データベースシステム 116－80
（1998. 7. 10）

不完全データの論理的解析

茨木 俊秀（京都大学情報学研究科）
牧野和久（大阪大学基礎工学研究科）
Endre Boros (Rutgers University)
Email: ibaraki@i.kyoto-u.ac.jp, makino@sys.es.osaka-u.ac.jp

Abstract

正例のデータ集合 $T \subseteq \{0,1\}^n$ と負例のデータ集合 $F \subseteq \{0,1\}^n$ からなる部分定義論理関数 (T,F) が与えられたとき、それと整合する論理関数 $f : \{0,1\}^n \to \{0,1\}$（すなわち拡大）を求める問題はデータの論理的解析の一形態である。本研究では、データに誤りや不完全ビットが存在する場合の対応として、誤りベクトル数を最小にするという基準、不完全ビットの扱いに関してロバスト性の度合に基づく3種の基準を提案し、それらの下で拡大を求める問題を計算の複雑さの観点から調べた。その結果、f の属する関数のクラス $\mathcal{C}$ に応じて多項式時間で解ける場合と NP 困難になる場合がどのように区別されるかが明らかになった。

Logical Analysis of Incomplete Data

Toshihide Ibaraki (Kyoto University)
Kazuhisa Makino (Osaka University)
Endre Boros (Rutgers University)

Abstract

Given a partially defined Boolean function (T,F) defined by a set of positive examples $T \subseteq \{0,1\}^n$ and a set of negative examples $F \subseteq \{0,1\}^n$, the problem of finding a Boolean function $f : \{0,1\}^n \to \{0,1\}$ consistent with it (i.e., an extension of (T,F)) has been studied as logical analysis of data. This paper considers the situation in which data may contain errors and/or incomplete bits. We introduce the criterion of minimizing the number of error vectors and three criteria of how to allow robustness of incomplete data. Under these criteria, we clarify the complexity of the problems to find the existence of extensions in several classes of functions $\mathcal{C}$. It turns out that some problems can be solved in polynomial time, while others are NP-complete.

1 まえがき

正例のデータ集合 $T \subseteq \{0,1\}^n$ と負例のデータ集合 $F \subseteq \{0,1\}^n$ が与えられたとき, 対 (T,F) を**部分定義論理関数** (partially defined Boolean function; pdBf) という. このとき T 内のベクトルに対しては値1を F 内のベクトルに対しては値0を出力する (完全定義) 論理関数 (Boolean function) $f : \{0,1\}^n \to \{0,1\}$ を構成する問題を考え, そのような f を pdBf (T,F) の**拡大** (extension) と呼ぶ.

この種の問題は, 大量のデータから意味のある情報を取り出すための研究として最近注目を集めている知識発見 (knowledge discovery) [8], データマイニング (data mining) [1, 13], データの論理的解析 (logical analysis of data; LAD) [3, 7, 12] といった分野における基本的なテーマであり, さらに学習理論 (learning theory) における学習可能性 [2, 15] やパターン認識における識別関数 (classifier) [10] などとも密接に関連している. また、論理設計の分野では、**不完全定義論理関数** (incompletely specified Boolean function) あるいはドントケア (don't care) のある場合の設計として、古くから研究がなされてきた。

pdBf (T,F) と拡大 f の具体的な意味を理解するため, 各ベクトル $a \in T \cup F$ が一人の患者の診察結果 $a = (a_1, a_2, \ldots, a_n)$ を表している場合を考えよう. ベクトルの各要素 a_j は j 番目の診察項目の結果を表しており, たとえばこれが血圧に対応しているならば, $a_j = 1$ は "血圧が高い", $a_j = 0$ は "高くはない"などの意味を持っている. 各要素 j は一般には**属性** (attribute) と呼ばれる. $a \in T$ はベクトル a を持つ患者が, この病気 (たとえばインフルエンザ) にかかっていると診断されたことを表し, $a \in F$ ならばその逆を表している. (T,F) の拡大 f は, 可能なすべてのデータベクトル $a \in \{0,1\}^n$ に対する診断結果を記述しており, f の中身を知ることは, 診断がどのようにしてなされたかの論理的説明を得ることに他ならない.

例 1 具体例として, 表 1 のデータが与えられたとしよう. この拡大の一つはたとえばつぎの論理関数で与えられる.

$$f = \bar{x}_1 x_2 \vee x_2 x_5 \qquad (1)$$

すなわち, f は属性1が値0で属性2が値1をとるか, あるいは属性2と5がともに値1をとればインフルエンザと診断し, そのどちらでもなければインフルエンザにはかかっていないと診断することを表している. □

表 1: pdBf (T,F) の例.

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
T	$a^{(1)} =$	0	1	0	1	0	1	1	0
	$a^{(2)} =$	1	1	0	1	1	0	0	1
	$a^{(3)} =$	0	1	1	0	1	0	0	1
F	$b^{(1)} =$	1	0	1	0	1	0	1	0
	$b^{(2)} =$	0	0	0	1	1	1	0	0
	$b^{(3)} =$	1	1	0	1	0	1	0	1
	$b^{(4)} =$	0	0	1	0	1	0	1	0

与えられた pdBf (T,F) に対し, その拡大 f は多数存在するのが普通である. 可能なすべての拡大の中から, (T,F) の論理的説明としてどの f を選ぶかが, 中心的課題の一つである.

拡大 f の決定に際し, つぎの2つの視点が重要と考えられる. その第一は,

- f の表現の簡潔さ,

である. これは, 「真理は単純で美しい」という期待を反映しており, 「ある事柄を説明するための仮説は必要最小限でなければならない」とするオッカムのかみそり (Occam's razor) の考え方にも通じる. もう一つの重要な視点は,

- f に関する構造的知識を具現,

すべきというものである. たとえば, 病気の診断において, 各属性の方向性は経験的あるいは病理学的理由によって明らかであることが多い. つまり, 体温は低いより高い方がインフルエンザにかかっていると診断され易い, などである. これは, 関数 f が体温という属性に関し正 (単調非減少) であることを要求していることに相当する. このような構造的知識をすべて満足する関数のクラスを $\mathcal{C}$ とすると, ここでの問題は「$f \in \mathcal{C}$ をみたす拡大を見出せ」という形に書かれる. なお, クラス $\mathcal{C}$ は, 上のような構造的知識からだけでなく, 応用に際しての要求から決定される場合もある. たとえば, 人工知能の分野でよく見られるように, f がホーン関数 (Horn function) であれば

ホーン規則 (Horn rule) によって処理できるので都合が良い, というような要求である. ここでは, $\mathcal{C}$ として,

1. すべての論理関数のクラス $\mathcal{C}_{ALL}$
2. 正 (positive, すなわち単調 monotone) 関数のクラス $\mathcal{C}_{+}$
3. ホーン (Horn) 関数のクラス $\mathcal{C}_{HORN}$
4. k-DNF 関数のクラス $\mathcal{C}_{k\text{-}DNF}$

を扱う (定義は 2 節).

ところで、データには誤りがつきものである. すなわち, $T \cup F$ の各ベクトル a の要素値が, 測定誤差などのために誤っている可能性, さらに a が T と F のどちらに属するかの判断の誤りが考えられる. また, データが不完全であり, 一部のビット値が不明 $a_j = *$ である場合もあろう. この理由には, 単にデータから欠落してしまった場合もあれば, 値を得るための測定が高価についたり危険であるため現時点では得られていないという状況も考えられる. 本論文では、主にこの観点から検討を加える。

2 諸定義

一般に $f : \{0,1\}^n \to \{0,1\}$ を (n 変数) **論理関数**という. $f(v) = 1$ をみたすベクトル $v \in \{0,1\}^n$ (正確には論理ベクトル) を f の**真ベクトル** (true vector), $f(v) = 0$ ならば v を**偽ベクトル** (false vector) という. f の真ベクトルの集合を $T(f)$, 偽ベクトルの集合を $F(f)$ と記す. 定義より, $T(f) \cap F(f) = \emptyset$ かつ $T(f) \cup F(f) = \{0,1\}^n$ である. ベクトル $v \in \{0,1\}^n$ に対し,

$$ON(v) = \{j \mid v_j = 1\}, \quad OFF(v) = \{j \mid v_j = 0\}$$

の記法を用いる.

n 変数論理関数 f がすべての $a, b \in \{0,1\}^n$ に対し, $a \leq b$ (つまり, $a_j \leq b_j$, $j = 1, 2, \ldots, n$) ならば $f(a) \leq f(b)$ という性質をもつとき**正関数** (positive function) であるという. 正関数はしばしば**単調関数** (monotone function) とも呼ばれる. 正関数のクラスを $\mathcal{C}_{+}$ と記す.

論理関数 f の DNF (論理和標準形)

$$\varphi = \bigvee_k t_k$$

において, 各項 t_k が多くとも一つの負リテラルしか持たないとき, それぞれを**ホーン項** (Horn term), その結果得られる φ を**ホーン DNF** (Horn DNF), さらにホーン DNF をもつ論理関数 f を **ホーン関数** (Horn function) といい [11, 14]、そのクラスを $\mathcal{C}_{HORN}$ と記す.

次の補題は, ホーン関数を特徴づける重要な性質である.

補題 1 [14] $f : \{0,1\}^n \to \{0,1\}$ がホーン関数であるための必要十分条件は $F(f) = C(F(f))$ であること (すなわち, $F(f)$ が $\wedge$ に関し閉じていること) である. □

DNF において, すべての項がたかだか k 個のリテラルしか含まないとき k-DNF であるという. 拡大 f が小さな k の k-DNF で表現できるとすれば, 表現の簡潔性を有していると理解できるので, そのような拡大を求めることは意味をもつ. k を与えられた定数 (正整数) とし, k-DNF をもつ関数のクラスを $\mathcal{C}_{k\text{-}DNF}$ と記す.

最後に, **部分定義論理関数** (pdBf) (T, F) に対し, $T(f) \supseteq T$ かつ $F(f) \supseteq F$ をみたす論理関数 f は (T, F) の**拡大** (extension) であるという. すでに述べたように、関数のクラス $\mathcal{C}$に対し次の問題が基本的である.

問題 EXTENSION($\mathcal{C}$)
入力: pdBf (T, F).
出力: (T, F) の拡大で $f \in \mathcal{C}$ をみたすものが存在すれば, そのような f を一つ; 存在しなければ "No".

幸いなことに、1節で述べたすべてのクラス $\mathcal{C}$に対し、問題 EXTENSION は多項式オーダー時間で解けることが分かっている [7, 4].

3 誤りを含むデータの処理

関数のクラス $\mathcal{C}$ において pdBf (T, F) の拡大は常に存在するわけではなく, 存在しない場合もある. もちろん, この関数クラス $\mathcal{C}$ が (T, F) の論理的説明を得るのにふさわしくなかったというのであれば, よりふさわしいクラスを見つけることが先決である. しかし, 何らかの正当な理由で $\mathcal{C}$ が選ばれているのであれば, 拡大 $f \in \mathcal{C}$ が存在しない理由が, データ集合 (T, F) に含まれる誤りによるという可能性もある. 誤りのタイプとしては, データベクトル $a \in T \cup F$ 内のいくつかの要素の値 0, 1 が間違って記録された場合, また, ベクトル a の T と F への分類が間違っ

たという場合も考えられる. したがって, このような場合, (T,F) に対する判定誤りの個数を最小にするような拡大 $f \in \mathcal{C}$ を求める問題が重要となる. そのような拡大を**最良適合拡大** (best-fit extension) といい, 正確には以下のように書かれる.

問題 BEST-FIT($\mathcal{C}$)
入力: pdBf (T,F) と重み関数 $w : T \cup F \to \mathbf{R}_+$.
出力: 次の性質をもつ pdBf (T^*,F^*) とその拡大 $f \in \mathcal{C}$.

1. $T^* \cap F^* = \emptyset$, $T^* \cup F^* = T \cup F$.
2. pdBf (T^*,F^*) は拡大 $f \in \mathcal{C}$ をもつ.
3. $w(T^* \cap F) + w(F^* \cap T)$ を最小にする.

なお, $\mathbf{R}_+$ は非負実数の集合であり, $w(X) = \sum_{a \in X} w(a)$ の記法を用いる. 上の条件は, 元の pdBf (T,F) のかわりに, pdBf (T^*,F^*) を用いるならばクラス $\mathcal{C}$ 内に拡大が存在することを述べている. このとき, $T^* \cap F$ は元の負例 $b \in F$ の中で (T^*,F^*) では正例と判定されるものの集合, $F^* \cap T$ は元の正例 $a \in T$ の中で (T^*,F^*) では負例と判定されるものの集合を示しているので, 両者を合わせて誤りベクトルの集合である. したがって, この問題では誤りベクトルの重み和を最小にすることが求められている. すべての $a \in T \cup F$ に対し $w(a) = 1$ であれば, 誤りベクトルの個数の最小化である.

定理 2 BEST-FIT($\mathcal{C}$) は $\mathcal{C}_{ALL}$ と $\mathcal{C}_+$ に対して多項式時間のアルゴリズムをもつ. しかし, $\mathcal{C}_{HORN}$, $\mathcal{C}_{k\text{-}DNF}$ に対しては NP-困難である.

証明: $\mathcal{C}_{ALL}$ と $\mathcal{C}_+$ についてのみ示す. NP-困難性については [4] の証明を参照のこと.

$\mathcal{C}_{ALL}$: (T,F) が拡大 $f \in \mathcal{C}_{ALL}$ をもたないならば, $T \cap F \neq \emptyset$ が成立する. T と F の両方に属している同じベクトル a は, それを正例と判定しても負例と判定しても誤りの重みは $w(a)$ と評価とされるので, たとえば

$$T^* = T - F, \quad F^* = F$$

とすれば, pdBf (T^*,F^*) は BEST-FIT($\mathcal{C}_{ALL}$) の正しい出力である.

$\mathcal{C}_+$: pdBf (T,F) が $\mathcal{C}_+$ において拡大を持たないとすれば, $a \leq b$ をみたす $a \in T$ と $b \in F$ が存在する. そこで, 次のグラフ $H_{(T,F)} = (T \cup F, E)$ を定義しよう.

頂点集合: T および F

辺集合: $(a,b) \in E \iff a \leq b,\ a \in T,\ b \in F$

すなわち, $H_{(T,F)}$ は T 側の頂点集合と F 側の頂点集合に分かれる 2 部グラフであって, 辺 (a,b) は常に T 側と F 側を接続している. $H_{(T,F)}$ において, 次の性質 1 を持つ頂点集合 U を $H_{(T,F)}$ の**頂点カバー** (vertex cover), さらに性質 2 もみたすならば**最小頂点カバー**という.

1. 任意の辺 (a,b) は $a \in U$ あるいは $b \in U$ をみたす.
2. U は性質 1 をもつ頂点集合の中で $w(U)$ を最小にする.

最小頂点カバーを求める問題は, 一般のグラフに対しては NP-困難であるが, 2 部グラフに対してはネットワークフローのアルゴリズムを用いて多項式時間 $O((|T|+|F|)^3)$ で解けることが知られている [9, ?]. $H_{(T,F)}$ の最小頂点カバー U はさらにつぎの性質 (i) と (ii) をもつ.

(i) BEST-FIT($\mathcal{C}_+$) の解 (T^*,F^*) に対し $w(T^* \cap F) + w(F^* \cap T) \geq w(U)$.

なぜなら, (T^*,F^*) の任意の拡大 $f \in \mathcal{C}_+$ において, 集合

$$\begin{aligned} W &= \{b \in F \mid f(b) = 1\} \cup \{a \in T \mid f(a) = 0\} \\ &\quad (= (T^* \cap F) \cup (F^* \cap T)) \end{aligned}$$

を作ると W は節点カバーとなっているからである (そうでなければ, $a \leq b$ かつ $f(a) = 1$, $f(b) = 0$ なる $(a,b) \in E$ が存在するので f が正関数であることに反する).

(ii) $T^* = (T - U) \cup (F \cap U)$, $F^* = (T \cap U) \cup (F - U)$ と定義すると, (T^*,F^*) は $\mathcal{C}_+$ において拡大をもち, しかも, $w(T^* \cap F) + w(F^* \cap T) = w(U)$ が成立する.

すなわち, 最小頂点カバー U が求まると, (ii) によって, BEST-FIT($\mathcal{C}_+$) の解 (T^*,F^*) をただちに構成できるわけである. □

例 2 pdBf (T,F) の例として,

$$\begin{aligned} T &= \{(01100), (01010), (00110), (00101), \\ &\quad (00111)\}, \\ F &= \{(01011), (11010), (01110), (00111)\} \end{aligned}$$

を考える. $n=5$ である. また重み関数 w にはすべての $a \in T \cup F$ に対し $w(a)=1$ とする.

まず, $\mathcal{C}_{ALL}$ における最良適合拡大は, $T \cap F = \{(00111)\}$ に注意して,

$$T^* = T - \{(00111)\}, \quad F^* = F$$

とすれば, pdBf (T^*, F^*) が求める出力である.

つぎに, $\mathcal{C}_+$ の最良適合拡大を求めるために, 2 部グラフ $G_{(T,F)}$ を作ると図 1 を得る. このグラフの最小頂点カバー U は, 容易にわかるように

$$U = \{(01010) \in T, (01110) \in F, (00111) \in F\},$$

とすればよい. こうすれば 3 頂点ですべての辺をカバーできるが, これ以外の頂点集合では少なくとも 4 頂点必要である. したがって, 定理 2 の証明の (ii) にしたがって

$$\begin{aligned} T^* &= (T-U) \cup (F \cap U) \\ &= \{(01100), (00110), (00101), (00111), \\ &\quad (01110)\} \\ F^* &= (T \cap U) \cup (F-U) \\ &= \{(01010), (01011), (11010)\} \end{aligned}$$

とすればよい. この (T^*, F^*) は, $a \leq b$ をみたす $a \in T^*, b \in F^*$ を持たないので, $\mathcal{C}_+$ における拡大が存在する. □

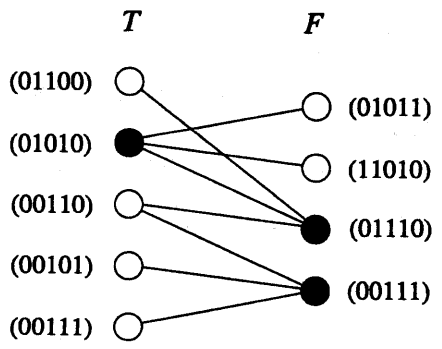

図 1: 例題 2 の (T,F) に対する 2 部グラフ $H_{(T,F)}$ (黒丸は最小頂点カバーを示す).

4 不完全データの処理

現実のデータには, 誤りが含まれているばかりでなく, ベクトルのいくつかの値が決定していないという意味で不完全なものも少なくない. データの中の不完全ビットを "$*$" の記号で示す. すなわち, 正例と負例のデータ集合は,

$$\tilde{T} \subseteq \{0,1,*\}^n, \quad \tilde{F} \subseteq \{0,1,*\}^n$$

なる対 $(\tilde{T}, \tilde{F})$ によって与えられる. これを以下, pBmb (partially defined Boolean function with missing bits) と記す.

pBmb $(\tilde{T}, \tilde{F})$ に対する拡大 $f \in \mathcal{C}$ の存在を議論するにあたって不完全ビット $*$ の扱いに, いくつかの可能性が考えられる. まず第一は, 各 $*$ が 0 と 1 の値のどちらであっても f が $(\tilde{T}, \tilde{F})$ の拡大になっているというものである. 一方, $*$ が 0 と 1 のどちらかの適当な値をとれば f は $(\tilde{T}, \tilde{F})$ の拡大になっているという定義も考えられる. 前者を **ロバスト拡大** (robust extension), 後者を **整合拡大** (consistent extension) という. ロバスト拡大は, 不完全ビットの値によらないという安全側に立った拡大である. ロバスト拡大が存在しなくても整合拡大は存在する場合がある. 実際には, すべての不完全ビットは 0 か 1 の値をとるべきという立場に立てば, 整合拡大の存在は実用的な意味からも重要である. また, $f \in \mathcal{C}$ の拡大の存在が何らかの理由で確かであれば, 整合拡大によって得られた不完全ビットへの 0, 1 の値の割り当てによって, 逆に, 不完全ビットの値を推定することもできる.

最後に, 一部の不完全ビットの値を適当に定めれば, 残りの不完全ビットは自由な値のままで拡大の存在が示せる場合もあろう. この場合, 値を固定する不完全ビットの個数を最小にするという問題が考えられるが, これを **最大ロバスト拡大** (most robust extension) という.

以上の問題を正確に記述するため, いくつか記号を導入する. ベクトル集合 $\tilde{S} \subseteq \{0,1,*\}^n$ に対し

$$AS(\tilde{S}) = \{(v,j) \mid v \in \tilde{S},\ v_j = *\}$$

と定義する. すなわち, $\tilde{S}$ 内の不完全ビットのベクトルとその位置を示している. $\tilde{S}$ が一つのベクトル v から成る場合は $AS(v)$ と記すこともある. $Q \subseteq AS(\tilde{S})$ に対し, 0 あるいは 1 への **割り当て** (assignment) $\alpha : Q \to \{0,1\}$ を考え, この割り当てによる結果を

$$\tilde{S}^\alpha = \{v^\alpha \mid v \in \tilde{S}\},$$

$$v_j^\alpha = \begin{cases} \alpha(v,j) & \text{if } (v,j) \in Q \\ v_j & \text{if } (v,j) \notin Q \end{cases}$$

のように記す. たとえば,

$$\tilde{S} = \{u = (1,*,0,1),\ v = (0,1,*,*),\ w = (1,1,*,0)\}$$

とすると, $AS(\tilde{S}) = \{(u,2),(v,3),(v,4),(w,3)\}$ である. $Q = \{(u,2),(v,4)\}$ に対し割り当て $(\alpha(u,2),\alpha(v,4)) = (1,0)$ を考えると, $\tilde{S}^\alpha = \{u^\alpha = (1,1,0,1),\ v^\alpha = (0,1,*,0),\ w^\alpha = (1,1,*,0)\}$ である. また, 与えられた pBmb $(\tilde{T},\tilde{F})$ に対し

$$AS = AS(\tilde{T} \cup \tilde{F})$$

と記す.

すなわち, 論理関数 $f : \{0,1\}^n \to \{0,1\}$ が pBmb $(\tilde{T},\tilde{F})$ のロバスト拡大であるとは, すべての割り当て $\alpha : AS \to \{0,1\}$ に対し

$$\begin{aligned} f(a^\alpha) &= 1, \quad a \in \tilde{T} \\ f(b^\alpha) &= 0, \quad b \in \tilde{F} \end{aligned} \tag{2}$$

が成立することである. また, (2) をみたすような割り当て α が一つでも存在すれば整合拡大である.

問題 RE($\mathcal{C}$) **(ロバスト拡大)**
入力: pBmb $(\tilde{T},\tilde{F})$.
出力: $(\tilde{T},\tilde{F})$ がロバスト拡大 $f \in \mathcal{C}$ をもてばそのような f の一つ; 存在しなければ "No".

問題 CE($\mathcal{C}$) **(整合拡大)**
入力: pBmb $(\tilde{T},\tilde{F})$.
出力: $(\tilde{T},\tilde{F})$ が整合拡大 $f \in \mathcal{C}$ をもてばそのような f の一つとそれを実現する割り当て $\alpha : AS \to \{0,1\}$; 存在しなければ "No".

問題 MRE($\mathcal{C}$) **(最大ロバスト拡大)**
入力: pBmb $(\tilde{T},\tilde{F})$.
出力: $(\tilde{T}^\alpha,\tilde{F}^\alpha)$ がロバスト拡大をもつような割り当て $\alpha : Q \to \{0,1\}$, $Q \subseteq AS$, の中で $|Q|$ を最小にするものを一つ; そのような α が存在しなければ "No".

定義から, RE($\mathcal{C}$) と CE($\mathcal{C}$) は EXTENSION($\mathcal{C}$) を特別な場合として含む. また, MRE($\mathcal{C}$) は RE($\mathcal{C}$) と CE($\mathcal{C}$) のどちらよりも一般的である. したがって, EXTENSION($\mathcal{C}$) が NP 困難であれば RE($\mathcal{C}$) と CE($\mathcal{C}$) はどちらも NP 困難であり, また, RE($\mathcal{C}$) と CE($\mathcal{C}$) のどちらかが NP 困難であれば MRE($\mathcal{C}$) も NP 困難である. また, 多項式時間アルゴリズムの存在については, 逆方向の議論が成り立つ.

2つのベクトル $a,b \in \{0,1,*\}^n$ に対し, ある割り当て $\alpha : AS(\{a,b\}) \to \{0,1\}$ が存在して $a^\alpha = b^\alpha$ とできるとき $a \approx b$ と記す. あるいは, $a^\alpha \le b^\alpha$ とできるならば $a \preceq b$ と記す. たとえば, $(0*1*) \approx (*110)$, $(0*1*) \preceq (*110)$ であるが, $(0*1*) \not\approx (1110)$, $(0*1*) \not\preceq (*10*)$ である. つぎに, $a \in \{0,1,*\}^n$ に対し, $AS(a)$ の不完全ビットをすべて 1 に置いて得られるベクトル $(\in \{0,1\}^n)$ を $a^{\mathbf{1}}$, すべて 0 に置いて得られるベクトル $(\in \{0,1\}^n)$ を $a^{\mathbf{0}}$ と記す.

定理 3 *pBmb* $(\tilde{T},\tilde{F})$ がクラス $\mathcal{C}_{ALL}$ においてロバスト拡大をもつ必要十分条件は, $a \approx b$ をみたす $a \in \tilde{T}$ と $b \in \tilde{F}$ が存在しないことである. また, この条件は $O(n|\tilde{T}||\tilde{F}|)$ 時間で判定可能である. □

次の補題は, 正関数のクラス $\mathcal{C}_+$ およびその部分クラスについて成り立つ性質である.

補題 4 *pBmb* $(\tilde{T},\tilde{F})$ が関数のクラス $\mathcal{C} \subseteq \mathcal{C}_+$ においてロバスト拡大をもつ必要十分条件は

$$T^- = \{a^{\mathbf{0}} \mid a \in \tilde{T}\}, \quad F^+ = \{b^{\mathbf{1}} \mid b \in \tilde{F}\}$$

によって定義される *pdBf* (T^-,F^+) が, クラス $\mathcal{C}$ において拡大をもつことである.

証明: まず, $(\tilde{T},\tilde{F})$ のロバスト拡大 $f \in \mathcal{C}$ が存在したとする. ロバスト拡大の定義によって, これは一つの割り当てから得られる pdBf (T^-,F^+) の拡大でもある.

逆方向を示すために, pdBf (T^-,F^+) が拡大 $g \in \mathcal{C}$ を持つとしよう. すると任意の $a \in \tilde{T}$ と割り当て $\beta : AS(a) \to \{0,1\}$ に対し $a^{\mathbf{0}} \le a^\beta$ であるので, $1 = g(a^{\mathbf{0}}) \le g(a^\beta)$ から $g(a^\beta) = 1$ が導かれる. 同様に, 任意の $b \in \tilde{F}$ と $\beta : AS(b) \to \{0,1\}$ に対し $g(b^\beta) = 0$ が得られる. これは $g \in \mathcal{C}$ が $(\tilde{T},\tilde{F})$ のロバスト拡大であることを意味する. □

例 3 つぎの pBmb $(\tilde{T},\tilde{F})$ を考えよう. ただし $n = 5$ である.

$$\tilde{T} = \{(01**0),(*1011)\},\ \tilde{F} = \{(**101),(0*1*1)\}.$$

容易に分かるように, どの $a \in \tilde{T}$ と $b \in \tilde{F}$ をとっても $a \approx b$ とはできないので, 定理 3 によってクラ

ス $\mathcal{C}_{ALL}$ に対してロバスト拡大 f が存在する．たとえば，

$$\begin{aligned} T(f) &= \{(01000),(01010),(01100),(01110),\\ &\quad (01011),(11011)\}\\ F(f) &= \{0,1\}^5 - T(f). \end{aligned}$$

によってそのような f を定義できる．

しかし，補題 4 の pdBf (T^-, F^+) を作ると

$$T^- = \{(01000),(01011)\},\ F^+ = \{(11101),(01111)\}$$

であるが，たとえば $a = (01000) \in T^-$ と $b = (11101) \in F^+$ は $a \leq b$ をみたすので，(T^-, F^+) は正関数のクラス $\mathcal{C}_+$ において拡大を持たない．したがって，$\mathcal{C}_+$ において $(\tilde{T}, \tilde{F})$ のロバスト拡大は存在しない． □

整合拡大に対しても上と類似の性質が成り立ち、その結果、次の定理がいえる。

定理 5 クラス $\mathcal{C} = \mathcal{C}_+, \mathcal{C}^+_{k\text{-}DNF}$ に対し，問題 RE($\mathcal{C}$) および CE($\mathcal{C}$) は多項式時間で解くことができる．□

多項式時間で解けるもう一つの場合として，ホーン関数に関する次の結果がある．

定理 6 問題 RE($\mathcal{C}_{HORN}$) は多項式時間で解くことができる．

証明： pBmb $(\tilde{T}, \tilde{F})$ を考え，$a \in \tilde{T}$ に対し

$$\tilde{F}_{\succeq a} = \{b \in \tilde{F} \mid b \succeq a\}$$

と定める．以下，$(\tilde{T}, \tilde{F})$ が $\mathcal{C}_{HORN}$ におけるロバスト拡大をもつための必要十分条件が，$\tilde{F}_{\succeq a} \neq \emptyset$ をみたす各 $a \in \tilde{T}$ に対し，

$$\begin{aligned} &a_j = 0 \text{ かつすべての } b \in \tilde{F}_{\succeq a} \text{に対し}\\ &b_j = 1 \text{ をみたす添字 } j \text{が存在する，} \end{aligned} \quad (3)$$

ことであることを示そう． この条件は明らかに $O(n|\tilde{T}||\tilde{F}|)$ 時間でチェックできるので，定理の証明が得られることになる．まず，各 $a \in \tilde{T}$ に対し条件 (3) の j の存在を仮定する．すなわち，任意の割り当て $\alpha : AS \to \{0,1\}$ において $a^\alpha_j = 0$ であり，しかもすべての $b \in \tilde{F}_{\succeq a}$ は $b^\alpha_j = 1$ をみたす．したがって，この j を用いてホーン項

$$t_a = (\bigwedge_{i \in ON(a)} x_i)\bar{x}_j$$

を作ると，すべての α に対し $t_a(a^\alpha) = 1$ が成立し，さらにすべての $b \in \tilde{F}_{\succeq a}$ は $t_a(b^\alpha) = 0$ をみたす．さらに，$b \in \tilde{F} - \tilde{F}_{\succeq a}$ については，$b \not\succeq a$ であることから，$b_i = 0$ をみたす $i \in ON(a)$ が存在し，やはり $t_a(b^\alpha) = 0$ である．したがって，ホーン DNF

$$\varphi = \bigvee_{a \in \tilde{T}} t_a \quad (4)$$

を作ると，これが表す関数は $(\tilde{T}, \tilde{F})$ のロバスト拡大となっている．

一方，$\tilde{F}_{\succeq a} \neq \emptyset$ をみたすある $a \in \tilde{T}$ に対し，条件 (3) が成立しないならば，割り当て $\alpha : AS \to \{0,1\}$ を次のように定める．ただし，$(a,i) \in AS(a)$ および $(b,i) \in AS(\tilde{F}_{\succeq a})$ に対しては

$$\alpha(a,i) = \begin{cases} \bigwedge_{b \in \tilde{F}_{\succeq a} s.t. b_i \neq *} b_i, & b_i \neq * \text{をみたす} \\ & b \in \tilde{F}_{\succeq a} \text{が存在する場合} \\ 1, & \text{その他の場合} \end{cases}$$

$$\alpha(b,i) = a^\alpha_i$$

をみたすが，他の $(v,i) \in AS$ については任意．こうすると，$(\tilde{F}^\alpha)_{\geq a^\alpha} = \{b^\alpha \in \tilde{F}^\alpha \mid b^\alpha \geq a^\alpha\}$ に対し

$$a^\alpha = \bigwedge_{b^\alpha \in (\tilde{F}^\alpha)_{\geq a^\alpha}} b^\alpha$$

が成立する．これは，割り当て α によって得られた pdBf $(\tilde{T}^\alpha, \tilde{F}^\alpha)$ がホーン拡大を持たないことを意味する．したがって，$(\tilde{T}, \tilde{F})$ は $\mathcal{C}_{HORN}$ におけるロバスト拡大を持たない． □

例 4 先の例 3 で用いた pBmb $(\tilde{T}, \tilde{F})$ を考えよう．$\tilde{T}$ の 2 つのベクトル a に対し $\tilde{F}_{\succeq a}$ を求めると

$$\begin{aligned} \tilde{F}_{\succeq(01**0)} &= \{(**101),(0*1*1)\}\\ \tilde{F}_{\succeq(*1011)} &= \{(0*1*1)\} \end{aligned}$$

を得るが，$(01**0)$ については $j = 5$, $(*1011)$ については $j = 3$ が上記の条件 (3) をみたしている．したがって，ホーン DNF (4)

$$\varphi = x_2\bar{x}_5 \vee x_2x_4x_5\bar{x}_3$$

を作ると，定理 6 の証明にあるように，$(\tilde{T}, \tilde{F})$ のロバスト拡大が得られる． □

なお，定理 3, 5, 6 以外の場合については，第 1 節で導入したクラスのすべてにおいて RE, CE, MRE は NP 困難 (問題によっては coNP-困難) であることが分かっている．それらの証明は [5] にある．

表 2: 計算の複雑さの結果のまとめ.

	BEST-FIT	RE	CE	MRE
C_{ALL}	P	P	NPH	NPH
C_{+}	P	P	P	NPH
C_{HORN}	NPH	P	NPH	NPH
$C_{k\text{-}DNF}$	NPH	NPH	NPH	NPH

P: 多項式時間, NPH: NP-困難 (あるいは, coNP-困難).

5 むすび

表 2 に, 本論文で扱ったさまざまな問題の計算の複雑さをまとめておく. 表からも明らかなように, この分野の多くの問題は NP 困難であり, このような問題を実用的に解くには厳密アルゴリズムではなく, 近似アルゴリズムによらざるを得ない. 我々も, この方向の努力を続けているが ([3] など), より大きな発展が望まれている.

なお, 本研究の一部は, 文部省科学研究費の重点領域研究および国際学術研究などから援助をうけて行ったものである.

参考文献

[1] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, *International Conference on Management of Data (SIGMOD 93)*, (1993) 207-216.

[2] D. Angluin, Queries and concept learning, *Machine Learning*, 2 (1988) 319-342.

[3] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz and I. Muchnik, An implementation of logical analysis of data, RUTCOR Research Report RRR 22-96, Rutgers University, 1996.

[4] E. Boros, T. Ibaraki and K. Makino, Error-free and best-fit extensions of a partially defined Boolean function, *Information and Computation*, 140 (1998) 254-283.

[5] E. Boros, T. Ibaraki and K. Makino, Extensions of partially defined Boolean functions with missing data, RUTCOR Research Report RRR 6-96, Rutgers University, 1996.

[6] E. Boros, T. Ibaraki and K. Makino, Boolean analysis of incomplete examples, in *Algorithm Theory - SWAT'96*, edited by R. Karlsson and A. Lingas, Springer Lecture Notes in Computer Science, 1097 (1996) 440-451.

[7] Y. Crama, P. L. Hammer and T. Ibaraki, Cause-effect relationships and partially defined boolean functions, *Annals of Operations Research*, 16 (1988) 299-326.

[8] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, The MIT Press, 1996.

[9] L. R. Ford and D. R. Fulkerson, *Flows in Networks*, Princeton University Press, 1962.

[10] R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley-Interscience. 1997 (Second Edition).

[11] A. Horn, On sentences which are true of direct unions of algebras, *Journal of Symbolic Logic*, 16 (1951) 14-21.

[12] 茨木俊秀、データの論理的解析とブール関数、離散構造とアルゴリズム V (藤重悟編)、第 4 章、近代科学社、1998.

[13] H. Mannila, H. Toivonen and A.I. Verkamo, Efficient algorithms for discovering association rules, *AAAI Workshop on Knowledge Discovery in Database*, edited by U.M. Fayyad and R. Uthurusamy, (1994) 181-192.

[14] J. C. C. McKinsey, The decision problem for some classes of sentences without quantifiers, *Journal of Sympbolic Logic*, 8 (1943) 61-76.

[15] L. G. Valiant, A theory of the learnable, *Communications of the ACM*, 27 (1984) 1134-1142.