下のように定める。

$$\begin{aligned} MLL(DM) &= \sum(n_{ij} \log n_{ij}) - N \log N & (5)\\ AIC(DM) &= -2 \times MLL(DM) + 2 \times 3 & (6)\\ MLL(IM) &= h \log h + k \log k \\ &\quad +(N-h)\log(N-h) \\ &\quad +(N-k)\log(N-k) \\ &\quad -2N \log N & (7)\\ AIC(IM) &= -2 \times MLL(IM) + 2 \times 2 & (8) \end{aligned}$$

ここで、2.1 で相関関係を認められた関係の中で、$AIC(DM)$ が $AIC(IM)$ より大きいものを取り除く。以上の操作で、取り除かれなかったものが求めるデフォルトルールである。さらに、先行研究では、式 (2) のように左辺が論理積になっている場合についても、デフォルトルールを求める方法を述べてある。

3. 本研究の方針

先行研究においては、式 (2) から求められる結論が一つの命題 q のときしか考慮されていない。そこで、本研究では、結論が複数の属性を表す命題の論理積、論理和、排他的論理和となるデフォルトルールを、求める方法を考案する。

ただし、注意しなければならないのは、トートロジーや前提が偽となるような病的なデフォルトルールを作らないようにすることである。

$Conf(R)$ の数学的拡張として、以下のようなものが考えられる。例えば、

$$\frac{\wedge_j p_j : \mathbf{M}(q \vee q')}{q \vee q'} \tag{9}$$

と表される関係 R の相関性を確かめるための式 $Conf(R)$ は、

$$\frac{\sum_{c_i}(\prod_{j=1}^{m} v_i(p_j) \cdot (v_i(q) + v_i(q') - v_i(q) \cdot v_i(q')))}{\sum_{c_i}(\prod_{j=1}^{m} v_i(p_j))} \tag{10}$$

と考えられる。また、AIC であるが、たとえば表 1 を、式 (9) の場合に拡張すると、$MLL(IM)$ は次のようになる。

$$\begin{aligned} MLL(IM) &= h \log h + k \log k + l \log l \\ &\quad t \log t + u \log u \\ &\quad (N-u)\log(N-u) \\ &\quad -2N \log N & (11) \end{aligned}$$

ただし $h = n_{11} + n_{12}, k = n_{13} + n_{14}, l = n_{22} + n_{23}, t = n_{23} + n_{24}, u = n_{12} + n_{22} + n_{14} + n_{24}, N = \sum_{ij} n_{ij}$, $n_{ij}(i = 1, 2, j = 1, 2, 3, 4)$ は次のように定める。$i = 1$ のとき p が観測され、$i = 2$ のとき p が観測される。$j = 1, 2$ のとき、q が観測される。j が奇数のとき、q' が観測される。

4. 今後の展望

3. で提案した手法を現実に基づくデータベースに用いて、デフォルトルールを得る。その後に得られたデータから、デフォルトルールを用いて結果を予測し、その予測の的中率から、$Conf(R), AIC$ の数学的拡張の妥当性を確かめる。

参考文献

[1] R. Reiter: *A Logic for Default Reasoning*, Artificial Intelligence pp. 81-132 vol. 13, 1980.

[2] 松本一則, 橋本和夫: *AIC* を用いたデフォルトルールの生成, 電子情報通信学会技術研究報告, pp. 25-30, AI97-27, 1997-11.

[3] R. Agrawal, R. Srikant: *Fast Algorithms for Mining Association Rules*, IBM Research Report, June 1993.

[4] H. Akaike: *A New Look at the Statistical Model Identification*, IEEE Transactions on Automatic Control, pp. 716-722, vol. ac-19, no.6, December, 1974.