1999年度日本オペレーションズ・リサーチ学会
春季研究発表会

データの論理的解析における階層的分解構造について

02004684　京都大学　*小野 廣隆　ONO Hirotaka
02601514　大阪大学　牧野 和久　MAKINO Kazuhisa
01001374　京都大学　茨木 俊秀　IBARAKI Toshihide

1　はじめに

本研究では, 数値的データ集合として正例の集合 P と, 負例の集合 Nの対 (P,N) が与えられたとき (ただし, $P,N \subseteq \mathbb{R}^d, P \cap N = \emptyset$), 論理関数の分解可能性を利用して, これらの属性値の間に成り立つ階層構造を発見することを考える.

そのため, まず各属性ごとにいくつかのカット点を導入し, 数値データ集合対 (P,N) を 2 値データ集合対 (T,F) に変換する (ただし, $T,F \subseteq \{0,1\}^n, T \cap F = \emptyset$ である). (T,F) を部分定義論理関数 (*partially defined Boolean function*, pdBf) と呼び, pdBf (T,F) と矛盾しない完全定義論理関数 fを拡大 (extention) と呼ぶ.

拡大 fを求めることは, (T,F) から論理的な形で知識獲得を行なっていると見なすことができ, ひいては元のデータ集合 (P,N) の論理的解析の一形式と考えられる.

ここでは拡大 fが分解構造 $f = g(x[S_0], h(x[S_1]))$ を持つ場合 (このときスキーム $F_0(S_0, F_1(S_1))$ を持つ, という) に着目する. これまでの研究により (S_0,S_1) が与えられたとき, 部分定義論理関数 (T,F) の $F_0(S_0,F_1(S_1))$-分解可能性の判定と (拡大可能である場合) その拡大を求めることは, 多項式時間で可能であるが [2], $F_0(S_0,F_1(S_1))$-分解可能な S_0, S_1の存在判定問題は NP-完全である. また, S_0, S_1が与えられている場合でもエラー最小の拡大 (BEST-FIT 拡大) を求める問題は NP 困難であることが知られている [1]. 本研究では, 全変数集合の分割 (S_0,S_1) 全てに対して $F_0(S_0,F_1(S_1))$-分解可能な BEST-FIT 拡大を求める問題を近似解法を使用して解き, 分解可能性の判定結果にしたがって, 変数間の関係を階層構造としてとらえることを試みる.

さらに, このアプローチの有効性を見るため, 人為的データ例と実データ例に適用し, その結果を検討する.

2　定義

2.1　部分定義論理関数の BEST-FIT 拡大

完全定義論理関数 (以下では, 単に関数と呼ぶ) $f : \{0,1\}^n \mapsto \{0,1\}$ に対して, $f(v) = 1$ である $v \in \{0,1\}^n$を真ベクトル, $f(v) = 0$ である $v \in \{0,1\}^n$を偽ベクトルと呼ぶ. fの真ベクトル集合を $T(f)$, fの偽ベクトル集合を $F(f)$ と記す. pdBf (T,F) に対し fが $T(f) \supseteq T$, $F(f) \supseteq F$を満たすとき, fをその**拡大**という.

与えられた完全定義論理関数のクラス $\mathcal{C}$に対し次の問題を考える.

問題 EXTENSION($\mathcal{C}$)
入力: pdBf (T,F), ただし, $T,F \subseteq \{0,1\}^n$.
出力: (T,F) の拡大 $f \in \mathcal{C}$が存在すれば yes, 存在しなければ no.

pdBf (T,F) と (必ずしもその拡大ではない) 関数 fが与えられたとき, $f(v) = 1$ であるベクトル $v \in T$, および $f(w) = 0$ であるベクトル $w \in F$は fによって正しく分類されているという. 逆に $f(v) = 0$ である $v \in T$, $f(w) = 0$ であるベクトル $w \in F$を fの誤りベクトルと呼ぶ. pdBf (T,F) に対する拡大が存在しないとき. 誤りベクトルの重みの和が最小な拡大 (BEST-FIT 拡大) を求めることは極めて自然である.

問題 BEST-FIT($\mathcal{C}$)
入力: pdBf (T,F), 重み関数 $w : T \cup F \mapsto \mathbb{R}_+$.
出力: 部分集合 T^* と F^*. ただし, $T^* \cap F^* = \emptyset$, $T^* \cup F^* = T \cup F$, さらに, pdBf (T^*,F^*) は $\mathcal{C}$ において拡大をもち, $w(T^* \cap F) + w(F^* \cap T)$ を最小にする.

2.2　関数の分解可能性

fが $\mathcal{S} = \{S_i \mid S_i \subseteq S,\ i = 0,1,\cdots,k\}$ に対して $F_0(S_0,F_1(S_1),F_2(S_2),\cdots,F_k(S_k))$ -分解可能であるとは, 次の条件を満足する関数 g: $\{0,1\}^{|\{S_0\}|+k} \to \{0,1\}$, h_i: $\{0,1\}^{|\{S_i\}|} \to \{0,1\}$, $i = 1,2,\cdots,k$, が存在することである [1,2].

全ての $v \in \{0,1\}^n$に対して
$f(v) = g(v[S_0], h_1(v[S_1]), \cdots, h_k(v[S_k]))$.

以下ではとくに $\mathcal{C} = F_0(S_0,F_1(S_1))$-分解可能関数のクラスに関して考察を加えるが, このクラスに対する問題 BEST-FIT($\mathcal{C}$) は NP 困難であることが知られている [1].

2.3 カット点

数値データ集合対 (P,N) に対して, i 番目の属性がとる値の領域を $\mathbb{D}_i = \{u_i \mid u \in P \cup N\}$ と書く. i 番目の属性にカット点$\alpha_{ij}, j = 1,2,\cdots,k_i$ を導入し, 次の規則に従って数値 $u_i \in \mathbb{D}_i$ をベクトル $(x_{i1},\cdots,x_{ik_i}) \in \{0,1\}^{k_i}$ に変える:

$$x_{ij} = \begin{cases} 1 & u_i \geq \alpha_{ij} \text{ のとき} \\ 0 & u_i < \alpha_{ij} \text{ のとき.} \end{cases}$$

導入されるカット点集合が満たすべき条件として, 2 値化の結果 (P,N) から得られる pdBf (T,F) が対象とする関数のクラス $\mathcal{C}$ において拡大を持つことが求められる. しかし取りうる全てのカット点を導入するのは冗長であり, 実用的ではない. その結果導入するカット点集合を最小化する問題が考えられるが, この問題は, 集合被覆問題に定式化できる. 一般には NP 困難であるが, 近似解法として欲張り法等が有効である.

以上からわかるようにカット点集合の選択には幅があるが, どのようなカット点集合を選択するかによって, 得られる pdBf (T,F) は異なってくる.

3 分解可能関数によるデータ解析

以下の手順によって, 数値データの属性間に存在する分解構造の発見が可能となる. (i) 適当な手法によりカット点を導入することによりデータの 2 値化を行ない, (ii) その結果得られる pdBf (T,F) に対して, 可能な分割 (S_0,S_1) を全て考慮し, それぞれにおける拡大の存在を調べる. (この目的にはクラス $F(S_0,F_1(S_1))$ に対する BEST-FIT 拡大を求める近似アルゴリズムを利用する.)

上の (ii) によって, ある分割 (S_0,S_1) に対して分解構造 $g(S_0,h(S_1))$ を持つという結果が得られた場合, その $h(S_1)$ を新たな変数と見ることができる. この結果得られる新たな変数集合上への pdBf (T,F) の射影を pdBf (T',F') とし, 再び上記の手法を適用することにより変数間に存在する階層的な分解構造を把握することが可能となる.

ところで, 上記の h には一般には自由度がある. 例えば, ある分割 (S_0,S_1,S_2) に対して $h_2(S_2)$ を考えたとき, ある h_2に対しては pdBf (T',F') は $F(S_0,F_1(S_1 \cup h(S_2)))$-分解可能であるが, 別の h_2に対してはそうではない, という現象が起こりうる. 3 変数集合への分割 (S_0,S_1,S_2) の場合, pdBf (T,F) が $F(S_0,F_1(S_1,F_2(S_2)))$-分解可能であるとしても, それを実現する $h_2(S_2)$ の構成問題は NP 困難であることを示すことができる. ただし, (T,F) に単調性があり, g, h_1, h_2がともに正関数の形を取る場合はこの割り当ての決定は多項式で可能である.

4 数値実験

3 節で述べた手法を元に計算実験を行なった. (i) における 2 値化は欲張り法に基づくアルゴリズムを使用し, k 個のカット点の導入により実現している. ただし, k は必ずしも最小なものが適当とは限らないので, 最小値付近のいくつかの kに対して調べる.

ここでは使用したデータのうち, 特に実データ (乳癌の診断データ[1]) に対する適用について紹介する. 各データベクトルは 9 つの属性を持ち, 細胞の大きさや形状の均一性等の状態を 1 から 10 の整数値によって表わしている (すなわち, 9 次元 10 値ベクトルの集合である). データは悪性腫瘍患者集合 $|P| = 239$, 良性腫瘍患者集合 $|N| = 444$ の合計 683 個のベクトルから成っている. 変数の意味を下の表に示す.

各変数の意味		各変数の意味	
1:	(患部) 集合の大きさ	6:	裸の核
2:	細胞サイズの均一性	7:	柔染色体
3:	細胞の形の均一性	8:	正常な核
4:	縁の癒着度	9:	有糸分裂
5:	一つの上皮細胞サイズ		

このデータから 600 個のベクトルを 10 通り抽出したのち, カット点の導入, 分解可能性の調査を行なった. ただし, データに誤りがある可能性があることを考慮し, アルゴリズムには, BEST-FIT を求める近似解法を適用し, 誤りベクトル数が全データの 1 %以内, 6 個以内に収まっている場合は, 分解構造が存在する, と判断している.

この結果, 分解構造を持つ可能性が大きいと判定された変数集合の組が存在することが確認された. それらの変数集合の組 (S_0,S_1) において S_1に同時に出現する変数の組は下表の通りである.

変数の組	出現回数 (全 185 組中)
(5 , 9)	130 回
(7 , 9)	109 回
(5 , 7)	106 回
(3 , 9)	105 回
(2 , 9)	101 回

この実験結果はこれらの変数は組になって別の概念を生成している可能性があることを示唆していると考えられる. なお, 他の実験結果については当日発表する.

参考文献

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

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

[1] ftp.ics.uci.edu:pub/machine-learning-databases/breast-cancer-wisconsin