1996年度日本オペレーションズ・リサーチ学会
秋季研究発表会

# 数値データの論理的分析

京都大学　*須田 高史　SUDA Takashi
02601514　京都大学　牧野 和久　MAKINO Kazuhisa
01001374　京都大学　茨木 俊秀　IBARAKI Toshihide

## 1　はじめに

ある事象を引き起こす例 (正例) のデータ集合 $P$ と, 引き起こさない例 (負例) のデータ集合 $N$ の組である $(P,N)$ が与えられたとき, なぜその事象が起こるのかということを説明する理論 $g$ (正確に言うと後述の判別関数のことである) を求める問題を考える. 例えば, 乳がんのデータにおいては, $P$ が悪性腫瘍, $N$ が良性腫瘍の患者のデータ集合を示し, それぞれのデータは細胞の大きさや形状の均一性, 細胞核, 核仁, そして有糸分裂などを示す属性に対する値からなるベクトルである. 理論 $g$ は次の基準を満たすことが望ましい.

1. $g$ の表現が簡潔である.
2. 新しいデータ集合 $(P',N')$ に対する誤り確率が小さい.

基準 1 は,「正しい理論は簡潔である」という原則を表しており, 換言すれば, 複雑な理論は, 現在までに得られたデータの表面的な数値に依存したものであり, その内部に存在する本質的な特徴を捉えていないと考えていることを表す. 2 は, 乳がんの例で考えると, 現在までに得られたデータによる理論を新しい患者に適用し, 正しい判断をすることが実際に役立つわけであるから, 当然求められる基準である.

## 2　カットポイントと決定木

データ集合の組 $(P,N)$ を考える. ただし, $P,N \subseteq \mathbb{R}^m$ とする. このとき, すべての $v \in P$ に対し $g(v) > 0$ を満たし, かつすべての $w \in N$ に対し $g(w) \leq 0$ を満たす関数 $g$ を判別関数と呼ぶ. $i$ 番目の属性の領域を $\mathbb{D}_i = \{u_i \mid u \in P \cup N\}$ と書く. このデータから論理的説明を得るため, $i$ 番目の属性上に $n_i$ 個のカットポイント $\alpha_{ij}$ を導入し, 次の規則に従って数値 $u_i \in \mathbb{D}_i$ を $n_i$ 個の $\{0,1\}$ の値 $x_{ij}$ $(j=1,2,\ldots,n_i)$ に変える:

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

この過程を二値化という. ここで $\mathbb{D}_i = \{u_i^{(0)}, u_i^{(1)}, \ldots, u_i^{(n_i)}\}$ であるとし, $u_i^{(0)} > u_i^{(1)} > \cdots > u_i^{(n_i)}$ を仮定する. $i$ 番目の属性に対するカットポイントの最大集合は, 各 $j=1,2,\ldots,n_i$ に対して $u_i^{(j-1)}$ と $u_i^{(j)}$ の間に 1 つのカットポイント $\alpha_{ij}$ を導入したときに得られる. このようにして得られる各属性の二値ベクトルをつなぎ合わせることによって, $P,N$ はそれぞれ $\{0,1\}^n$ の部分集合 $T,F$ に変換される (ただし $n = \sum_i n_i$). 一般に $T,F \subseteq \{0,1\}^n$ であるとき, $(T,F)$ を部分定義論理関数 (pdBf) と呼ぶ. pdBf $(T,F)$ に対し, $T \subseteq T(f)$ かつ $F \subseteq F(f)$ を満たす論理関数 $f$ (つまり, $(T,F)$ の判別関数) を $(T,F)$ の拡大と呼ぶ. 上述のようにカットポイントの最大集合を導入した結果得られる pdBf $(T^*,F^*)$ を, $(P,N)$ の master pdBf と呼ぶ.

論理関数は, 決定木によって表現できる. あるベクトル $v \in \{0,1\}^n$ が与えられると, まず決定木の根に対応する変数の値にしたがって, 当てはまる枝へ進む. 次に, その枝の先の節点に対応する変数の値にしたがって, 当てはまる枝へ進む. この過程を葉に到達するまで反復し, 到達した葉の値 (真あるいは偽) を与えられたベクトルの結果 $f(v)$ として返す.

## 3　問題定義

ここでは, 導入されるカットポイント数が最小であるという意味において, 最も簡潔な拡張を求める問題を考える.

**MIN-CP**
入力：数値データの集合 $(P, N)$ (master pdBf $(T^*, F^*)$ は拡張を持つとする).
出力：二値化の結果得られる pdBf $(T, F)$ が拡張を持つカットポイント集合の中で最小のもの.

MIN-CP は集合被覆問題として表され [1], 一般的に MIN-CP は NP-困難であることが証明できる. したがって, 以下では, MIN-CP に対する発見的解法を考察する.

## 4 解法と実験結果

MIN-CP に対する 2 つの発見的方法を比較した. 1 つは集合被覆問題に対する欲張り法 [2] を適用したものであり, もう 1 つは情報量をもとに決定木を構成する ID3 という方法 [4] である. 与えられたデータから標本を選び, 2 つの方法でカットポイント集合を求め, 集合の大きさ及びその有意性を比較した. その結果, カットポイント数に限定して評価した場合, ID3 による解法よりも, 集合被覆問題に対する欲張り法から考えた解法の方が良い結果を出すことが分かった (図 1). もう 1 つの基準である誤り確率 (与えられたカットポイント集合による最適な拡大が得られたとして) については, 同じ標本の大きさでは ID3 による解法が (図 2), そして同じ簡潔さ (カットポイント数) ならば欲張り法による解法が (図 3), それぞれ良い結果を出すことが分かった.

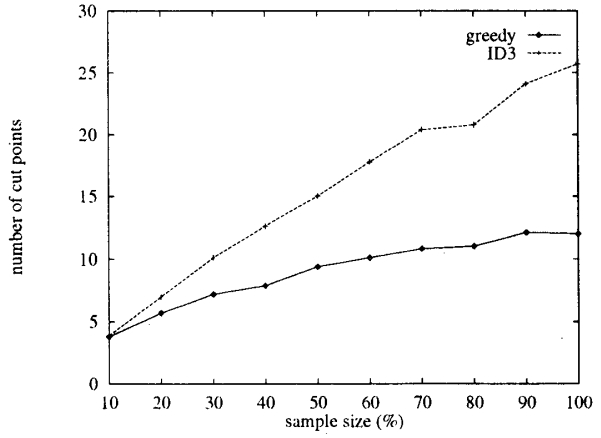


図 1: 乳がんのデータにおける標本の大きさとカットポイント数

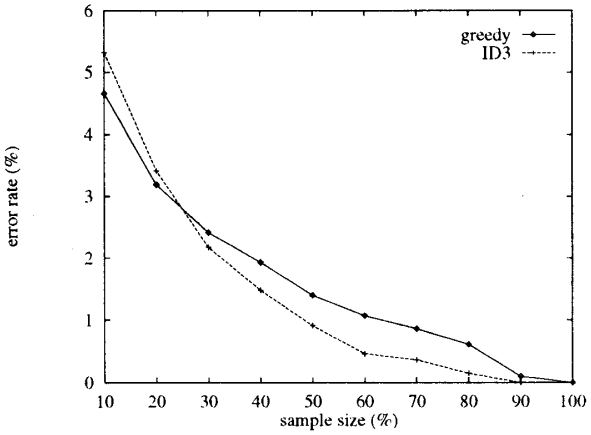


図 2: 乳がんのデータにおける標本の大きさと誤り確率

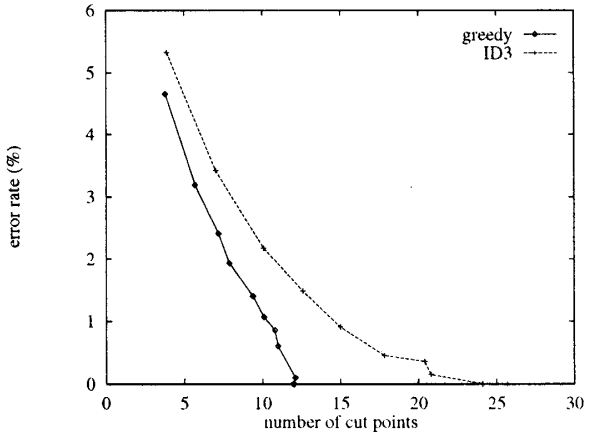


図 3: 乳がんのデータにおけるカットポイント数と誤り確率

## 参考文献

[1] E. Boros, P. L. Hammer, T. Ibaraki and A. Kogan, Logical analysis of numerical data, *unpublished manuscript*, 1995.

[2] V. Chavatal, A greedy heuristic for the set-covering problem, *Mathmatics of Operations Research*, 4 (1979) 233-235.

[3] Y. Crama, P. L. Hammer and T. Ibaraki, Cause-effect relationships and partially defined Boolean functions, *Annals of Operations Research*, 16 (1988) 299-325.

[4] J. R. Quinlan, Induction of decision trees, *Machine Learning*, 1 (1986) 81-106.