

# ルール導出検証モデルへのCART法の適用と考察

川口翔也(島根大学), 加藤裕一(島根大学), 佐伯徹郎(山口大学)

## 1.研究概要

現在のデータマイニング研究において学習器の評価は判別問題が主であり、学習器内の構造、**ルール導出性能**に着目されることは少ない。本研究では、**ルール導出性能検証法:データ発生・ルール導出検証モデル**を用い、木構造手法の代表CART法によるルール導出性能を考察する。特にシミュレーション実験結果が実データ分析の結果の解釈・評価に有効であること示す。更にSTRIMによる結果との比較も考察する。これらの考察からCART法には、ルール導出法としての基本的欠陥があり、適用する上での注意点が明らかとなる。

## 3.CART法

木構造は最終的には、if-then形式のプロダクションルールに整理でき、本研究ではif-thenルール導出法の一つとして使用

| An example of the output by rpart |                                                                                              |
|-----------------------------------|----------------------------------------------------------------------------------------------|
| List No.                          | Information of output Node<br>(node), split, n, loss, yval, (yprob), * denote terminal node) |
| 1                                 | 1) root 5000 4139 5 (0.17 0.16 0.17 0.17 0.17 0.16)                                          |
| 2                                 | 2) C3=1,2,5,6 3402 2760 1 (0.19 0.18 0.13 0.14 0.18 0.17)                                    |
| 3                                 | 4) C4=1,2,3,4,5 2827 2263 1 (0.2 0.2 0.14 0.14 0.19 0.13)                                    |
| 4                                 | 8) C3=1 676 443 1 (0.34 0.14 0.12 0.14 0.12 0.13)                                            |
| 5                                 | 16) C4=1 143 10 1 (0.93 0.014 0.014 0 0.021 0.021) *                                         |
| 6                                 | 17) C4=2,3,4,5 533 433 1 (0.19 0.18 0.15 0.18 0.14 0.16)                                     |
| 7                                 | 34) C1=2,3,5 276 209 2 (0.15 0.24 0.2 0.13 0.16 0.12)                                        |
| 8                                 | 68) C2=1,3,6 147 106 3 (0.14 0.2 0.28 0.15 0.088 0.14) *                                     |
| 9                                 | 69) C2=2,4,5 129 92 2 (0.16 0.29 0.1 0.12 0.25 0.093)                                        |
| ...                               | ...                                                                                          |
| 136                               | 123) C2=6 7 2 6 (0 0.14 0 0.14 0 0.71) *                                                     |
| 137                               | 31) C3=4 130 11 4 (0.0077 0.015 0.023 0.92 0.015 0.023) *                                    |

16)→8)→4)→2):(C(4)=1)∧(C(3)=1)∧,...,=(C(4)=1)∧(C(3)=1  
→ If (C(4)=1)∧(C(3)=1)then D=1

68)→34)→17)→8)→4)→2):if (C(1)=2∨3∨5)∧(C(2)=1∨3∨6)∧(C3)=1)∧  
(C(4)=2∨3∨4∨5) then D=3

RL=4のプロダクションルールを36個を導出

|    |                                                                                        |
|----|----------------------------------------------------------------------------------------|
| 結果 | RL=2のルールが7つのみ、予め設定したルールは6つのみで全てのルールを導出することができない。                                       |
| 原因 | 集合を再帰的に二分割するため矛盾データ、無関心データを取り除けない。二分木であるため矛盾データを表現できない。導出されたRLの短いルールから分岐してルールを導出してしまう。 |

## 5.実データへの適用

楽天トラベルデータ:2013~2014年度の約400.000件のアンケートデータ

{C(1)="立地",C(2)="部屋",C(3)="食事",C(4)="風呂",C(5)="サービス",C(6)="設備"}  
D=総合評価

V<sub>a</sub>={1(不満),2(やや不満),3(普通),4(満足),5(大変満足)}

出現頻度にかんりの偏りがあった為、各D=d毎に抽出し、データベースを作成、シミュレーション実験と同様に、ランダムに5000件を取り出して決定表とした。

### STRIM

| Estimated rule R(i) | Condition Part | D | p-vale(z)        | (n <sub>1</sub> ,...,n <sub>6</sub> ) | accuracy | coverage |
|---------------------|----------------|---|------------------|---------------------------------------|----------|----------|
| 1                   | 005050         | 5 | 0.0(46.31)       | (6,4,6,74,662)                        | 0.88     | 0.65     |
| 2                   | 055000         | 5 | 0.0(44.28)       | (9,9,9,46,594)                        | 0.89     | 0.59     |
| 3                   | 005005         | 5 | 0.0(41.28)       | (6,3,5,27,488)                        | 0.92     | 0.48     |
| CART:RL=2           | 050050         | 5 | 5.77E-281(35.80) | (1,2,24,147,511)                      | 0.75     | 0.51     |
| 4                   | 000010         | 1 | 0.0(40.72)       | (634,175,26,4,0)                      | 0.76     | 0.64     |
| 5                   | 040040         | 4 | 5.90E-183(28.82) | (7,22,39,338,60)                      | 0.73     | 0.34     |
| 6                   | 000044         | 4 | 3.92E-166(27.44) | (8,35,57,380,110)                     | 0.64     | 0.39     |
| 7                   | 030030         | 3 | 1.45E-164(27.31) | (43,109,407,88,2)                     | 0.63     | 0.40     |
| 8                   | 004004         | 4 | 6.28E-138(24.97) | (10,19,39,296,81)                     | 0.67     | 0.30     |
| 9                   | 020000         | 2 | 1.02E-84(19.47)  | (146,344,168,28,1)                    | 0.50     | 0.34     |

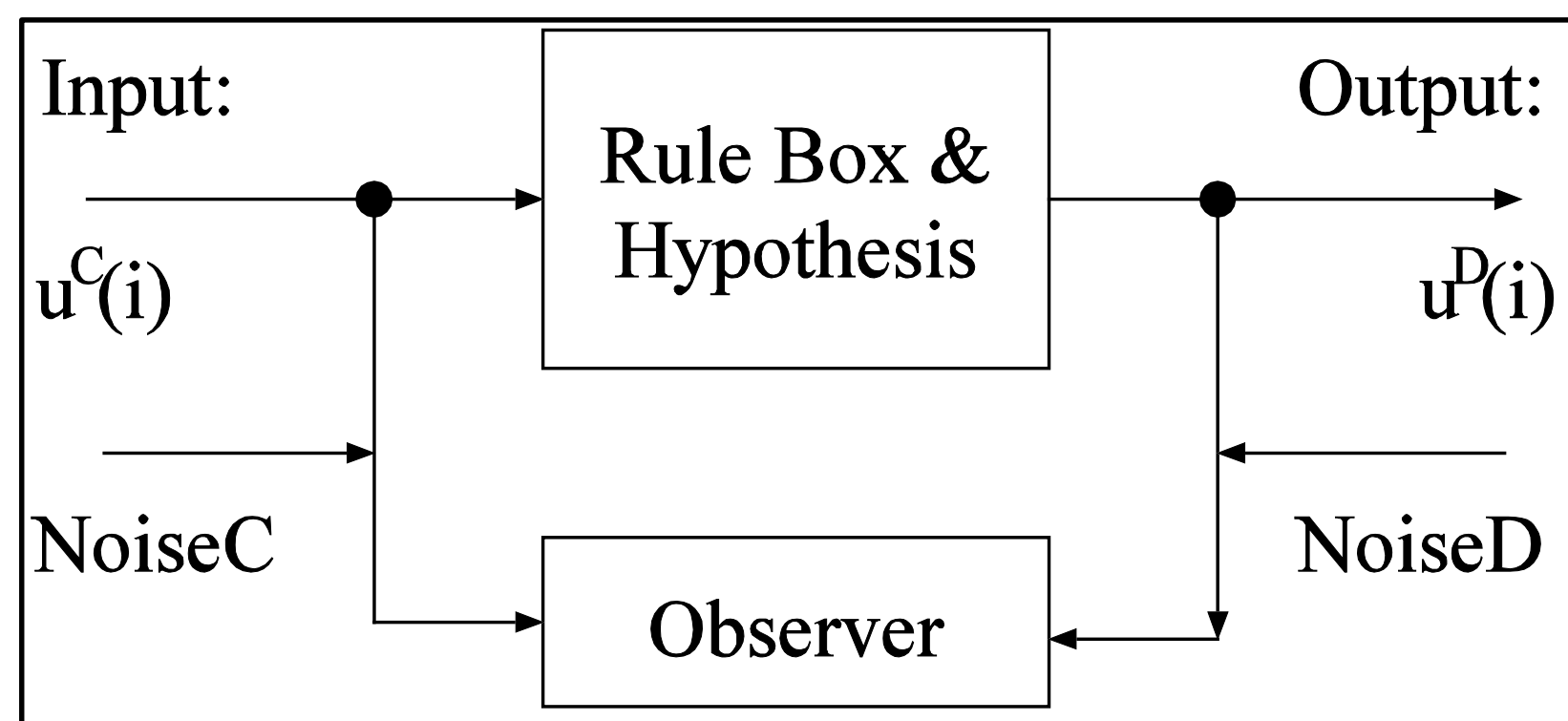
## 6.まとめ

本研究では以下の木構造法の問題点が考察された:

- 無関心データや矛盾データを含んだまま分割を行う。
- 木構造で表現するため矛盾データを表現できない
- 細かく分割されたRLの長いルールを多数導出する

実データに適用した結果、シミュレーション実験との傾向とほぼ合致する。両手法とも有用なルールを人間にわかりやすい形で提供できた。

## 2.データ発生・ルール導出検証モデル



決定表発生:

Rule Box:if-then ルール, Hypothesis:完全一致,  
u<sup>C</sup>(i):ルール条件部データ, u<sup>D</sup>(i);決定部データ



検定用データ:{u(i)=(u<sup>C</sup>(i),u<sup>D</sup>(i))|i=1,...,N}  
↓検証対象のルール導出法

if-then ルールを導出

↑比較検討

Rule Box の設定ルール

Rule Box: if CP(d,k) then D=d

| R(d, k) | CP(d, k) | D=d |
|---------|----------|-----|
| R(1, 1) | 110000   | D=1 |
| R(1, 2) | 001100   | D=1 |
| R(2, 1) | 220000   | D=2 |
| R(2, 2) | 002200   | D=2 |
| R(3, 1) | 330000   | D=3 |
| R(3, 2) | 003300   | D=3 |
| R(4, 1) | 440000   | D=4 |
| R(4, 2) | 004400   | D=4 |
| R(5, 1) | 550000   | D=5 |
| R(5, 2) | 005500   | D=5 |
| R(6, 1) | 660000   | D=6 |
| R(6, 2) | 006600   | D=6 |

Hypothesis

|         | u <sup>C</sup> (i) | D    |
|---------|--------------------|------|
| 一意決定データ | 113456             | 1    |
| 無関心データ  | 123456             | 1~6  |
| 矛盾データ   | 112256             | 1or2 |

決定表:S={U,A=C∪D,V,ρ},

U={u(i)|i=1,...,N=|U|},

A=Attribute Set,

C={C(j)|j=1,...,|C|}:Condition Attributeset,

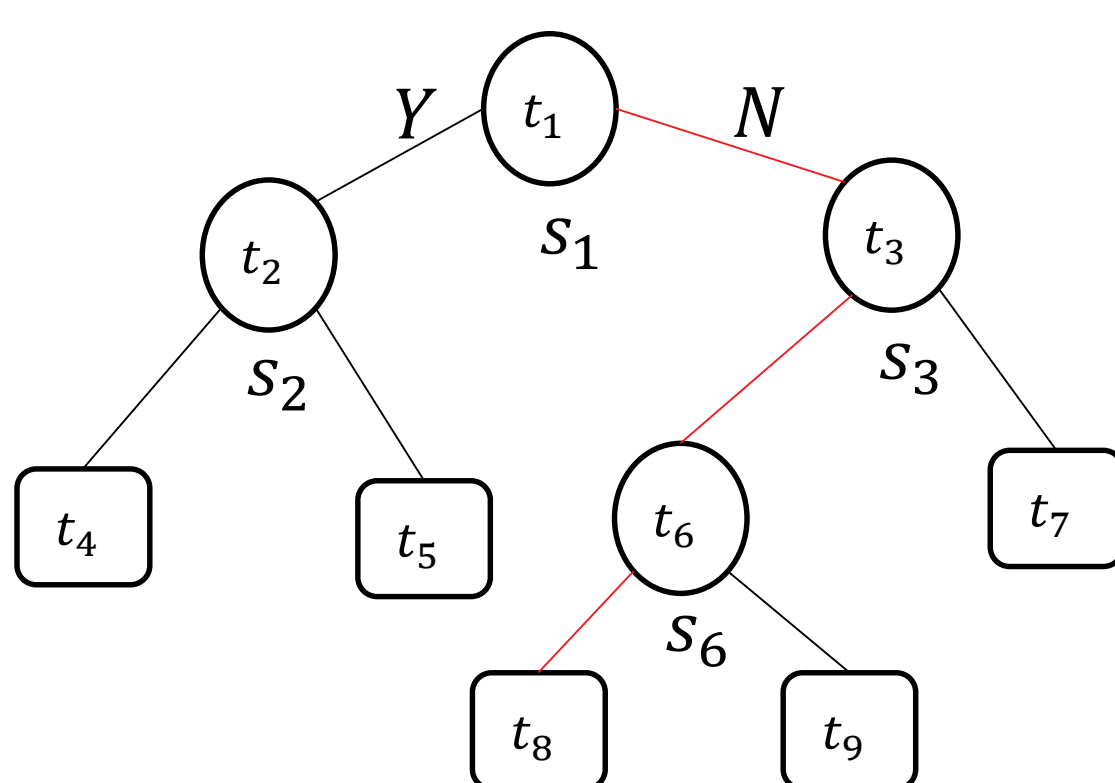
C(j):Condition Attribute,

D: Decision Attribute,

V=∪<sub>a∈A</sub>V<sub>a</sub>: Set of Attribute Value,

ρ:U×A→V: Information Function

|C|=6,|V<sub>C(i)</sub>|=6,|V<sub>D</sub>|=6,N=5000



t<sub>i</sub>:対象集合

S<sub>t</sub>:分岐規則

S<sub>1</sub>∧S<sub>3</sub>∧S<sub>6</sub>:条件部

t<sub>8</sub>の最大度数となる目的変数値:決定部

下の表は導出されたルールの個数を決定属性値、ルール長毎にまとめたものである

| D=d   | Number of rules by Rule Length |   |   |     |     |      |       |
|-------|--------------------------------|---|---|-----|-----|------|-------|
|       | 1                              | 2 | 3 | 4   | 5   | 6    | Total |
| 1     | 0                              | 1 | 0 | 86  | 458 | 792  | 1337  |
| 2     | 0                              | 1 | 0 | 159 | 94  | 1296 | 1550  |
| 3     | 0                              | 2 | 0 | 75  | 222 | 0    | 299   |
| 4     | 0                              | 1 | 0 | 150 | 96  | 1296 | 1543  |
| 5     | 0                              | 1 | 4 | 237 | 36  | 0    | 278   |
| 6     | 0                              | 1 | 0 | 44  | 60  | 504  | 609   |
| Total | 0                              | 7 | 4 | 751 | 966 | 3888 | 5616  |

## 4.STRIM

STRIM(Statistical Test Rule Induction Method)は古典的ラフ集合理論を統計的に解釈したif-thenルール導出法である。下の表はシミュレーションデータのルール導出結果

| Estimated rule R(i) | Condition Part C | D | p-vale(z)        | (n <sub>1</sub> ,...,n <sub>6</sub> ) | accuracy | coverage |
|---------------------|------------------|---|------------------|---------------------------------------|----------|----------|
| 1                   | 660000           | 6 | 1.38E-160(26.98) | (1,2,2,3,1,149)                       | 0.9430   | 0.1865   |
| 2                   | 330000           | 3 | 5.45E-151(26.15) | (0,0,139,1,1,1)                       | 0.9789   | 0.167    |
| 3                   | 002200           | 2 | 3.17E-145(25.63) | (2,138,1,2,1,1)                       | 0.952    | 0.1673   |
| 4                   | 006600           | 6 | 1.27E-137(24.94) | (6,1,4,2,3,137)                       | 0.895    | 0.171    |
| 5                   | 110000           | 1 | 8.81E-133(24.49) | (126,1,1,0,1,0)                       | 0.977    | 0.148    |
| 6                   | 005500           | 5 | 2.16E-131(24.36) | (1,1,1,2,133,3)                       | 0.943    | 0.154    |
| 7                   | 001100           | 1 | 1.24E-130(24.29) | (133,2,2,0,3,3)                       | 0.930    | 0.156    |
| 8                   | 550000           | 5 | 1.35E-129(24.19) | (1,0,0,2,130,4)                       | 0.949    | 0.151    |
| 9                   | 003300           | 3 | 1.59E-127(23.99) | (3,3,130,2,2,3)                       | 0.909    | 0.157    |
| 10                  | 220000           | 2 | 6.93E-124(23.64) | (0,116,1,1,2,1)                       | 0.959    | 0.141    |
| 11                  | 440000           | 4 | 4.35E-123(23.57) | (2,2,0,123,2,4)                       | 0.925    | 0.147    |
| 12                  | 004400           | 4 | 1.74E-117(23.01) | (1,2,3,119,2,3)                       | 0.915    | 0.143    |

Condition Part C:条件属性集合 p-vale(z):P値(観測値を標準化した値) accuracy:正確度  
D:決定属性値 (n<sub>1</sub>,...,n<sub>6</sub>):決定属性値の度数分布 coverage:支持度

## CART法

| D=d   | Number of rules by Rule Length |   |    |     |     |     |       |
|-------|--------------------------------|---|----|-----|-----|-----|-------|
|       | 1                              | 2 | 3  | 4   | 5   | 6   | Total |
| 1     | 1                              | 0 | 2  | 20  | 22  | 0   | 45    |
| 2     | 0                              | 0 | 20 | 95  | 112 | 306 | 533   |
| 3     | 0                              | 0 | 3  | 43  | 24  | 324 | 394   |
| 4     | 0                              | 0 | 5  | 43  | 75  | 48  | 171   |
| 5     | 0                              | 1 | 0  | 17  | 24  | 0   | 42    |
| Total | 1                              | 1 | 30 | 218 | 257 | 678 | 1185  |

|    |                                                                                                   |
|----|---------------------------------------------------------------------------------------------------|
| 結果 | 両手法ともシミュレーションデータ実験と同じ傾向の結果になる。CART法はRL=1のルールはSTRIMのものと同じになったがRL=2のルールは、Z-値やaccuracyの低いルールを導出している。 |
| 考察 | 楽天トラベルデータからはC(1),C(4)は総合評価に影響を与えないことがわかる。                                                         |

## 7.今後の課題

・CART法の発展形であるMART法, Bagging法, Random Forest法にデータ発生・ルール導出検証モデルを適用し、判別問題の観点からルール導出性能の有効性を検討する。

・今回のCART法の比較対象として利用したSTRIMの判別問題での性能評価をBagging 法などの結果と比較・検討する