# Tabu Search による樹形モデルの区分点探索に関する研究

大塚賢太、関 庸一
群馬大学工学部情報工学科
〒376 群馬県桐生市天神町 1-5-1

連続値をとる説明変量と応答変量の関係を調べる統計解析の一手法として樹形モデル分析がある。これは、説明変量で整序したサンプルを、応答変量がなるべく等質になるような区間に分割し、各区間の応答変量レベルを推定するものである。この手法は柔軟にデータの構造を説明できるが、分割位置の選択において、分割評価基準と組合せ最適化の難しさがあった。本研究ではこの問題をモデル選択問題として定式化し、MDL 基準に基づきタブーサーチによる最適化を行なうことで、これらの問題点を避ける手法を提案した。シミュレーション実験の結果から、一変量多分割の樹形モデルに対しては妥当な計算量で、データの誤差分散に見合った推定結果を得られることが明らかとなった。

# Tabu Search Optimization for Partitioning Problem in Tree-Based Model Analysis

Kenta Ohtsuka, Yoichi Seki
Department of Computer Science, Faculty of Engineering, Gunma University
1-5-1 Tenjin-cho kiryu Gunma, 376 Japan.

The Tree-Based Model Analysis is one of statistical methods for uncovering structure between a response variable and predictor variables. It partitions the samples sorted by the predictor variables into homogeneous response groups, and estimates levels of response. It is able to estimate a flexible structure, however, has difficulty in the selection of partitioning criterion and combinatorial optimization. We formulate this as a model selection problem, and propose a method around these problems, using Minimum Description Length (MDL) criterion and Tabu Search Method. Monte Carlo experiments have revealed that the method gives considerably good fit for the models of one predictor variable, within reasonable computational quantity.

## 1 研究目的

連続値をとる説明変量と応答変量の組からなるデータセットが与えられた時に、その両変量間の関係を調べる統計解析の一手法として樹形モデル分析がある [1, 2]。これは、説明変量の大きさで整序したサンプルを応答変量がなるべく等質になるような区間に分割し、それぞれの区間の応答変量のレベルを推定することによって両変量の関係を説明するモデルである。樹形モデルでは、高次線形回帰モデルよりも柔軟にデータの構造を説明できるが、分割位置の選択において、分割の評価基準の選択と組合せ最適化の難しさがあるため敬遠されてきた。

そこで、本研究ではこれらの問題点に対して、MDL 基準を評価基準としてタブサーチを用いて最適化を行なう方法を提案するとともに、一変量多分割の樹形モデルに関しシミュレーション実験を行ない、提案方法の計算効率や推定結果の良さを検討する。

## 2 MDL 原理と樹形モデル

### 2.1 樹形モデル

樹形モデルは、サンプル $i$ を説明変数 $x_i$ の値に従って分類する決定木 [4, 6, 7, 11, 12] を考える。決定木の葉以外のノード (決定節) では説明変数の値に従ってサンプルが分類される。木の葉は、その葉よりも root に近いところにある決定節により分類されたサンプルカテゴリとなる。

図 1では、$a_1, a_2$ 2つの閾値と説明変数の値で3つのサンプルカテゴリ $g_1, g_2, g_3$ を作る様子を示している。一般に事象数の説明変量を考える場合、決定節も複数となるが、本研究では単一の説明変数の場合を扱うので決定節は1つとなる。

樹形モデル分析は回帰分析の一種であり、応答変量 $y_i$ の構造は次のようなモデルで表現される。

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

ここで、

| | |
|---|---|
| $D$ | :サンプルの分割点数 |
| $(a_0, \cdots, a_{D+1})$ | :説明変量の閾値パラメータ |
| ただし、$a_0 = \min x_i$, | $a_{D+1} = \max x_i$ |
| $\boldsymbol{y} = (y_1, \cdots, y_N)^t$ | :応答変量値ベクトル |
| $\boldsymbol{Z} = (z_{ij})_{\substack{i=1,\cdots,N \\ j=1,\cdots,D+1}}$ | :$z_{ij} = 1$ if $x_i \in [a_{j-1}, a_j]$ |
| | $z_{ij} = 0$ otherwise |
| $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{D+1})^t$ | :各説明変量区間に対する応答変量レベルパラメータ |
| $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_N)^t$ | :各要素が独立に $N(0, \sigma^2)$ に従う誤差項 |

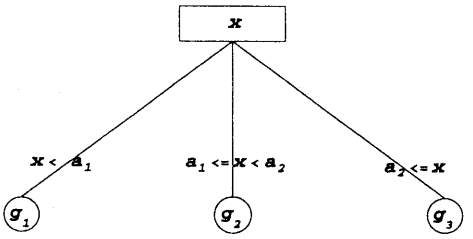


図 1: 決定木の例

このとき、最小二乗法による $\boldsymbol{\beta}$ の推定値は次のように与えられる。

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t\boldsymbol{y} = (\hat{y}_1, \cdots, \hat{y}_{D+1})^t$$

ただし、

$$\hat{y}_j = \frac{1}{n_j}\sum_{x_i \in [a_{j-1}, a_j]} y_i \quad (j = 1, \cdots, D+1)$$
$$n_j = \#\{x_i | x_i \in [a_{j-1}, a_j]\}$$

以上で問題は説明変数の閾値 (分割位置, knot) の個数と位置の決定に集約される。

応答を細かく分類すると、決定木の構造は複雑になるが残差は小さくなる。しかし、偶然誤差まで説明している場合も考えられる。決定木の構造と残差との間にトレードオフ関係があるので、最適な決定木を選択するために MDL 原理を用いる。

### 2.2 MDL 原理

MDL 原理 (Minimum Description Length Principle) は、Rissanen [5, 8, 9, 10] により情報源符号化のモデルから提案されたモデル選択基準である。データに対して「1. モデルのクラスからモデルを選択し、2. そのモデルの下でのデータの符号化をする」という二段階符号化を想定することで、「与えられたデータを、モデル自身の記述も含めて最も短く符号化できるような確率モデルが最良のモデルである」とする原理である。

まず、モデル選択の対象となる確率モデルのクラス $\mathcal{C}$ を次で表す。

$$\mathcal{C} = \{P(\boldsymbol{y} : \boldsymbol{\beta}, M), \boldsymbol{\beta} \in \Theta(M), M \in \mathcal{M}\}$$

| | | |
|---|---|---|
| $\mathcal{M}$ | : | 対象となるモデルの可算集合 |
| $\Theta(M)$ | : | モデル番号 $M$ が定まった下でのパラメータの集合 |
| $P(\boldsymbol{y} : \boldsymbol{\beta}, M)$ | : | モデル番号 $M$ で指定され、パラメータ $\boldsymbol{\beta}$ を持つ確率モデル |

データが与えられた際の MDL 基準は、

$$-\log P(\boldsymbol{y}:\boldsymbol{\beta},M)+l(\boldsymbol{\beta}|M)+l(M) \qquad (2)$$

となり、モデルのクラスの中からこの値が最小なモデルを選択する。ただし、本論文では対数の底は2とする。ここで、(2)式の第一項はモデルを選んだもとでのデータの符号長であり、データの負の対数尤度となる。また、第二項、第三項はそれぞれパラメータとモデル番号の符号長である。MDL原理は両者のトレードオフ関係をバランスさせるモデルを最適なものとして選ぼうとするものである。

(1)式の樹形モデルの場合は、残差平方和を

$$RSS(\boldsymbol{y})=\sum_{j=1}^{D+1}\sum_{x_i\in[a_{j-1},a_j]}(y_i-\hat{y}_j)^2$$

とすると、誤差分布の仮定からMDL基準は

$$L_{MDL}(\boldsymbol{y},\boldsymbol{M})=\frac{1}{2\sigma^2}RSS(\boldsymbol{y})\log e+\sum_{j=1}^{D+1}\log\frac{\sqrt{n_j e}}{\sigma}+l(M) \qquad (3)$$

となる。ただし、ここではモデル選択に無関係な定数項は省いてある。

### 2.3 木の符号長 $l(M)$

決定木の符号長をカテゴリ数と各カテゴリのサンプル数の記述長とする。ここで、カテゴリ数の符号長は次式のRissanenの自然数の符号化を用いる。

$$L^*(x)=\log c+\log x+\log\log x+\cdots$$

ただし、$c\simeq 2.865064$ であり、上式の和は正の項のみについて取る。また、カテゴリのサンプル数の符号長は $(D+1)\log(N+1)$ とする。よって木の符号長は以下となる。

$$l(M)=(D+1)\log(N+1)+L^*(D+1) \qquad (4)$$

## 3 Tabu Searchによる最適化算法

### 3.1 Tabu Search

選択対象となる決定木つまり閾値ベクトルの組合せは $\sum_{D=0}^{n-1}\binom{n-1}{D}=2^{n-1}$ 通りあり、$n$ の増加とともに手に負えない量となる。そこで、タブサーチ[3]によりヒューリスティック解を求める方法を提案する。

タブサーチは、以下のような手続きにより、組合せ最適化問題において大域的な最適解を得るメタヒューリスティックなアルゴリズムである。ある可能解 $s$ から何らかの方法により生成される解を $s$ の近傍と呼び、$NB(s)$ で表す。タブサーチは、基本的には、暫定解の近傍 $NB$ の中で最も評価関数値の良い解に移行していく操作を繰り返し、この操作系列の全過程中で最良の解を返す算法である。

一般のGreedyアルゴリズムに対し、タブサーチアルゴリズムの異なるのは、局所最適解から脱出するためにタブリストと呼ばれるリストを用いて、最近の一定回数の間に選択された解には移行しないように制限する点である。すなわち、現在の暫定解を $s$ とした場合、その後の一定回数の反復の間 $s$ に戻るようないかなる操作も禁止操作とする。

以下にタブサーチアルゴリズムの概要を示す。

```
function Tabu_Search (s);
    s_0 := s ;
    best := s ; { 最良の解 }
    t := 0; {Loop 回数 }
    best_t := 0; { 最良の解が見つかった時の t}
    TL := φ ;
    while t − best_t < Stop_Time do
        t := t + 1;
        s_t := min_{s∈NB(s_{t−1})\TL} f(s);
        TL := TL\{s_{t−L}} ∪ s_t;
        if f(s_t) < f(best) then
            best := s_t;
            best_t := t;
return best ;
```

ここで、

| | | |
|---|---|---|
| $f$ | : | 最小化目的関数 |
| $NB(s)$ | : | 解 $s$ の近傍 |
| $TL$ | : | タブリスト |
| $L$ | : | タブリストの長さ |

である。また、Stop_Timeは、このStop_Time回の反復の間、最良の解 $best$ が更新されなければ探索を終了するという定数である。

### 3.2 暫定解の近傍の生成法

タブサーチアルゴリズムを2節の樹形モデル分析のモデル選択問題へ適用する。具体的にはMDL基準で最適な説明変量の閾値 $\{a_j\}$ をタブサーチによって探索する。

近傍生成操作としては以下を採用する。

**移動** $D$ 個の分割点の内、任意の1つを $P$ 個以内 + 方向、− 方向に動かして得られる分割。

**削除** 現在の分割位置のうち一つの削除。
例：現在の分割位置が $(s_1,\ s_2,\ s_3)$ のとき、削除近傍は $(s_1,\ s_2)$, $(s_1,\ s_3)$, $(s_2,\ s_3)$ となる。

**追加** 現在の分割位置とは重ならない新たな分割位置の追加。

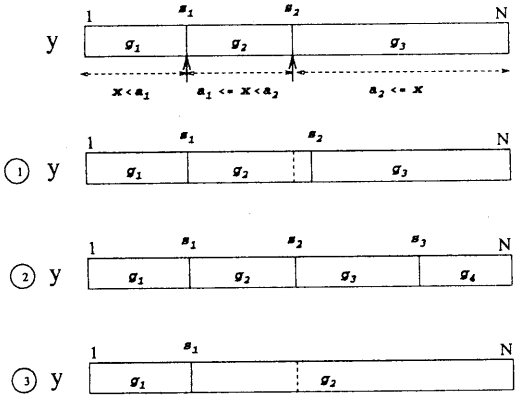


図 2: 分割法 $s$ の (1) 移動、(2) 追加、(3) 削除

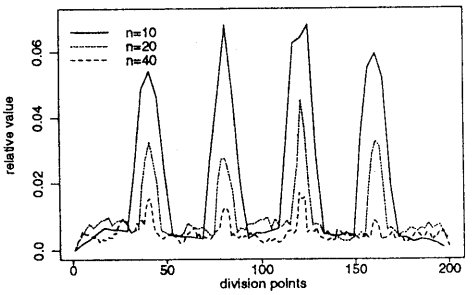


図 3: $D=4,\ K=3$ の $w_i$ の例

例：現在の分割位置を $(s_1,\ s_2)$ とする。$s_1,\ s_2$ と同じではない乱数で得られた新たな分割位置を $x$ とするとき、新しい分割位置は $s_1,\ s_2,\ x$ をソートして得られる $(s'_1,\ s'_2,\ s'_3)$ となる。

このようすを図2に示す。分割点数が $D$ の時の近傍は $D(P+1)+1$ 個となる。

追加分割位置の生成に関しては、第1に事前情報を用いない方法として、$[1,\ N-1]$ 一様乱数を用いる方法が考えられる。第2に、分割点としての適切さに関する情報があればこれを利用し、適当な重み付き乱数で生成することが考えられる。

後者の場合重みが適切であれば結果として計算量の減少が期待できる可能性がある。利用可能な情報としては、応答変量が与えられているので、そのバラツキ程度が考えられる。そこで、$x_i$ を分割点として選ぶ確率を入力データの $K$ 近傍 (K-knots) 誤差分散の2乗和 $w_i$

$$w_i = \sum_{k=-K}^{K} (y_{i+k} - \bar{y}_i)^2$$

$$\bar{y}_i = \frac{1}{2K+1} \sum_{k=-K}^{K} y_{i+k}$$

に比例させて、重み付き乱数により決定する方法を提案する (図3)。確率となるよう総和基準化した $w_i$ を $\tilde{w}_i$ とする。ただし、この際、すでに選択されている点の近傍については、以下のように、その点からの距離に応じて確率を減少させるものとする。

Loop: $\tilde{w}_i$ 重みつき乱数 $d$ を生成
　if ( 分割点 $d$ が既存の分割点 $s_j$ の近傍 ) {
　　$[0,1)$ 一様乱数 $u$ を生成
　　if ( $u > \frac{|d-s_j|}{K}$ )
　　　Goto Loop　{ $d$ を棄却 }
　}
　$d$ を採用

## 4 数値実験

提案法が真の母平均モデルと異なるモデル選択を行なう場合として、1, タブサーチが局所最適解から脱出できない場合、2, MDL 基準が真の決定木で最適とならない場合の二つが考えられる。この両者について、シミュレーション実験で検討する。

### 4.1 タブサーチの能力についての実験

サンプル数や真の分割数で代表される問題の規模に対して、提案手法のタブリスト長 $L$、$Stop\_Time$ はどの程度で十分かについて MDL 値の最小値からの探索誤差の観点から評価する。また、このときの計算量も評価する。ただし、計算量は MDL を評価した回数とする。

用いるサンプルデータは、$x_i \in [a_{j-1}, a_j]$ のとき

$$y_i \sim N(\mu_j,\ 1^2) \quad \left( \begin{array}{l} i=1,2,\cdots,N, \\ j=1,\cdots,D+1 \end{array} \right)$$

によって生成し、真のカテゴリ数 $D+1=2,5,9$、各カテゴリのサンプル数 $n=10,20,40$ $[N=(D+1)n]$ という問題規模で各 60 ケースを用意した (図4)。

各カテゴリの母平均は、隣合うカテゴリの間での差 $\Delta\mu$ を $\frac{10}{\sqrt{n}}$ として設定した。これにより、カテゴリの分割点が正しい際のカテゴリ平均値の差の推定誤差分散が、規模に関わらず、等しくなるように設定したこととなる。

また、最適化パラメータは $Stop\_Time=400$、$P=1$、$K=3$ として、$L=20,50,100$ について検討した。探索中の MDL 値の変化の様子を図5に示す。

以上の実験組合せについて、異なる 10 乱数系列を用いて探索を行なうことで、1回の Tabu Search による結果の不安定さも検討する。また、真の最小 MDL 値を大きな問題に対しては求めることが困難であるので、以

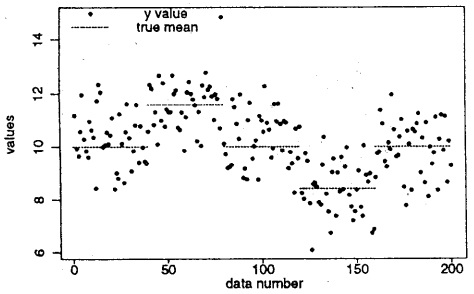


図 4: $D = 4, n = 40$ のデータの一例

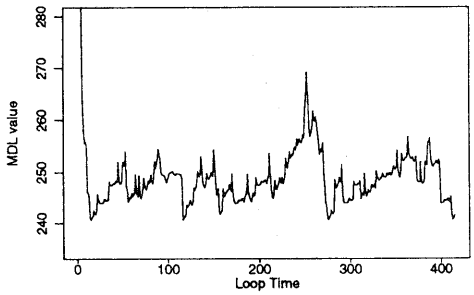


図 5: $D = 4, n = 20, L = 100$ のときのタブサーチ手続き内のループでの最小 MDL 値の推移例

上の解析によって得られた MDL 値 $10 \times 3 \times 2$ (個/データセット) のうちで、最小のものを暫定最適解として、探索誤差 = (該当 MDL 値) − (暫定最適解) を用いた。

図 6(a) に示すように分割数が小さいとき ($D = 1$) には、タブリスト長に関係なく常に最適解が得られている。分割数 $D$ やサンプル数 $n$ が大きい時には、タブリスト長の増加とともに探索誤差は小さくなるが、0 にはならないことがわかる。この場合には、長いタブリスト長を用意したり、探索の際に用いる乱数系列を変えて、反復実行する必要があることがわかる。

また、分割点追加の際用いる乱数に関しては、大規模問題では重み付き乱数を用いた方が探索誤差が小さくなることがわかる (図 6(b))。

計算量は一般に問題規模、タブリスト長の増加につれて増えるが、大規模問題には重み付き乱数を用いた方が計算量も少なくて済む (図 7)。これは、重みつき乱数により良い分割点を早期に探索できるためと考えられる。

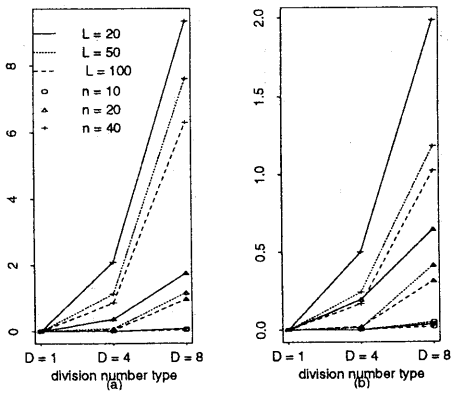


図 6: 探索誤差の比較

**(a):** 一様乱数を用いた場合の探索誤差

**(b):** 一様乱数を用いた場合と重み付き乱数を用いた場合との探索誤差の減少分

表 1: 真のモデルより少ないモデルを採択した回数 (重みあり, $K = 3$)

| | $n = 20$ |
|---|---|
| $\Delta\mu = 2$ | 56/4/0/0/0/0 |
| $\Delta\mu = 5$ | 6/26/70/6/0/0 |
| $\Delta\mu = 10$ | 0/0/0/0/78/2 |

表はそれぞれ分割数 0/1/2/3/4/5 のモデルを選択した回数を示す。

## 4.2 MDL 基準についての実験

誤差分散 $\sigma^2$ に対して、カテゴリ間の母平均の差 $\Delta\mu$ が小さい場合、MDL 基準の与える解の傾向について分割数や分割位置の真のモデルとのずれから調べる。

データは前節同様の方式で、$D = 4$ の場合について $\Delta\mu = \frac{2}{\sqrt{n}}, \frac{5}{\sqrt{n}}, \frac{10}{\sqrt{n}}$ として、60 ケース用意した。また、タブサーチの実験パラメータは前節の結果から $L = 100$ などとした。

真のモデル $D = 4$ に対し採択された分割数ごとのケース数を表 1に示す。これより、グループ平均の差がサンプル数に比べ少ないときには、MDL 基準の性質からより単純なモデルを選ばれる傾向があることが明らかとなった。

## 4.3 連続型データについての実験

真のモデルが連続的に変化するデータについても、本手法がどの程度の結果を与えるかを $y$ の母平均が正規分布の密度関数のような滑らかな変化をする場合の例で示

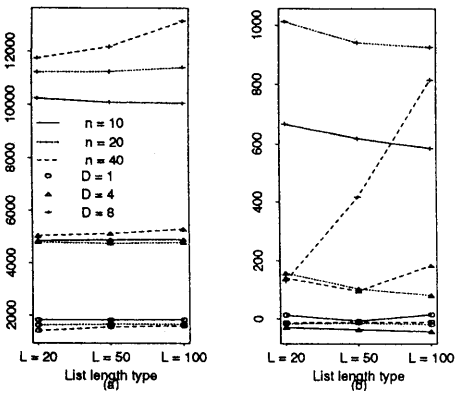


図 7: 計算量の比較

(a): 一様乱数を用いた場合の計算量

(b): 重み付き乱数を用いた場合の計算量と (a) との減少分

す。母平均のレンジと誤差分散の比が $h = 5, 20$ の場合の結果を図 8に示す。

これより、誤差分散に比べデータ数が多くなるにつれて、母平均の変化に追従してより細かい分割が行なわれる様子がわかる。

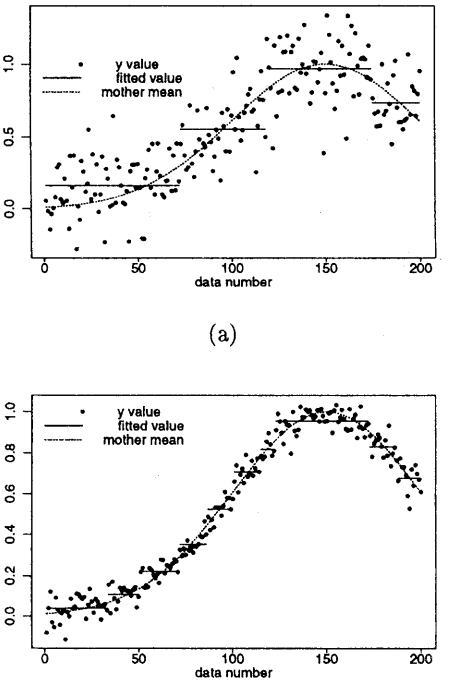


図 8: $n = 200$ の場合の連続型データについての結果
(a): $h = 5$ のとき, (b): $h = 20$ のとき

## 5 結論

樹形モデルを推定する方法として、MDL 基準をタブーサーチにより最適化する手法を提案した。シミュレーション実験の範囲で、提案手法は、その重みつき乱数による分割点追加などの算法により、かなり大きな問題に対しても、MDL 基準の意味で最適な解を妥当な計算量で与える。また、用いた MDL 基準はデータの誤差分散に見合った推定結果を得られることが明らかとなった。

多変数モデルへの拡張が今後の課題となる。

## 参考文献


[1] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., *Classification and Regression Trees*, Wadsworth International Group, Belmont, C.A., 1984

[2] Chambers,J.M., and Hastie,T.J., *Statistical Models in S*, Wadsworth & Brooks/Cole Advanced Books & Software, A Division of Wadsworth, Inc., Pacific Grove, C.A., 1992

[3] 藤沢克樹, 久保幹男, 森戸晋, Tabu Search のグラフ分割問題への適用と実験的解析, 電学論 C, vol.114, No.4, pp.430-437, 1994

[4] 伊藤秀一, MDL 入門:MDL のパターン認識への応用, 人工知能学会誌, vol.7, no.4, pp.608-614, 1992

[5] 韓太舜, 山西健司, MDL 入門: 情報理論の立場から, 人工知能学会誌, vol.7, no.3, pp.427-434, 1992

[6] Quinlan, J.R., and Rivest, R.L., Inferring decision trees using the minimum description length principle, *Information and Computation*, 80, pp.227-248, 1989

[7] Quinlan, J.R., Decision trees and Decision-making, *IEEE Trans. System, Man and Cybernetics*, vol.20, no. 2, pp.339-346, 1990

[8] Rissanen,J., Modeling by shortest data description, *Automatica*, vol.14, pp.465-471, 1978

[9] Rissanen,J., A universal prior for integers and estimation by minimum description length, *Annals of Statistics*, vol.11, no.2, pp.416-431, 1983

[10] Rissanen,J., Universal coding, information, prediction and estimation, *IEEE Trans. on Information Theory*, vol.IT-30, pp.629-636, 1984

[11] 鈴木秀男, 圓川隆夫. MDL 基準による判別木の生成, 人工知能学会誌, vol.10, no.4, pp.572-579, 1994

[12] 山西健司, MDL 入門: 計算論的学習理論の立場から, 人工知能学会誌, vol.7, no.3, pp.435-442, 1992