数理解析研究所講究録 1325 巻 2003 年 93-97

分解可能なしきい関数を用いたデータ解析

片岡 博幸 (Hiroyuki Kataoka)* 小野 廣隆 (Hirotaka Ono)†
山下 雅史 (Masafumi Yamashita)†

1 はじめに

大量のデータから意味のある情報を関数として取り出すことは様々な分野において現れる問題であり, 知識発見 [4, 12], データマイニング [1, 9] などの基本的なテーマであると言える. このような問題において, 抽出される関数が合成関数として表されるかどうかを知ることはデータの表す情報の論理構造を知る上で有益である. これは, ある情報を表現する関数が合成関数であるならば, その情報は複数の情報へと分解できることを意味するからである.

本稿では, ある現象が起こったことを表す n 次元 0-1 ベクトルの正データ集合と, 起こらなかったことを表す負データ集合が与えられたとき, 正データである x に対しては $f(x)=1$, 負データ x に対しては $f(x)=0$ を満たす論理関数 f が以下で定義されるある合成関数としての性質を満足するかどうかを判定する問題の計算時間を示す.

2 準備

ある現象が起こったことを表す正データベクトル集合 $T \subseteq \{0,1\}^n$ と起こらなかったことを表す負データベクトル集合 $F \subseteq \{0,1\}^n$ の対 (T,F) を部分定義論理関数 (partially defined Boolean Function; pdBf) と呼ぶ. ただし, ここでは $T \cap F = \emptyset$ を仮定する. (完全定義) 論理関数 $f : \{0,1\}^n \to \{0,1\}$ が条件 $a \in T \Rightarrow f(a)=1, b \in F \Rightarrow f(b)=0$ を満すとき, (T,F) の拡大 (extension) であるという. 部分定義論理関数 (T,F) に対する拡大 f は, 対象とする現象が起こった／起こらないを判別するため, 拡大 f の発見は知識獲得の一形態ととらえることができる. このような知識獲得の手法はデータの論理的解析 (Logical Analysis of Data; LAD) と呼ばれ, 研究されている [7, 3, 8].

データの論理的解析において一般に拡大 f は多数存在する. このため拡大発見の際, 関数のクラス $\mathcal{C}$ を指定することが多い. ここで関数のクラスとして, データの持つ性質, もしくは得たい知識の形式などから, 正関数, ホーン関数, しきい関数, 分解可能関数など, 様々なものが考えられる.

n 変数論理関数 f が, 任意の $x,y \in \{0,1\}^n$ に対し, $x \leq y \Rightarrow f(x) \leq f(y)$ を満すとき正関数であると言い, ある定数 θ と n 次元 (行) ベクトル w が存在して, $w \cdot x^T \geq \theta \Leftrightarrow f(x)=1$, $w \cdot x^T < \theta \Leftrightarrow f(x)=0$ であるときしきい関数であるという [11]. (T,F) が与えられたとき, 正関数, しきい関数に属する拡大 f の発見は多項式時間で可能であることが知られている [3].

変数 x_i の添字集合 $\{1,2,\ldots,n\}$ に対し, その部分集合 $S_0, S_1, S_2, \ldots, S_k$ を考える ($S_i \cap S_j = \emptyset$ である必要はない). 変数ベクトル x の S_i への射影を $x|_{S_i} = (x_j | j \in S_i)$ で表し, h_i $(i=1,2,\ldots,k)$ を変数 $x|_{S_i}$ をもつ論理関数 $\{0,1\}^{|S_i|} \to \{0,1\}$, さらに g を $x|_{S_0}$ と $y_1, y_2, \ldots, y_k$ を変数とする論理関数 $\{0,1\}^{|S_0|+k} \to \{0,1\}$ とした時, 拡大 f が合成関数

$$f(x) = g(x|_{S_0}, h_1(x|_{S_1}), \ldots, h_k(x|_{S_k}))$$

と書けるならば, f は分解可能であるという. また, そのような関数全体をスキーム

$$C_A(S_0, C_A(S_1), \ldots, C_A(S_k))$$

によって表わす. ここで, 記号 C_A は g や h_i として任意の論理関数が許されることを表わしており, g や h_i

*九州大学 大学院システム情報科学府情報工学専攻 (Department of Computer Science and Communication Engineering,Graduate School of Information Science and Electrical Engineering, Kyushu University), hiroyuki@tcslab.csce.kyushu-u.ac.jp
†九州大学 大学院システム情報科学研究院情報工学部門 (Department of Computer Science and Communication Engineering,Graduate School of Information Science and Electrical Engineering, Kyushu University), {ono, mak}@csce.kyushu-u.ac.jp

が正関数あるいはしきい関数から選ばれる場合には, それぞれそのことを C_P あるいは C_{TH} で表わすことにする. 例えば, g がしきい関数, $h_i(i=1,2,...,k)$ が正関数に制約されている場合は,

$$C_{TH}(S_0, C_P(S_1), ..., C_P(S_k))$$

と表わす. (S_0, S_1) が与えられたとき, (T,F) に対する分解可能な拡大の存在性判定(発見)については, スキーム $C_A(S_0, C_A(S_1))$, $C_P(S_0, C_P(S_1))$ では多項式時間アルゴリズムが存在すること, また $C_A(S_0, C_A(S_1), C_A(S_2))$ では NP 完全(困難)であることが知られている [8, 2].

本論文では以上を踏まえ, pdBf(T,F) と変数の 2 分割 (S_0, S_1) が与えられたとき, (T,F) の拡大 f で以下のような分解可能スキームに属するものが存在するかどうかを判定する問題について計算時間を示す. なお, Makino たちは正関数とホーン関数について同様の考察を行っている [10].

スキーム	計算量
$C_{TH}(C_A(S_0), C_A(S_1))$	$O(m^2n)$
$C_{TH}(C_P(S_0), C_P(S_1))$	$O(m^2n)$
$C_{TH}(C_A(S_0), C_P(S_1))$	$O(m^2n)$
$C_A(C_{TH}(S_0), C_{TH}(S_1))$	NP-complete
$C_P(C_{TH}(S_0), C_{TH}(S_1))$	NP-complete
$C_{TH}(C_{TH}(S_0), C_{TH}(S_1))$	NP-complete
$C_P(S_0, C_{TH}(S_1))$	$O(m^2n + LP(\|S_1\|, m))$

$m = |T| + |F|, n$: 変数の数,
$LP(a,b)$: a 変数, b 制約式の線形計画問題の解を求める時間

以下の節では上記の結果について説明する.

3 $C_{TH}(C_A(S_0), C_A(S_1))$ $C_{TH}(C_P(S_0), C_P(S_1))$ $C_{TH}(C_A(S_0), C_P(S_1))$

定理 1 pdBf(T,F) に対する, 以下の分解可能スキームに属する拡大の存在性は $O(m^2n)$ で判定できる.

- $C_{TH}(C_A(S_0), C_A(S_1))$
- $C_{TH}(C_P(S_0), C_P(S_1))$
- $C_{TH}(C_A(S_0), C_P(S_1))$ □

ここでは紙面の都合上, $C_{TH}(C_A(S_0), C_P(S_1))$ に関する証明のみ紹介する.
略証 $a \in T, b \in F$ に対して,

- 全てが $a|_{S_0} \neq b|_{S_0}$ のとき

  拡大は常に存在. なぜなら, $h_1(a|_{S_1})$, $h_1(b|_{S_1})$ がとり得る値に関係なく, $h_0(a|_{S_0}) = 1$, $h_0(b|_{S_0}) = 0$ としてやれば他の制約を全て満たす.

- 少なくとも 1 つは,$a|_{S_0} = b|_{S_0}$ のとき,

  h_1' は $a|_{S_1} < b|_{S_1}$ ならば $(h_1'(a|_{S_1}), h_1'(b|_{S_1})) = (0,1)$, $a|_{S_1} > b|_{S_1}$ ならば $(h_1'(a|_{S_1}), h_1'(b|_{S_1})) = (1,0)$ とし, この操作と正関数の条件より決定する値も各ベクトルに与える関数とする. このとき, 拡大が存在する必要十分条件は $c, d \in T \cup F$, $c \leq d$ ならば,$h_1'(c|_{S_1}) \leq h_1'(d|_{S_1})$ が成立し, かつ $a, a' \in T$, $b, b' \in F$ に対して, $h_1'(a|_{S_1}) < h_1'(b|_{S_1})$ と $h_1'(a'|_{S_1}) > h_1'(b'|_{S_1})$ が同時に成立しないことである. 前半部分は正関数の制約であり, 後半部分はしきい関数の制約を満たすためである.

これらの判定は各 T と F のベクトル対に対する比較により行なうことができるため, $C_{TH}(C_A(S_0), C_P(S_1))$ に属する拡大の存在性判定は $O(m^2n)$ の計算時間で行なうことができる. □

4 $C_A(C_{TH}(S_0), C_{TH}(S_1))$ $C_P(C_{TH}(S_0), C_{TH}(S_1))$ $C_{TH}(C_{TH}(S_0), C_{TH}(S_1))$

定理 2 pdBf(T,F) に対する, 以下の分解構造スキームに属する拡大の存在性判定は NP-完全である.

- $C_A(C_{TH}(S_0), C_{TH}(S_1))$
- $C_P(C_{TH}(S_0), C_{TH}(S_1))$
- $C_{TH}(C_{TH}(S_0), C_{TH}(S_1))$ □

証明. ここでは紙面の都合上, $C_{TH}(C_{TH}(S_0), C_{TH}(S_1))$ に対する証明のみを紹介する.

ハイパーグラフの2彩色問題[5]から帰着することにより証明する. $V=\{1,2,...,n\}$ 上の3次ハイパーグラフ $\mathcal{H}=(V,E)$ に対して, $T,F\subseteq\{0,1\}^{6V}$ を次のように取る.

$$\begin{aligned} T &= \{u_e^{(1)},u_e^{(2)},u_e^{(3)} \mid e\in E\} \\ F &= \{v_e^{(1)},v_e^{(2)},v_e^{(3)} \mid e\in E\}. \end{aligned}$$

ただし $u_e^{(i)},v_e^{(i)}$, $i=1,2,3$ は $e=\{p,q,r\}$ に対して,

$$\begin{aligned} u_e^{(1)} &= (x^{\{3p,3q,3r\}},y^{\{3p',3q',3r'\}}), \\ u_e^{(2)} &= (x^{\{3p+1,3q+1,3r+1\}},y^{\{3p'+1,3q'+1,3r'+1\}}), \\ u_e^{(3)} &= (x^{\{3p+2,3q+2,3r+2\}},y^{\{3p'+2,3q'+2,3r'+2\}}), \\ v_e^{(1)} &= (x^{\{3p,3p+1,3p+2\}},y^{\{3p',3p'+1,3p'+2\}}), \\ v_e^{(2)} &= (x^{\{3q,3q+1,3r+2\}},y^{\{3q',3q'+1,3q'+2\}}), \\ v_e^{(3)} &= (x^{\{3r,3r+1,3r+2\}},y^{\{3r',3r'+1,3r'+2\}}) \end{aligned}$$

と定義される. ここで, $(x,y)=(a|_{S_0},a|_{S_1})$, $a\in T\cup F$, $S_0\cap S_1=\emptyset$ であり, x^W は変数の添字集合 W に対し, $x_j=1$ $(j\in W)$, $x_j=0$ $(j\notin W)$ なるベクトルを表わす. 今, $C_{TH}(C_{TH}(S_0),C_{TH}(S_1))$ に属する論理関数 $f(x)=g(h_0(x),h_1(x))$ を考える. g は2変数論理関数であり, 定義域に相当するベクトル集合 $\{0,1\}^2$ の正負 (これを T',F' とする) に分ける組合せは16通りである.

まず $(T',F')=(\{(1,1)\},\{(0,0),(0,1),(1,0)\})$ について議論をする. (T,F) が $C_{TH}(C_{TH}(S_0),C_{TH}(S_1))$ の拡大を持つと仮定すると $T'=\{(1,1)\}$ であるから,

$$\begin{aligned} w_{3p}+w_{3q}+w_{3r} &\geq \theta \\ w_{3p+1}+w_{3q+1}+w_{3r+1} &\geq \theta \\ w_{3p+2}+w_{3q+2}+w_{3r+2} &\geq \theta \end{aligned}$$

が成り立つ. これより必ず $(h_0(v_e^{(1)}|_{S_0}),h_1(v_e^{(1)}|_{S_1}))$, $(h_0(v_e^{(2)}|_{S_0}),h_1(v_e^{(2)}|_{S_1}))$, $(h_0(v_e^{(3)}|_{S_0}),h_1(v_e^{(3)}|_{S_1}))$ のいずれかは $(1,0)$ の値を持つ. なぜならば, $(0,1),(0,0)$ のみであるとすると,

$$\begin{aligned} w_{3p}+w_{3p+1}+w_{3p+2} &< \theta \\ w_{3q}+w_{3q+1}+w_{3q+2} &< \theta \\ w_{3r}+w_{3r+1}+w_{3r+2} &< \theta \end{aligned}$$

が成立し,

$$\begin{aligned} w_{3p}+w_{3p+1}+w_{3p+2}+w_{3q}+w_{3q+1} \\ +w_{3q+2}+w_{3r}+w_{3r+1}+w_{3r+2} &\geq 3\theta \end{aligned}$$

と

$$\begin{aligned} w_{3p}+w_{3p+1}+w_{3p+2}+w_{3q}+w_{3q+1} \\ +w_{3q+2}+w_{3r}+w_{3r+1}+w_{3r+2} &< 3\theta \end{aligned}$$

が同時に成り立ち矛盾が生ずるためである. 同様に

$$\begin{aligned} w_{3p'}+w_{3q'}+w_{3r'} &\geq \theta \\ w_{3p'+1}+w_{3q'+1}+w_{3r'+1} &\geq \theta \\ w_{3p'+2}+w_{3q'+2}+w_{3r'+2} &\geq \theta \end{aligned}$$

が成立し, これにより $(h_0(v_e^{(1)}|_{S_0}),h_1(v_e^{(1)}|_{S_1}))$, $(h_0(v_e^{(2)}|_{S_0}),h_1(v_e^{(2)}|_{S_1}))$, $(h_0(v_e^{(3)}|_{S_0}),h_1(v_e^{(3)}|_{S_1}))$ のいずれかは $(0,1)$ の値を取る. すなわち,

$$\begin{aligned} \overline{h_0(v_e^{(1)}|_{S_0})}\vee\overline{h_0(v_e^{(2)}|_{S_0})}\vee\overline{h_0(v_e^{(3)}|_{S_0})} &= 1 \\ h_0(v_e^{(1)}|_{S_0})\vee h_0(v_e^{(2)}|_{S_0})\vee h_0(v_e^{(3)}|_{S_0}) &= 1 \end{aligned}$$

が成立し, $h_0(v_e^{(1)}|_{S_0})$, $h_0(v_e^{(2)}|_{S_0})$, $h_0(v_e^{(3)}|_{S_0})$ をそれぞれ p,q,r への塗り分けと見なすと, 各 $e=\{p,q,r\}$ は $0,1$ の2色を含む. すなわち, ハイパーグラフは2色に塗り分けられる.

次にハイパーグラフが2色で塗り分けられると仮定する. まず, $e=\{p,q,r\}$ に対し, r が他の2点と異なる色が塗られているとする. 今, $(v_e^{(1)}|_{S_0},v_e^{(2)}|_{S_0},v_e^{(3)}|_{S_0})=(p_1,q_1,r_1)$ とすると,

$$\begin{pmatrix} h_0(p_1) \\ h_0(q_1) \\ h_0(r_1) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} w_{3p} & w_{3p+1} & w_{3p+2} \\ w_{3q} & w_{3q+1} & w_{3q+2} \\ w_{3r} & w_{3r+1} & w_{3r+2} \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \theta=2$$

または,

$$\begin{pmatrix} h_0(p_1) \\ h_0(q_1) \\ h_0(r_1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} w_{3p} & w_{3p+1} & w_{3p+2} \\ w_{3q} & w_{3q+1} & w_{3q+2} \\ w_{3r} & w_{3r+1} & w_{3r+2} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2 & 2 & 2 \end{pmatrix}, \theta = 2$$

とすることにより, 確かに拡大を持つ. これは p,q,r の対称性より p,q,r が入れ替わっても問題はない. よって, (T,F) は $C_{TH}(C_{TH}(S_0), C_{TH}(S_1))$ の拡大を持つ. 同様の証明により他の16通りの内の7つの場合においてNP-完全性が証明できる.

また, 16通りの内, 残りの8通りは明らかに拡大を持たない. なぜなら, それらは任意の $a \in T$, $b \in F$ に対して,

- $(h_1(a|_{S_1}), h_1(b|_{S_1})) = (0,1)$
- $(h_0(a|_{S_0}), h_0(b|_{S_0})) = (0,1)$
- $(h_0(a|_{S_0}), h_0(b|_{S_0})) = (1,0)$
- $(h_1(a|_{S_1}), h_1(b|_{S_1})) = (1,0)$
- $0 > \theta, w_{h_0} + w_{h_1} > \theta, w_{h_0} < \theta, w_{h_1} < \theta$
- $w_{h_0} < \theta, w_{h_1} < \theta, 0 < \theta, w_{h_0} + w_{h_1} < \theta$
- $T = \emptyset$
- $F = \emptyset$

のどれか一つが成立条件であるが, 上記の条件が成立しないことは明らかである. 16通り以外の各要素が4個より少なくなる場合 (例えば, $(T',F') = (\{(1,1)\},\{(1,0),(0,1)\}), (\{(1,1)\},\{(0,0)\})$) を想定する必要もあるが, それらは全て16通りのどれかに含まれる. □

5 $C_P(S_0, C_{TH}(S_1))$

定理 3 pdBf(T,F) に対する, 以下の分解構造スキームに属する拡大の存在性判定は $O(m^2n + LP(|S_1|, m))$ の計算時間, すなわち多項式時間で行うことができる.

- $C_P(S_0, C_{TH}(S_1))$ □

証明. 正拡大 f が存在するための必要十分条件は $a \leq b$ を満たす $a \in T, b \in F$ が存在しないことである. したがって, 拡大 $g(S_0, h_1(S_1))$ において g が正関数であることの必要十分条件は $a \in T, b \in F, a|_{S_0} < b|_{S_0}$ を満す任意の a,b に対して, $h_1(a|_{S_1}) > h_1(b|_{S_1})$ が成立することである. 言い換えると, 制約条件が $w \cdot a^T > \theta$ および $w \cdot b^T < \theta$ である線形計画問題が解を持つことが必要十分である. よって, 全ての $a \in T$ と $b \in F$ のペアについて比較を行い, その上で線形計画問題を解くことで本分解構造スキームに属する拡大を求めることができる. ここで得られる線形計画問題は $|S_1|$ 変数, m 制約式からなるため, 定理は証明できた. □

6 考察と今後の課題

本論文では, (T,F) が与えられたとき, 分解可能なしきい関数に関連した判別関数の存在判定にかかる計算時間を示した. 分解可能性を示す情報はデータ集合における属性間の関係, 階層構造を表すことができ, 知識発見の観点から有益である. また, 一方しきい関数も多くの現実的な場面で現れる関数である (分解可能なしきい関数の単純な応用としては, 政策に対する政党ごとの賛成, 反対が法案の可決, 否決などが考えられる [6]). 残念ながら, 今回とり上げた問題のいくつかはNP-完全のクラスに属し, それらを実用的に解くには, 近似あるいはヒューリスティックなアルゴリズムを考慮していく必要がある. また, 今回の結果も, 全ての論理関数, 正関数, しきい関数の全ての組合せについては明らかにしておらず, まずこれらを解決する必要がある.

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,* 93, 207-216, 1993.

[2] E. Boros, V. Gurvich, P. L. Hammer, T. Ibaraki and A. Kogan, "Decompositions of partially defined Boolean functions," *Discrete Applied Mathematics,* 62, 51-75, 1995.

[3] Y. Crama, P.L. Hammer, and T. Ibaraki , "Cause-effect relationships and partially defined

Boolean functions," *Annals of Operations Research,* 16, 299-325, 1988.

[4] R. Dechter and J. Leiserson, "Structure identification in relational data," *Artificial Intelligence,* 58, 237-270, 1992.

[5] M. R. Garey, D. S. Johnson, *Computers and Intractability,* Bell Telephone Laboratories, Incorporated, 1979.

[6] 船木 由喜彦, "エコノミックゲームセオリー," サイエンス社, 43-67, 2001.

[7] P.L Hammer, "Partially defined Boolean functions and cause-effect relation-ships," *In Proceedings of the International Conference on Multi-Attribute Decision Making via OR-Based Expert Systems,*,University of Passau, Passau, Germany, 1986.

[8] 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, 近代科学社, 147-202, 1998.

[9] H. Mannila, H. Toivonen, and A.I. Verkamo, "Efficient algorithms for discovering association rules," *AAAI Workshop on Knowledge Discovery in Database,*, 181-192, 1994.

[10] K. Makino, K. Yano, and T. Ibaraki, "Positive and horn decomposability of partially defined boolean functions," *Discrete Applied Mathematics,* 74, 251-274, 1997.

[11] 室賀三朗, 茨木俊秀, 北橋忠宏, "しきい論理," 産業図書, 1971.

[12] J. R. Quinlan, "Induction of decision trees," *Machine Learning,* 1, 81-106, 1986.