4Q−4

事例からのルール獲得における中間仮説抽出手法

秋葉 泰弘　金田 重郎
NTT 情報通信網研究所

1 はじめに

ID3 等の診断型ルール獲得アルゴリズムで獲得された知識は、論理学的には積和標準型とみなせる。しかし、人間は背景知識を用いて推論を積み重ねて判断するため、獲得された知識を1段の積和標準型を用いて表現すると、ルールの持つ意味が理解困難となる場合がある。

そこで、事例から2段のルールを生成し、この1段目の結論を中間仮説として人間に提示して知識獲得を支援する手法を提案する。本手法は、例外を除去した事例から1段のルールを獲得するステップと、1段ルールを2段化するステップから構成され、記述量が(準)最小化された2段ルールを生成するように例外事例を決定する。これにより、例外事例の影響を抑えた2段ルールの獲得が期待される。

2 準備

Definition 1 (notation)
(1) 論理変数をアルファベットの大文字で表す。

(2) 論理関数をアルファベットの小文字で表す。

(3) 論理関数 f に対し、& を × に、or を + に置き換えて出来る整式で、論理関数 f を表現する。
—例 —
論理式

$$A\&B\&CorA\&DorE$$

を整式

$$ABC+AD+E$$

で表現する。

Definition 2 (weak division)
論理関数 f,g に対して、論理関数 q,r が存在して、論理関数 f を論理関数 g で代数的に除法をした時に、商が論理関数 q で剰余が論理関数 r であるとする。この時、論理関数 f は、論理関数 g で *weak division* 可能であると言う。

—例 —
論理関数 $ABC+AD+E$ は、論理関数 $BC+D$ で *weak division* 可能である。

Definition 3 (広義の kernel)
論理関数 f が、*weak division* 可能な論理関数であってかつ、代数式として既約な多項式とみなせる時、論理式 f を(狭義の) *kernel* という。また、論理関数 f が、*weak division* 可能な論理関数であってかつ、代数式として単項式としてみなせる論理関数である時、論理式 f を *pseudo kernel* と言う。さらに、(狭義の) *kernel* と *pseudo kernel* を総称して、広義のカーネルと言う。

—例 —
論理関数 $ADF+ADG+AE+BF+BG+CF+CG$ は狭義の *kernel* として $DF+DG+E$ と $F+G$ を持ち、*pseudo kernel* としては AD、AF 等を持つ。

Definition 4 (MDL 原理)
「ルールの記述量とそのルールの例外の記述量」(以下この量を‖値と呼ぶ。)の和が最小になるルールが、未知事例に対する決定能力を最大とする。

3 問題とその解決アルゴリズム

problem 1
n 個のクラス $\{C_i\}_{1\leq i\leq n}$ に分類される事例集合 $\mathfrak{D}$ から *MDL* 原理を満たす2段ルールを求めよ。

基本方針
次の処理(*)により、事例集合 $\mathfrak{D}$ の任意の部分集合 $\mathfrak{S}$ に対応する2段ルールを生成し、その2段ルールとその2段ルールの例外事例(補集合 $\overline{\mathfrak{S}}$ の部分集合)の中から、*MDL* 原理を満たす2段ルールを探索する。(以下探索 *I* と呼ぶ。)

処理(*)
部分集合 $\mathfrak{S}$ から ID3 により各クラス C_i を説明する1段ルールを生成し、その各1段ルールの広義の kernel 集合 $K_i(1\leq i\leq n)$ から任意の kernel を選び、2段ルールの候補1を生成する。さらに、残りのその各1段ルールの広義の kernel から任意の kernel を選び、2段ルールの候補2を作る。以降この操作を可能な限り試行し、最後の候補を部分集合 $\mathfrak{S}$ に対する2段ルールの候補にする。あらゆる広義の kernel の組合せに対する2段ルールの候補の中で、ルール記述量最小の2段ルールを探索し(以下探索 II と呼ぶ。)、それを部分集合 $\mathfrak{S}$ に対する2段ルールとする。

.....

例えば、クラス C_1 に対する1段ルールが

$$f=ADF+ADG+AE+BF+BG+CF+CG$$

である時、f の kernel $DF+DG+E$ を選んで、

$$kernel1=DF+DG+E \tag{1}$$

と置き、f を kernel1 で weak division し、

$$f=A(kernel1)+BF+BG+CF+CG \tag{2}$$

と変形し、式(1)を1段目の候補とし、式(2)を2段目の候補1とする。次に f の kernel $F+G$ を選んで、f の kernel $F+G$ を選んで、

$$kernel2=F+G \tag{3}$$

と置き、f を kernel2 で weak division し、

$$f=A(kernel1)+B(kernel2)+C(kernel2) \tag{4}$$

と変形し、式(1)と式(3)を1段目の候補2とし、式(4)を2段目の候補2とする。これ以上試行が出来ないので、1段目の候補2を最終的な1段目の候補、2段目の候補2を最終的な1段目の候補とする。また、部分集合 $\mathfrak{S}$ に対する2段ルールにより説明されない集合 $\overline{\mathfrak{S}}$ の要素を部分集合 $\mathfrak{S}$ に対する2段ルールの例外事例とする。

.....

†Two-stage rule induction using MDL principle
Yasuhiro Akiba , Shigeo Kaneda
NTT Network Information Systems Laboratories

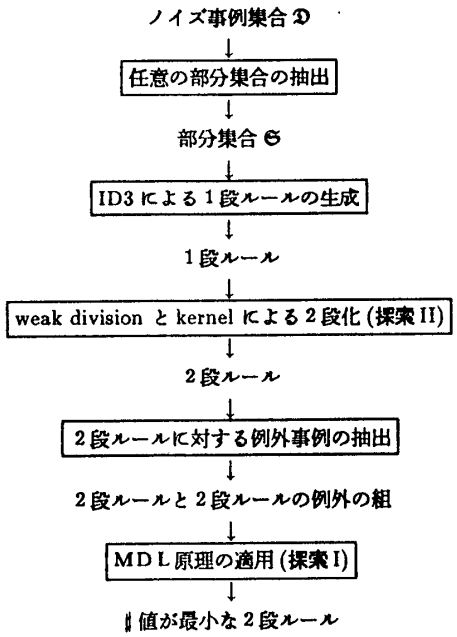

図 1. 各処理の関係

一般に、探索 I 及び探索 II を全探索すると組合せ的爆発を起こすので、次の様なヒューリスティクを提案する。

—探索 I に対するヒューリスティク —
探索の始点を全事例集合 𝔇 とし、濃度が 𝔇 の濃度より 1 小さい部分集合にのみ探索を進め、その中で ‖ 値が探索始点の ‖ 値より小さいものがあれば、その集合を探索の始点として先と同様の操作を繰り返すと言う様に探索を深めていき、出来なくなったところで探索を終る。

—探索 II に対するヒューリスティク —
ルールの記述量を一番減らす (狭義の) kernel を優先的に採用し、weak division 可能でなくなったら、探索をやめる。

4 アルゴリズムの評価

データセット 1: 扶養家族に係わる控除金額決定事例 ..... 128 事例

データセット 2: ISDN 問診 ES からの事例 ..... 84 事例

を基に次の 5 項目についてアルゴリズムの性質を評価した。

項目 1: 記述量の減少傾向
データセット 1、2 を本アルゴリズムに通したところ、記述量を三分の一から二分の一削減

項目 2: 記述量最小化によるノイズ除去動作
データセット 1 128 事例からランダムに選択した 3 事例にクラスのノイズを混入して、本アルゴリズムに通したところ、ノイズをほぼ正確に除去可能

表 1. ノイズ除去結果

ノイズ除去	非ノイズ除去	
3 個	0 個	2 ケース
2 個	1 個	2 ケース
2 個	0 個	5 ケース
0 個	0 個	1 ケース

項目 3: ノイズ除去による未知事例に対する正当率の推移
データセット 1 中から 75% をランダムに抽出し、ID3 と本アルゴリズムで各々ルールを生成し、先の残りの 25% でテストしたところ、前者は 80% の正解率であるのに対し、後者は 92%

項目 4: トレーニングデータのノイズの有無による中間仮説の違い
データセット 1 128 事例からランダムに選択した 3 事例にクラスのノイズを混入したデータセットと生のデータセット 1 を各々本アルゴリズムに通したところ、出力された中間仮説はほぼ同一

表 2. データの差異による中間仮説の違い

トレーニングデータ	生成された中間仮説				
	k1	k2	k3	k4	k5
データセット 1	○	○	○	○	
データセット 1 の 75%		○	○		
データセット 1 に 3 つクラスノイズ混入 (その 1)	○	○	○	○	
データセット 1 に 3 つクラスノイズ混入 (その 2)	○	○	○	○	○

k1 : 20 歳又は配偶者 / 子どもで 80 歳
k2 : 別居又は同居健常者
k3 : 扶養対象同居障害者
k4 : 扶養対象同居健常者
k5 : 配偶者又は子供

項目 5: 生成された中間仮説 (2 段ルールの 1 段目のルール) の妥当性
データセット 1 を本アルゴリズムに通したところ、所得税法上意味のある中間仮説 (2 段ルールの 1 段目) をある程度生成

5 おわりに

従来のルールインダクション (ID3) により専門家に判りやすいルールを、中間仮説を生成することにより生成した。また、本アルゴリズムは、ノイズをほぼ正確に除去可能であった。

参考文献

[1] Brayton R.K.,McMullen C.(1982) : "The decomposition and factorization of Boolean expressions",in Proc Int. Symp. Circ. Syst.(ISCAS-82),Rome,May 1982. Algorithms for VLSI Synthesis.

[2] Quinlan,J.R. (1986):"Induction of decision trees",Mach. Learning 1,81-106

[3] Quinlan J.R.,Rivest R.L. (1989):"Inferring Decision Trees Using the Minimum Description Length Principle",Infomation and Computation 80,227-248.