記述長最小基準と状態分割の立場からみた確率モデルの選択方法について†☆

鈴木 譲††☆☆ 大嶽康隆††☆☆☆ 平澤茂一††

本論文では，具体的な訓練例系列から例外を許容する不確実な知識を学習する問題（確率的知識の学習）について検討している．入出力の系列からの学習の問題に限定すれば，確率的な関係をもつ入出力の系列から，各入力を前提にした各出力の条件付確率（確率的知識）を推定する問題として定式化できる．この問題を記述長最小基準から検討する場合に，状態分割の立場からの枠組みは既存のものである．本論文は，その修正・拡張およびその際に生ずる問題について解決している．具体的には，まず系列が入出力の形式になっている前提を除去した一般的な場合を検討している．すなわち，多次元の属性値の確率的な因果関係を見いだす問題についてである．しかしながら，学習の対象を広げることによって，一般には学習の計算量は増加する．本論文では，この視点にたって学習の対象の広さと学習の計算量の相互関係を議論している．そして，学習の対象を若干狭めても学習の計算量を十分に低減させる方策について検討している．その一例として，問題を近似して学習対象を Dendroid 分布に限定した場合の記述長最小基準における最適解法を導出している．導出した学習アルゴリズムは，C. K. Chow らのアルゴリズムに基づいているが，属性間の依存関係が単一の木構造になる仮定を排除する一般性がある．

1. はじめに

本論文では，具体的な訓練例系列から例外を許容する不確実な知識を学習する問題（確率的知識の学習）について検討する[2),3)]．入出力の系列からの学習の問題に限定すれば，この検討は

1. 関数的な関係をもつ入出力から，各入力に対して一意的に決定される出力（確定的知識）を見いだす問題．
2. 確率的な関係をもつ入出力から，各入力を前提にした各出力の条件付確率(確率的知識☆☆☆☆)を見いだす問題．

の後者についての考察に相当する．

この問題について，記述長最小基準[4),5)]から検討する場合に，状態分割の立場からの枠組みは既存のものである[1)]．本論文では，状態分割の枠組みの修正・拡張およびその際に生ずる問題の解決を検討する．

まず，状態分割は確率的な入出力関係の学習を前提にして提案されており，任意の属性値間といった一般的な関係の学習を想定するものではない[1)]．本論文の最初のねらいは，訓練例が入出力の形式になっている前提を除去した一般的な場合を検討することである．すなわち，多次元の属性値の確率的な因果関係を見いだす問題についてである．

しかしながら，学習の対象を広げることによって，一般には学習の計算量は増加する．本論文の2番目のねらいは，学習の対象の広さと学習の計算機の相互関係を確認するとともに，学習の対象を若干狭めても学習の計算量を十分に低減させる方策を検討することである．この一例として，問題を近似して学習対象を Dendroid 分布[6)]に限定した場合について，記述長最小基準における最適解法を導出する．提案アルゴリズムは，C. K. Chow らのアルゴリズム[7)]に基づくが，属性間の依存関係が単一の木構造になる仮定を排除する一般性がある．

2. 準備

2.1 確率的な入出力関係の学習

確率的な属性値関係の学習について検討する前に，確率的な入出力関係の結果[1)]について整理しておこう．

時点 $-n+1$ から 0 まで ($n=1,2,\cdots$) の長さ n の実際の入出力の系列

$$z_{-n+1}=(x_{-n+1}^{(1)}, x_{-n+1}^{(2)}, \cdots, x_{-n+1}^{(N)}, y_{-n+1})$$
$$z_{-n+2}=(x_{-n+2}^{(1)}, x_{-n+2}^{(2)}, \cdots, x_{-n+2}^{(N)}, y_{-n+2})$$
$$\cdots \quad \cdots$$
$$z_0=(x_0^{(1)}, x_0^{(2)}, \cdots, x_0^{(N)}, y_0)$$

† Selection of the Stochastic Model Based on the Minimum Description Length Principle and State Decomposition by JOE SUZUKI, YASUTAKA OHDAKE and SHIGEICHI HIRASAWA (Department of IE and Management, School of Science and Engineering, Waseda University).

†† 早稲田大学理工学部工業経営学科

☆ 本研究は，一部本学平成3年度特定課題研究 91A-170，および平成3年度電気通信普及財団研究の助成によっている．

☆☆ 現在 青山学院大学理工学部経営工学科
☆☆☆ 現在 (株)東芝 システム・ソフトウェア生産技術研究所

☆☆☆☆ 本論文では確率的な知識だけを扱うので，以後，単に知識とよぶことにする．

(z^n で表記する) から，将来の入出力の系列

$$z=z_1z_2\cdots, z_i=(x_i^N, y_i),$$
$$x_i^N=(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(N)}),$$
$$N=1, 2, \cdots$$

の知識を学習する．具体的には，入力 x_i^N のもとでの出力 y_i の条件付確率 $P(y_i|x_i^N)$ を推定する．ここで，実現値 $x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(N)}$ のとる値はそれぞれ無限集合の要素でよいが，実現値 y_i のとる値は有限集合 $A=\{0, 1, \cdots, J-1\}$ の要素であるとする．また，以下では $x_i^N=(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(N)})$，$y_i$ をそれぞれ時点 i に対しての属性値ベクトル，およびクラスということがある．また，本論文では，一時点の入出力 z_i の生起が過去の系列 $z_{-n+1}z_{-n+2}\cdots z_{i-1}$ には依存しないことを仮定する．

仮定1 系列の各値 $z_{-n+1}, z_{-n+2}, \cdots, z_0, z_1, \cdots$ は，時点 i ごとに独立に生起する．

さらに，本論文では，状態分割の枠組み[1]を前提としてこの問題を検討する．状態分割の枠組みは，条件付確率の値自身である確率パラメータ，および確率パラメータを設定する構造であるモデルから構成される(定義1参照)．したがって，真の知識 K から発生した入出力の系列 $z^n=z_{-n+1}z_{-n+2}\cdots z_0$ を用いて，真のモデルおよび真の確率パラメータを推定することは知識 K を学習すること等価なのである．以下では，推定された知識（推定されたモデル，および推定されたパラメータ）と区別するために，必要に応じて，真の知識（真のモデル，および真の確率パラメータ）という表現を用いる．

定義1 G をモデル g の集合とする．モデル $g\in G$ は，属性値ベクトル x_i^N の空間を $S(g)$ 個の状態に分割する．これにより，各属性値ベクトル x_i^N は $s=1, 2, \cdots, S(g)$ のいずれかの状態に属すことになる．そして，モデル g は各状態 $s=1, 2, \cdots, S(g)$ において，クラスの空間である集合 $A=\{0, 1, \cdots, J-1\}$ を $J(s, g)$ 個の群に分割する[8]．これにより，状態 s において各クラスは $j=0, 1, \cdots, J(s, g)-1$ のいずれかの群に属すことになる．そして，状態 s で同じ群 j に含まれる $m[j, s, g]$ 個のクラスは，各々等確率で生起するものとする．

したがって，モデル g の各状態 $s=1, 2, \cdots, S(g)$ について，(1)が成立する．

$$\sum_{j=0}^{J(s, g)-1} m[j, s, g]=J \qquad (1)$$

すなわち，モデル $g\in G$ が特定されれば，各状態 $s=1, 2, \cdots, S(g)$ を前提にした各群 $j=0, 1, \cdots, J(s, g)-1$ の生起確率 $p[j, s, g]$ (確率パラメータ) を決定することにより知識が定まるのである．

また，以下では，

$$k(g)=\sum_{s=1}^{S(g)}[J(s, g)-1] \qquad (2)$$

をモデル g の確率パラメータの総数とよぶことがある．各状態 $s=1, 2, \cdots, S(g)$ において，$p[0, s, g]$ の値が次式によって決定されるからである．

$$p[0, s, g]=1-\sum_{j=1}^{J(s, g)-1} p[j, s, g] \qquad (3)$$

補題1[1] 定義1で示した枠組みでは，長さ n の入出力の系列 z^n を，モデル g の記述およびそのモデル g のもとでの系列の記述の2段階で記述する際に，高々以下の長さ $length_L(z^n)$ で記述する言語 L が存在する．ここで，クラフトの不等式 $\Sigma_{z^n\in Z^n}2^{-length_L(z^n)}\leq 1$ が成立する．また，各状態 s のもとで最終的に群 $j=0, 1, \cdots, J(s, g)-1$ が生起した頻度，および各状態 $s=1, 2, \cdots, S(g)$ が生起した頻度をそれぞれ $n[j, s, g]$, $n[s, g]$ とおいた．さらに，C_1 を具体的な定数とおいた．

$$length_L(z^n)=H[g](z^n)+\frac{k(g)}{2}\log n+C_1 \qquad (4)$$

$$k(g)=\sum_{s=1}^{S(g)}[J(s, g)-1] \qquad (5)$$

$$H[g](z^n)=\sum_{s=1}^{S(g)}\sum_{j=0}^{J(s, g)-1} -n[j, s, g]\log\frac{n[j, s, g]}{m[j, s, g]n[s, g]} \qquad (6)$$

このときの記述長最小基準[4),5)]に基づくアルゴリズムは，以下[1)]で与えられる

アルゴリズム1

procedure *Algorithm* 1(N)

begin

N 次の属性値ベクトル x_i^N を前提にした各クラス y_i の条件付確率 $P(y_i|x_i^N)$ を表現するモデル $g\in G$ のうち，(4)を最小にするモデルを選択する

end.

上記では，高々 $M-1$ 回の記述長の比較によってモデルを推定している．

仮定2 真のモデル g_K は，有限個 (N 個) のモデル $g_1, g_2, \cdots, g_M$ のいずれかと一致する．

モデルが推定された後，各確率パラメータ $p[j, s, g]$ $(j=1, 2, \cdots, J(s, g)-1, s=1, 2, \cdots, S(g))$ は状態ごとの各群の頻度 $n[j, s, g]$ によって推定される．確率パラメータの推定には，どのような性能の保証をもたせる

かによって種々の方法がある．例えば，最尤推定量 $n[j,s,g]/n[s,g]$，確率パラメータの事前分布を一様分布とおいた推定量 $(n[j,s,g]+1)/(n[s,g]+J(s,g))$，確率パラメータの事前分布を Dirichlet 分布とおいた推定量 $(n[j,s,g]+1/2)/(n[s,g]+J(s,g)/2)$ などがある[9]．

定義2 冗長度 長さ n の系列（確率変数）Z^n の各実現値 z^n に対して，以下の値を言語 L を用いたときの真の知識 K に対しての冗長度 (redundancy)[10] という．

$$\sum_{z^n\in Z^n} r(z^n|K)[length_L(z^n)-\{-\log r(z^n|K)\}] \tag{7}$$

ここで，$r(z^n|K)$ は知識 K を前提にした系列 z^n の起こる確率とした．冗長度は，知識 K が未知であるときの平均の記述長と既知であるときの平均の記述長との差異である．系列長 n を十分に大きくとることによって，1入出力あたりの冗長度は0に収束する．冗長度を推定誤差に用いる学習基準は，系列をより短く圧縮できることを確率的な関係を学習できたことの帰結とする基準ということができる．

また，(7)を知識 K の事前確率 $\pi(K)$ で平均した値(8)を学習アルゴリズムの平均冗長度という．

$$\sum_K \pi(K)\sum_{z^n\in Z^n} r(z^n|K)[length_L(z^n) -\{-\log r(z^n|K)\}]\geq 0 \tag{8}$$

一般には，モデル g の記述長およびモデル g のもとでの系列の記述長（モデル g のもとで確率パラメータの事前確率が反映される）は，言語 L が想定した事前確率（プライア）$\pi'(K)$ に基づいて決定される．各知識 K についての事前確率 $\pi(K)$ が既知であれば（すなわち，事前確率 $\pi(K)$ がプライア $\pi'(K)$ と一致していれば），任意の n に対して平均冗長度が最小になり，trivial な問題となる．事前確率が未知である一般的な場合については，次の結果が得られている．

補題2[1] アルゴリズム1は，入出力対の形式になっている知識 K を学習する問題について，C_2 を n によらない定数として定義2で示された冗長度を $\{k(g_K)/2\}\log n+C_2$ まで低減させる．

ここで，種々の推定誤差が考えられるが，各々を用いる絶対的な理由はないものと思われる．ここでは，事前確率が未知の場合でも補題2の性能が保証されるという点で，冗長度を推定誤差に用いた．

一方，学習基準は多くの場合，Kullback-Leibler 情報量☆，Hellinger 距離，2乗距離等，真の知識 K と長さ n の訓練例系列から学習した知識 $\hat{K}$ の間の距離 $d(K,\hat{K})$ で定義される[3),11)]．補題2の推定誤差は，この意味で距離の形式にはなっていないが，記述長最小基準の基本性能（真の知識への収束性[3),12)]）を示す上で重要であり，各距離の性能を保証する基礎をなすものである．

仮定3 訓練例を得る前の各知識の事前確率は未知である．

2.2 Dendroid 分布近似

分布が既知である多次元の確率分布

$$P(x^R)=P(x^{(1)},x^{(2)},\cdots,x^{(R)}) =\prod_{N=1}^{R} p(x^{(N)}|x^{(0)}x^{(1)}\cdots x^{(N-1)}) \tag{9}$$

を Dendroid 分布[6]

$$P'(x^R)=\prod_{N=1}^{R} p(x^{(N)}|x^{(q[N])}),\quad 1\leq q[N]\leq N-1,\quad q[1]=0 \tag{10}$$

で近似する問題を検討する．ここで，$x^{(0)}$ は空の属性値で，属性 $p(x^{(N)}|x^{(0)})$ は他の属性とは独立に発生する $x^{(N)}(N=1,2,\cdots,R)$ の確率分布であるとした．すなわち，属性 $N=2,3,\cdots,R$ は本来であれば他の $N-1$ 属性に依存するが，高々1属性にだけに依存させるのが Dendroid 分布近似である．直観的には**図1**のように依存関係が（ループをつくらない）木で表現される．このとき各 $N=2,3,\cdots,R$ について，$1\leq q[N]\leq N-1$ をいかにおくかによって $(R-1)!$ 通りの近似が考えられる．C. K. Chow らは両者の Kullback-Leibler 情報量[13]

$$\begin{aligned} D(P\|P')&=\sum_{x^R}P(x^R)\log\frac{P(x^R)}{P'(x^R)}\\ &=-\sum_{x^R}P(x^R)\sum_{N=1}^{R}\log p(x^{(N)}|x^{(q[N])})\\ &\quad+\sum_{x^R}P(x^R)\log P(x^R)\\ &=-\sum_{N=2}^{R}I(X^{(N)},X^{(q[N])})+\sum_{N=1}^{R}H(X^{(N)})\\ &\quad+\sum_{x^R}P(x^R)\log P(x^R) \end{aligned} \tag{11}$$

を最小にする Dendroid 分布を効率よく探索するために，相互情報量 $I(X^{(N)},X^{(q[N])})$ を枝のコストにおく，コスト最小極大木アルゴリズム[14]を提案している[7]．ここで，以下のようにおいた．

$$I(X^{(q)},X^{(q')}) =-\sum_{x^{(q)},x^{(q')}}p(x^{(q)},x^{(q')})\log\frac{p(x^{(q)}|x^{(0)})}{p(x^{(q)}|x^{(q')})} \tag{12}$$

☆ Kullback-Leibler 情報量は，距離の公理を満足していない．

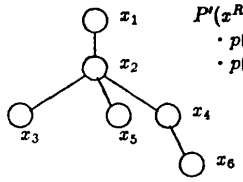

図1 木 の 例
Fig. 1 An example of a tree.

$$H(X^{(q)})=\sum_{x^{(q)}} p(x^{(q)}|x^{(0)})\log p(x^{(q)}|x^{(0)}) \quad (13)$$

ここで，(11)式の第2項，第3項の値は一定値である．

具体的な Dendroid 分布は，アルゴリズム2で最終的に木 T のデータ構造によって，モデルが表現される．また各確率パラメータについては，この依存関係を用いて容易に計算できる．アルゴリズム2では，順序付キュー Q はあくまで作業領域であり，集合 T が最終的な出力になる．

アルゴリズム2

begin

1. $T:=\{\}$;
2. (q,q') のすべての辺に対して，相互情報量 $I(X^{(q)},X^{(q')})$ を計算して，この値の降順に分類して，順序付キュー Q に格納する；
3. $q=1,2,\cdots,R$ に対して，集合 $\{q\}$ の集合を VS とおく；
4. **while** $\|VS\|>1$ **do**
   begin
   (a) Qの中で相互情報量 $I(X^{(q)},X^{(q')})$ を最大にする辺 (q,q') を取り除く；
   (b) **if** 頂点 q と頂点 q' とが VS の異なる集合 W_1 と W_2 とに属する **then**
   begin
   i. VS の中の集合 W_1 と W_2 とを $W_1\cup W_2$ でおきかえる；
   ii. 辺 (q,q') を集合 T に加える
   end
   end

end.

アルゴリズム2を要約すると以下のようになる．各属性は各頂点に対応する．相互情報量の大きい順に各頂点を辺で結んでいく．このとき，ループを形成させる辺は結ばないようにする（同じ集合に属すか否かの検査は，この条件の検査に対応している．）．最終的には，どの頂点どうしも一組の複数の辺をたどることによって結ばれることになる．アルゴリズム2は，相互情報量の大きい辺から単純に結んでいるように思われるが，Kullback-Leibler 情報量[13] $D(P\|P')$ の最適解を必ず見つけられるアルゴリズムになっている．

また，C. K. Chow らは同じ論文[7]で，有限長の訓練例系列から Dendroid 分布の範囲でモデルを選択する問題も扱っている．しかしこのアルゴリズムでは相互情報量 $I(X^{(q)},X^{(q')})$ を，(系列から推定した) 相互情報量

$$\hat{I}(X^{(q)},X^{(q')})=-\sum_{x^{(q)},x^{(q')}}\hat{p}(x^{(q)},x^{(q')})\log\frac{\hat{p}(x^{(q)}|x^{(0)})}{\hat{p}(x^{(q)}|x^{(q')})} \quad (14)$$

におきかえて，確率パラメータの数が一定 ($2N-1$ 個) のモデルの集合から最尤なモデルを選択することになる．ここで，$\hat{p}(\cdot|\cdot)$ (あるいは，$\hat{p}(\cdot,\cdot)$) を，訓練例の相対頻度から得られた各 $p(\cdot|\cdot)$ (あるいは，$p(\cdot,\cdot)$) の最尤推定量とした．本論文で提案する学習アルゴリズムは，このアルゴリズムの拡張になっている．

2.3 学習する対象の広さと学習の計算量

本論文の目的は，入出力関係の学習を属性値関係という一般的な場合に拡張し，記述長最小基準の適用範囲を広げることにある．しかしながら，学習する対象を広げれば学習のための計算量が莫大になることが予期できる．したがって，本来なら学習する対象の広さと学習の計算量の両面で評価する必要があると考えることができる．本論文では，その一例としてDendroid分布を前提にすることによって学習する対象を狭める問題を考察する．具体的には，次の2項目を検討する．

1. 状態分割の概念[1),15)]から，入力 x_i^{N-1} から出力 y_i の入出力の関係という特別な設定を排除し，多次元の属性値ベクトル x_i^N の要素間の因果関係についての知識を学習できないだろうか．
2. アルゴリズム2を有限長の訓練例系列から出発させても[7)]，確率パラメータの数が一定であり，その範囲で最も尤度の高いモデルを選択するしかない．このオーバフィッティングの問題[8)]を解決するために，記述長最小基準を用いて集合Tの最終結果が全頂点を含むひとつの木だけではなく，一部の頂点からなる複数の木(森)を表現する一般的な形式に修正できないだろうか．また，その結果として 1. の特別な場合を導くことができないだろうか．

3. 確率的な属性値関係の学習

3.1 学習の計算量に関する考察

時点$-n+1$から0まで$(n=1,2,\cdots)$の長さnの実際の属性値ベクトルの系列

$$z_{-n+1}=(x_{-n+1}^{(1)}, x_{-n+1}^{(2)}, \cdots, x_{-n+1}^{(R)})$$
$$z_{-n+2}=(x_{-n+2}^{(1)}, x_{-n+2}^{(2)}, \cdots, x_{-n+2}^{(R)})$$
$$\cdots \quad \cdots \quad \cdots$$
$$z_0=(x_0^{(1)}, x_0^{(2)}, \cdots, x_0^{(R)})$$

(z^nで表記する)から，将来の属性値ベクトルの系列

$z=z_1z_2\cdots$, $z_i=x_i^R=(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(R)})$, $R=1,2,\cdots$

の知識を学習する．具体的には，属性値ベクトルx_i^Rの確率分布$P(x_i^R)$を推定する．

ここでは，次の仮定を満足する問題に限定して，以下の検討にはいる．

仮定4 属性値ベクトルx_i^Rの各要素$x_i^{(q)}$ $(q=1,2,\cdots,R)$の値が，有限集合$A=\{0,1,\cdots,J-1\}$のいずれかの値をとる．

まず，提案アルゴリズムから示そう．

アルゴリズム3

begin

for $N=1$ **to** R **do**

1. $y_i:=x_i^{(N)}, i=-n+1,-n+2,\cdots,0$;
2. Algorithm 1 $(N-1)$

end.

アルゴリズム1では，Algorithm 1(N)を1回だけ適用して入出力関係を表現するモデルを選択した．アルゴリズム3では，Algorithm 1$(N-1)$, $N=1,2,\cdots,R$を順次適用し，入出力関係を表現するモデルを順次選択している．アルゴリズム3が結果的に属性値関係を表現するモデルを選択していることは，属性値ベクトルx^Rの分布が$\prod_{N=1}^{R} p(x^{(N)}|x^{(0)}x^{(1)}\cdots x^{(N-1)})$で表現されることからも明らかである．すなわち，$x^{(0)}x^{(1)}\cdots x^{(N-1)}$を入力，$x^{(N)}$を出力にする入出力関係の学習を$N=1,2,\cdots,R$の範囲で繰り返しているのである．

このとき，アルゴリズム3におけるモデルの候補の数の各段階の和および全比較回数を各々$M(J,R)$, $\tilde{M}(J,R)$としたときに次の定理が成立する☆．

定理1 各モデル$g\in G$に対して，各段階$N=1,2,\cdots,R$および各状態$s=1,2,\cdots,S(N,g)$で$J(s,N,g)=J$としたときに，

$$M(J,R)=\sum_{N=1}^{R} f[J^{N-1}] \tag{15}$$

$$\tilde{M}(J,R)=\sum_{N=1}^{R}\{f[J^{N-1}]-1\}=\sum_{N=1}^{R} f[J^{N-1}]-R \tag{16}$$

ここで，関数$f[m]$はm個の要素をもつ集合の要素を任意数$S=1,2,\cdots,m$の状態$s=1,2,\cdots,S$に分割する組合せの数を意味し，次のように定義できる．

$$f[m]=\sum_{S=1}^{m}\sum_{T=1}^{S}\frac{T^m(-1)^{S-T}}{(S-T)!\,T!} \tag{17}$$

(証明は付録を参照のこと．)

ところで，このときの記述長の最小値は，アルゴリズム3の各段階$N=1,2,\cdots,R$について，独立に計算した最小値の和と一致する．したがって，次の定理が成立する．

定理2 $C_1'=RC_1$として，各モデル$g\in G$に対して，長さnの訓練例の系列z^nを高々以下の長さで記述する言語L'が存在する．ここで，各段階$N=1,2,\cdots,R$の各状態$s=1,2,\cdots,S(N,g)$のもとで最終的に群$j=0,1,\cdots,J(s,N,g)-1$が生起した頻度(各群は等確率で生起する$m[j,s,N,g]$個のクラスを含む)，および各状態sが生起した頻度をそれぞれ$n[j,s,N,g]$, $n[s,N,g]$とおいた．

$$length_{L'}[g](z^n)=H[g](z^n)+\frac{k(g)}{2}\log n+C_1' \tag{18}$$

$$k(g)=\sum_{N=1}^{R}\sum_{s=1}^{S(N,g)}[J(s,N,g)-1] \tag{19}$$

$$H[g](z^n)=\sum_{N=1}^{R}\sum_{s=1}^{S(N,g)}\sum_{j=0}^{J(s,N,g)-1} -n[j,s,N,g]\cdot\log\frac{n[j,s,N,g]}{m[j,s,N,g]n[s,N,g]} \tag{20}$$

(証明略)

定理3 $C_2'=RC_2$として，アルゴリズム3は，仮定1～仮定4を同時に満足する多次元の属性値ベクトルの要素間の知識Kを学習する問題について，冗長度を$\{k(g_K)/2\}\log n+C_2'$まで低減させる．

(証明) 多次元の属性値の場合の冗長度は，記述長を各段階で独立に計算できることから，入出力関係の場合の冗長度の累積になる．各段階$N=1,2,\cdots,R$で補題2を適用し，この値は$\sum_{N=1}^{R}\left\{\sum_{s=1}^{S(N,g_K)}[\{J(s,N,g_K)-1\}/2]\log n+C_2\right\}$となる．

(証明終)

☆ 通常は，計算量低減をねらいとしてモデルの候補を事前に限定する．

従来は，特定の入出力対の知識のみを学習することが可能であったが，この一般化によってより広いクラスの知識，すなわち，属性値ベクトルの次元R以下の任意数の属性間の因果関係を学習することができる．

しかしながら，定理1の結果は，本論文の主たる結論ではない．むしろ，確率的な属性値関係の学習ではこのように膨大な計算量が必要になる，という問題を提起しているのである．したがって，学習の対象を若干狭めても，学習の計算量を十分に低減させることができれば，大変ありがたいのである．

3.2 Dendroid 分布近似を仮定した場合の検討

ここでは，さらに次の2つの仮定を満足する問題に限定して，以下の検討にはいる．

仮定5 モデル $g\in G$ の各段階 $N=1,2,\cdots,R$ および各状態 $s=1,2,\cdots,S$ に対して，

$$J(s,N,g)=J=2 \tag{21}$$

仮定6 各モデル $g\in G$ の各段階 $N=1,2,\cdots,R$ に対しての各状態分割は，高々他の1属性の値による．

まず，アルゴリズムから示そう[16]．アルゴリズム4では，順序付キュー Q はあくまで作業領域であり，集合 T が最終的な出力になる．

アルゴリズム4

begin

1. $T:=0$;
2. (q,q') のすべての辺に対して，相互情報量 $\hat{I}(X^{(q)},X^{(q')})$ を計算して，この値の降順に分類して，順序付キュー Q に格納する；
3. $q=1,2,\cdots,R$ に対して，集合 $\{q\}$ の集合を VS とおく；
4. **while** Q の中の相互情報量 $\hat{I}(X^{(q)},X^{(q')})$ の最大値$>(\log n)/2n$ **do**

begin

(a) Q から相互情報量 $\hat{I}(X^{(q)},X^{(q')})$ を最大にする辺 (q,q') を取り除く；

(b) **if** 頂点 q と頂点 q' とが VS の異なる集合 W_1 と W_2 とに属する **then**

begin

i. VS の中の集合 W_1 と W_2 とを $W_1\cup W_2$ でおきかえる；

ii. 辺 (q,q') を集合 T に加える

end

end

end.

定理4 アルゴリズム4は，仮定1～仮定6を同時に満足する，多次元の属性値ベクトルの知識Kを学習する問題について，記述長を最小にするモデルを選択する．

(証明) 結果は，定理3の特別な場合となる．したがって，アルゴリズム3に仮定5，仮定6をおいた特別な場合がアルゴリズム4になることを示せば十分である．

段階 $N=2,3,\cdots,R$ では，一般には $x^{(N)}$ の値の確率分布が $x^{(1)}x^{(2)}\cdots x^{(N-1)}$ の $N-1$ 個の属性値に依存することになるが，仮定6より $x^{(N)}$ の値は高々1属性 $1\le q[N]\le N-1$ の $x^{(q[N])}=0,1$ に依存して決定される．

まず，属性 N が他の属性 $1\le q[N]\le N-1$ と辺を結ぶ場合を考える．仮定5より属性値 $x^{(q[N])}$ は0，1の2値をとり得る $(S(N,g)=2)$ ので，$x^{q[N]}=0,1$ に対応して状態 $s=1,2$ をおく．また，仮定5より属性値 $x^{(N)}$ も0，1の2値をとり得る $(J(s,N,g)=2)$ ので，$x^{(N)}=0,1$ に対応して群 $j=0,1$ をおく．このとき，確率パラメータの数は，

$$\sum_{s=1}^{S(N,g)}[J(s,N,g)-1]=2 \tag{22}$$

となる．一方，仮定5より

$$m[j,s,N,g]=1 \tag{23}$$

また，$\hat{p}(\cdot|\cdot)$ および $\hat{p}(\cdot,\cdot)$ の定義から

$$\frac{n[j,s,N,g]}{n[s,N,g]}=\hat{p}(x^{(N)}|x^{(q[N])}) \tag{24}$$

$$\frac{n[j,s,N,g]}{n}=\hat{p}(x^{(N)},x^{(q[N])}) \tag{25}$$

が成立するので，

$$\begin{aligned}
&\sum_{s=1}^{S(N,g)}\sum_{j=0}^{J(s,N,g)-1}-n[j,s,N,g]\log\frac{n[j,s,N,g]}{m[j,s,N,g]n[s,N,g]}\\
&=n\sum_{x^{(N)}=0}^{1}\sum_{x^{(q[N])}=0}^{1}-\hat{p}(x^{(N)},x^{(q[N])})\log\hat{p}(x^{(N)}|x^{(q[N])})\\
&=n\sum_{x^{(N)},x^{(q[N])}}\hat{p}(x^{(N)},x^{(q[N])})\cdot\log\frac{\hat{p}(x^{(N)}|x^{(0)})}{\hat{p}(x^{(N)}|x^{(q[N])})}\\
&\quad-n\sum_{x^{(N)}}\hat{p}(x^{(N)}|x^{(0)})\log\hat{p}(x^{(N)}|x^{(0)}) \qquad (26)\\
&=-n\hat{I}(X^{(N)},X^{(q[N])})+n\hat{H}(X^{(N)}) \qquad (27)
\end{aligned}$$

が導出される．ここで，

$$\hat{H}(X^{(N)})=-\sum_{x^{(N)}}\hat{p}(x^{(N)}|x^{(0)})\log\hat{p}(x^{(N)}|x^{(0)}) \tag{28}$$

とおいた.

次に，属性 N が1から $N-1$ までの属性と辺を結ばない場合（$q[N]=0$ の場合）を考える．$x^{(N)}$ の値は，1から $N-1$ までの属性の値とは無関係に生起する（$S(N,q)=1$）ので，確率パラメータの数は，

$$\sum_{s=1}^{S(N,q)}[J(s,N,q)-1]=1 \tag{29}$$

となる．一方，(26)で $q[N]=0$ とおいて

$$\sum_{s=1}^{S(N,q)}\sum_{j=0}^{J(s,N,q)-1}-n[j,s,N,q]\log\frac{n[j,s,N,q]}{m[j,s,N,q]n[s,N,q]}=n\hat{H}(X^{(N)}) \tag{30}$$

が導出される.

以上，$q[N]\neq 0$，$q[N]=0$ の各々の場合と考察，および定理2から，記述長 $length_{L'}(x^n)$ は，

$$length_{L'}=-n\sum_{q^{(N)}\neq 0}\hat{I}(X^{(N)},X^{(q[N])})+n\sum_{N=1}^{R}\hat{H}(X^{(N)})+\frac{k(q)}{2}\log n+C_1' \tag{31}$$

$$k(q)=N+(q[N]\neq 0 \text{ なる属性の数}) \tag{32}$$

で計算できる.

(31)の第2項の値は，記述長の計算とは無関係に一定なので，辺 $(q[N],N)$ を結ぶか否かを決定する際には

$$-\hat{I}(X^{(N)},X^{(q[N])})+\frac{\log n}{2n} \tag{33}$$

の値を見てやればよい.

（証明終）

アルゴリズム2（訓練例から出発させた場合）が確率パラメータ数が $k(q)=2N-1$ なるモデル q の集合に適用可能であるのに対し，提案アルゴリズムが $N\leq k(q)\leq 2N-1$ の拡張されたモデル q の集合に適用可能であることに注意したい.

Dendroid 分布近似の問題を有限個の訓練例系列から出発させる場合に，種々の方法が考えられる．例えば，(11)の値の不偏推定量を計算し，この値の最小化をはかる方法（AIC[17]）[18]が考えられる．AIC 自身有効な性質を満足している[19]が，学習の問題に関する限り Consistency を満足していない点が致命的である．ここで，任意の n に対して，真のモデルを誤って選択する確率を一定値以下にできないことを，consistency を満足していないといっている.

表1 各属性値間の相互情報量 $\hat{I}(X^{(q)},X^{(q')})$
Table 1 Mutual information between attribute values.

属性値 q ＼ 属性値 q'	1	2	3	4	5	6	7
1		0.0174	**0.1930**	0.0375	0.0331	0.0807	0.0402
2			0.0800	0.0037	0.0102	**0.1230**	0.0115
3				0.0512	0.0097	**0.0957**	0.0476
4					0.0063	0.0512	**0.1580**
5						**0.1230**	0.0293
6							0.0206
7							

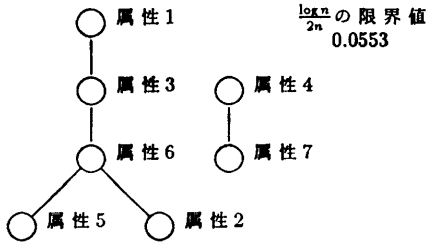

図2 適用例の結果
Fig. 2 Results of application example.

4. 適用例

以下，アルゴリズム4に実データを入れて，適用例を示すことにする.

大学生31人（$n=31$）に対して，次のアンケート調査を行った.

属性1 ウィスキーが好きか？ yes no
属性2 ビールが好きか？ yes no
属性3 ワインが好きか？ yes no
属性4 日本酒が好きか？ yes no
属性5 焼酎が好きか？ yes no
属性6 酎ハイが好きか？ yes no
属性7 カクテルが好きか？ yes no

この調査結果を提案アルゴリズムで解くと以下のような結果となる．相互情報量が $(\log n)/2n=0.0553$ を越えている属性値の組 $(q[N],N)$ $(1\leq q[N]\leq N-1)$ を相互情報量の大きい順に挙げると，（属性1，属性3），（属性4，属性7），（属性2，属性6），（属性5，属性6），（属性3，属性6），（属性1，属性6）のようになる（表1）．しかし，（属性1，属性6）を結ぶとループをつくるので，この辺だけは結ばない.

したがって，（属性1，属性3），（属性4，属性7）（属性2，属性6），（属性5，属性6），（属性3，属性6）に従属関係があるといえる．この様子を図2に

示す.

5. おわりに

以上，入出力関係を仮定しない一般的な確率的な因果関係について，記述長最小基準の立場から知識の学習方法について検討した．このために，まず文献1)で示した入出力関係における状態分割の概念が，属性値関係に対しても有効であることを示した．そして，学習の対象の広さと学習の計算機の間の相互関係を議論した上で，学習の対象をほとんど狭めることなく学習の計算量を大幅に低減できるアルゴリズムを提示した．

今後の課題として，Dendroid 分布近似のようにモデルに制約を加えて計算量を低減させる場合に，少ないモデルのいずれかに近似することの誤差とモデルを比較する計算量との間の関係をさらに明確にすることがあげられる．また，本論文では，真の知識と長さ n の訓練例系列から学習された知識 $\hat{K}$ の間の距離 $d(K, \hat{K})$ の性能については言及していない．しかし，この問題は学習結果の性能の保証という立場からは重要であり，いずれ検討する予定である．

謝辞 有益なご意見をいただいた，本学理工学研究所特別研究部会「経営工学における知識情報処理に関する研究」(責任者石渡徳彌教授) の諸氏に感謝します．また，一部検討に加わっていただいた湘南工科大学稲積宏誠博士に感謝します．

参考文献

1) Suzuki, J.: Generalization of the Learning Method for Classifying Rules with Consistency Irrespective of the Representation Form and the Number of the Classified Patterns, *ISITA 90*, Waikiki, Hawaii, pp. 495-498 (Nov. 1990).
2) Haussler, D.: Decision Theoretic Generalizations of the Pac Learning Model, *The 1st Workshop on Algorithmic Learning Theory*, Tokyo, pp. 21-41 (Oct. 1990).
3) Yamanishi, K.: A Learning Criterion for Stochastic Rules, *The 3rd Workshop on Computational Learning Theory*, Morgan Kaufman, pp. 67-81 (Aug. 1990).
4) Rissanen, J.: Universal Coding, Information, Prediction, and Estimation, *IEEE Trans. Inform. Theory*, Vol. IT-30, No. 4, pp. 629-636 (1984).
5) Rissanen, J.: Stochastic Complexity and Modeling, *The Annals of Statistics*, Vol. 14, No. 3, pp. 1080-1100 (1986).
6) 堀部安一：情報エントロピー論，森北出版 (1989).
7) Chow, C. K. and Liu, C. L.: Approximating Discrete Probability Distributions with Dependence Trees, *IEEE Trans. Inform. Theory*, Vol. IT-14, No. 3, pp. 462-467 (1968).
8) 伊藤秀一：ユニバーサル量子化の情報源モデルと符号化，情報理論とその応用学会第10回シンポジウム資料，pp. 611-616 (1987).
9) Suzuki, J.: *MDL Principle—From Viewpoints of Minimax Redundancy—*, Hokkaido, pp. 40-48 (Aug. 1991).
10) Davisson, L. D.: Universal Noiseless Coding, *IEEE Trans. Inform. Theory*, Vo. IT-19, No. 6, pp. 783-795 (1973).
11) Abe, N. and Warmuth, M.: On the Computational Complexity of Approximating Distributions by Probabilistic Automata, *The 3rd Workshop on Computational Learning Theory*, Morgan Kaufman, pp. 52-56 (Aug. 1990).
12) Barron, A. R.: Logically Smooth Density Estimation, Ph.D. thesis, Department of Electrical Engineering, Stanford University (1985).
13) Kullback, S.: *Information Theory and Statics*, John Wiley and Sons, New York, Chapman and Hall, London (1959).
14) Aho, A. V., Hopcroft, J. E. and Ullman, J. D.: *The Design and Analysis of Computer Algorithm*, Chap. 5, Addison-Wesley (1974).
15) 鈴木 譲：有限系列に対しての情報源の推定方法の改善とモデル学習への応用，第12回情報理論とその応用シンポジウム資料，pp. 515-520 (1989. 12).
16) 大嶽康隆，中條 健，鈴木 譲，平澤茂一：概念学習に関する一考察，電子情報通信学会情報理論研究会資料，IT 90-49 (1990. 7).
17) Akaike, H.: A New Look at the Statistical Model Identification, *IEEE Trans. Automatic Control*, Vol. AC-19, No. 6, pp. 716-723 (1974).
18) 篠原靖志：情報量の見積による帰納学習，人工知能学会第3回全国大会資料，pp. 95-98 (1989. 7).
19) Shibata, R.: Selection of the Order of an Autoregressive Model by Akaike's Imformation Criterion, *Biometrica*, Vol. 63, pp. 117-126 (1976).

付録 定理1の証明

まず，状態数を S に固定する．m 個の要素が S 個の状態のいずれかに含まれ，各状態 $s=1, 2, \cdots, S$ にいずれも1個以上の要素が含まれる組合せ $f_S[m]$ は，

$$f_S[m]=\sum_{T=1}^{S}\binom{S}{T}T^m(-1)^{S-T} \tag{34}$$

となる．実際，m 個の各要素が S 以下の状態に m 個の要素が含まれる組合せ

$fs[m, S]=S^m$

を考えると，$fs[m, S]$ には $S-1$ 以下の状態に m 個の要素が含まれる組合せ

$$fs[m, S-1]=\binom{S}{S-1}(S-1)^m$$

が重複して加算されている．さらに，$fs[m, S-1]$ には $S-2$ 以下の状態に m 個の要素が含まれる組合せ

$$fs[m, S-2]=\binom{S}{S-2}(S-2)^m$$

が重複して加算されている．以上を繰り返して次式から(34)が得られる．

$$\begin{aligned} fs[m]&=fs[m, S]-(fs[m, S-1]\\ &\quad -(fs[m, S-2]\cdots(fs[m, 2]\cdots(fs[m,\\ &\quad 1]))))\\ &=\sum_{T=1}^{S} fs[m, T](-1)^{S-T} \end{aligned} \tag{35}$$

そして，関数 $f[m]$ の値を求めるためには，各状態の順序は区別しないので $fs[m]$ の値を $S!$ で除し，$S=1, 2, \cdots, m$ で加えてやればよい．したがって，次式から(17)が得られる．

$$f[m]=\sum_{S=1}^{m} \frac{fs[m]}{S!} \tag{36}$$

また，各段階でモデルの候補の数から1を減じた値がモデルを選択する比較の回数になるので，(16)が成立する．

(平成3年5月20日受付)
(平成4年9月10日採録)

鈴木 譲

1960年生．1984年早稲田大学理工学部工業経営学科卒業．1986年同大学院理工学研究科修士課程修了．1989年同大学院理工学研究科博士課程単位取得退学．同年同大学理工学部工業経営学科助手．1992年青山学院大学理工学部経営工学科助手．MDL（記述長最小基準），計算論的学習理論，ユニバーサルデータ圧縮，最近は遺伝的アルゴリズムの理論的解析に没頭している．電子情報通信学会，人工知能学会，日本経営工学会，情報理論とその応用学会各会員．

大嶽 康隆

1966年生．1990年早稲田大学理工学部工業経営学科卒業．1992年同大学院理工学研究科修士課程修了．同年(株)東芝入社．在学中知識情報処理，特に計算論的学習理論の研究に従事，現在に至る．

平澤 茂一（正会員）

1938年生．1961年早稲田大学第一理工学部数学科卒業．1963年同電気通信学科卒業．同年三菱電機(株)入社．1963年同電気通信学科卒業．1981年早稲田大学理工学部工業経営学科教授・工学博士．この間，1978年米UCLA計算機科学科客員研究員．1985年ハンガリー科学アカデミー，伊トリエステ大学客員教授．データ伝送方式，計算機応用，情報システムの開発，ならびに情報理論とその応用の研究に従事．電子情報通信学会，IEEE各会員．共著「理工系のための計算機工学」(昭晃堂)．