論理的データ解析における階層的分解構造について*

6AA－4

小野 廣隆　(京都大学大学院工学研究科) †
牧野 和久　(大阪大学大学院基礎工学研究科) ‡
茨木 俊秀　(京都大学大学院工学研究科) §

1　はじめに

本研究では，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，$P,N\subseteq\mathbf{R}^d, P\cap N\neq\emptyset$)，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関係を求めることを考える．

そのため，まず各属性ごとにいくつかのカットポイントを導入し，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，$T,F\subseteq\{0,1\}^n, T\cap F\neq\emptyset$である)．$(T,F)$を部分定義論理関数(*partially defined Boolean function*, pdBf)と呼ぶ．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$f=g(x[S_0],h(x[S_1]))$を持つ場合に着目し，このときスキーム$F_0(S_0,F_1(S_1))$を持つ，という．これまでの研究により部分定義論理関数(T,F)の$F-0(S_0,F_1(S_1))$-分解可能性と(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]，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]．従って本研究では$F_0(S_0,F_1(S_1))$-分解可能なBEST-FIT 拡大を求めるためには近似解法を使用する．これをすべてのS_0,S_1の組み合わせに対して適用するという手順を再帰的に実行すれば，全変数集合間の分解構造を階層的にとらえることが可能となる．

本研究では，このアプローチの有効性を見るため，実データ例に適用し，その結果を検討した．

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な問題である．

問題 BEST-FIT($\mathcal{C}$)
入力: pdBf (T,F)，重み関数$w:T\cup F\mapsto\boldsymbol{R}_+$．
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3つの条件を満足することである[1,2]．

(i)　全ての$v\in\{0,1\}^n$に対して
　　$f(v)=g(v[S_0],h_1(v[S_1]),\cdots,h_k(v[S_k]))$，
(ii)　各h_iはS_i上の変数のみに依存，
(iii)　g: $\{0,1\}^{|\{S_0\}|+k}\to\{0,1\}$．

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するBEST-FIT 拡大を利用するが，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　カットポイント

数値データ集合対(P,N)に対して，i番目の属性の領域を$\mathbf{D}_i=\{u_i\mid u\in P\cup N\}$と書く．データの$i$番目

*A hierarchial decomposability in logical data analysis
†Hirotaka Ono, Kyoto University
‡Kazuhisa Makino, Osaka University
§Toshihide Ibaraki, Kyoto University

の属性上にカットポイント$\alpha_{ij}, j = 1,2,\cdots,$を導入し, 次の規則に従って数値$u_i \in \mathbb{D}_i$を$\{0,1\}$の値$x_{ij}$に変える:

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

導入されるカットポイント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)が全関数のクラス$\mathcal{C} = \mathcal{C}_{ALL}$において拡大を持つこと(つまり$T \cap F = \emptyset$)が求められる. しかし取りうる全てのカットポイント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ポイント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ポイント集合の選択には幅があるが, どのようなカットポイント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 数値実験

3.1 実験の手順

§2.2では変数集合の部分集合S_0, S_1があらかじめ与えられた場合について分解可能関数の存在について述べた. 実際には, (S_0,S_1)の組は与えられているわけではなく, むしろ, 分解可能な(S_0,S_1)を探すことが重要である.

本研究では, カットポイント集合により実データ集合を2値化したのち, (S_0,S_1)の候補として, 変数集合の2分割をすべて列挙し, それぞれに対してBEST-FIT$(C_{F_0(S_0,F_1(S_1))})$を近似解法を用いて解くことによって分解可能性を調べた. ただし, カットポイント集合の選択には欲張り法, および, それに確率的な要素を加えた方法(被覆される要素数に乱数をかけて欲張り法を実行)を利用した.

3.2 実験のデータ

今回の実験では, 町ごとの住宅価格の査定用のデータ (ftp.ics.uci.edu:pub/machine-learning-databases/housing)を用いた. ただし, 各変数値が実数値をとることから, カットポイントを導入し, 2値化してから実験を行なっている.

このデータは14個の属性をもち, それぞれの意味は次の通りである.

1	犯罪率比(対市民)	8	雇用地への距離
2	居住面積の比	9	高速道路接続性
3	ビジネス街面積比	10	固定資産税率
4	ダミー変数	11	学校の先生数
5	NOX濃度	12	黒人の比率
6	町の平均部屋数	13	低所得者層数
7	築年数	14	住宅の価格

ここでは住宅価格が21万ドル以上のデータベクトルを正例, 21万ドル未満を負例とした. 一般にはカットポイント集合の選択によって, 得られるpdBf (T,F)は異なり, それによって分解構造も影響を受ける. しかし, 分解構造が存在するような(S_0,S_1)の組は, カットポイント集合の取り方によらないと予想される. この点を10通りのカットポイント集合の導入法(確率的な欲張り法による)によって得られたpdBf (T,F)の分解構造を調べることにより, 検討した.

3.3 実験のまとめ

以上の実験において, 分解構造を持つと判断された(S_0,S_1)の組(誤り率1%未満)において同じS_1に入っている変数の組の出現頻度の上位6個をとると次のようになった.

変数の組	出現頻度	変数の組	出現頻度
(1,5)	117回	(1,10)	94回
(1,8)	107回	(5,10)	89回
(5,8)	97回	(8,10)	82回

さらに$S_1 = \{1,5,8\}, \{1,5,10\}, \{1,8,10\}, \{5,8,10\}$の出現頻度は10通りのカットポイント集合の取り方中9回, $S_1 = \{1,5,8,10\}$の出現頻度は10回中6回, となっており, かなりの頻度で同じ分解構造が現れていることが分かる.

ここで変数間の関係を知るため, 例えば$S_1 = \{1,5,8\}$を$h(S_1)$という一つの変数として見て, 同様の操作を加えると(ただし誤差の影響を考えて誤り率2%未満まで考慮) $S_1' = \{S_1, 10\}$の出現頻度は9回中7回となった. これよりS_1に含まれている属性1,5,8,10はそれぞれ独立しては住宅の価格の決定に大きな影響を持たないが, まず属性1,5,8が一つの状態を区別し, さらに属性10が影響をおよぼす, というように, 町の住環境に関係する特徴が階層的に集まって全体として一つの概念を作り出している可能性がある.

参考文献

[1] E.Boros, T.Ibaraki, and K.Makino, Error-free and best-fit extensions of partially defined Boolean function, RUTCOR Research Report RRR 14-95, Rutgers University,1995(To appear in Information and Computation).

[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.