ID3-GAによる非独立性属性データ問題へのアプローチ

本堂直浩[†] 成瀬継太郎[††] 嘉数侑昇[††]

ID3は，分類のための決定木生成アルゴリズムとして知られているが，データの性質によっては分類精度が低下する特性を持つ．特に，データ中の属性がセマンティクス的に非独立である場合，その傾向は顕著である．本論ではこれに対し従来の情報エントロピーによる属性情報の獲得のほかに，属性のセマンティクスを表現する重み付けパラメータを設定し，決定木の生成を行うID3-GAによるアプローチを行っている．特に，本論ではこの最適パラメータの決定には遺伝的アルゴリズムを用いている．ここでは計算機実験において，非独立性属性を含むデータに対し実験を行い，ID3-GAの有効性，挙動，特性の検証を行っている．

ID3-GA for the Dependent Attribute Problem

NAOHIRO HONDO,[†] KEITAROU NARUSE[††] and YUKINORI KAKAZU[††]

ID3 is an algorithm of making decision trees for classification and has a wide use and a high quality mechanism of concept acquisition. However, it has some features that the classificational accuracy becomes lower in some data. Especially, if the data have dependence attributes on semantics, the tendency becomes more conspicuous. In this paper, we propose a new parameter that shows conceptual weight of attribute, in addition to conventional information entropy criterion. The parameter works to make trees with the entropy in ID3. The optimal parameters are acquired using genetic algorithms in this paper. Experiments show that proposed ID3 makes it possible to classify the data more carefully. Finally, the nature of approach is discussed through several experiments.

1. はじめに

ID3[1]は属性データからなる知識を木構造へと抽象化し，知識の構造化，属性の重要度判定，およびクラスへの分類を行う帰納学習・分類システムとして知られ，様々な工学応用が行われているが，トレーニングデータ中のイルプロブレムな属性データ特性が存在し，分類精度低下の一因になっていることが課題として指摘されている[1),2)]．後述するように，この特性に起因する多くの問題は，それぞれの問題向きに拡張されたID3ファミリーにより解決されてきた[1)~6)]．しかし，トレーニングデータの中にセマンティクス的に非独立な属性が存在する場合は未解決の課題として残っている[2)]．この課題に対し本論では，データの概念的な性質を表す重み付けパラメータを付加したID3-GAによるアプローチを試みる．このパラメータの調整には，組み合わせ最適化問題において高い性能を示す遺伝的アルゴリズム（Genetic Algorithms; GA）[7)]を用いる．

帰納学習の一手法であるID3は，過去の事例に基づいて決定木を生成し，この決定木を用いてクラスの決定していない事例の分類を行うものである．すなわち，ID3は，事例群であるトレーニングデータの一般化を行い，その結果をツリー構造を用いて表現することにより，未知データの分類を可能とする．この決定木生成過程においてID3は，事例群のクラス分布から情報理論に基づき各属性の情報量を計算し，節点のラベルとして属性を選択する．選択された属性である節点から次の節点への枝が伸び，そこで同様の手続きにより再帰的に節点が決定され，決定木が成長する．このID3のトップダウン的な生成戦略，およびセット学習戦略の結果として，ID3は対象データの特性に対して敏感となり，しばしば分類精度が低下することが指摘されている[1),2)]．そのため，対象データが有する様々な特徴に応じて分類精度の向上のための拡張が行われ

† 北海道大学工学部情報工学科
Division of Information Engineering, Hokkaido University
†† 北海道大学工学部精密工学科
Department of Precision Engineering, Hokkaido University

ている．たとえば，ノイズに対する拡張[1)~3)]，インクリメンタルな環境に対しての拡張[4)]，あるいは，高い分類精度と計算コストの減少の両立が求められる環境に対しての拡張[5),6)]などがすでに存在し，それぞれ有効性が確認されている．しかしながら，非独立性属性を含むデータ特性に対するID3の弱点克服に関する有効な手法はまだ明らかにされていない．ここでいう非独立性属性とは，複数の属性がセマンティクス的に従属関係を持つ属性であり，クラスに関する情報を獲得するには，これらの属性を複合的に観察する必要があるものである．現実的にこのような属性がデータに存在すると考えるのは自然であるが，ID3における決定木の生成では，機構上扱う属性はセマンティクス的に独立であることを前提にしているため，ID3が非独立性属性を含む対象データの分類を行う際には，属性選択機構が十分に機能せず，データの一般化に失敗し，結果として分類精度は低くなる可能性が高くなる．

ここで問題とされているのは，非独立性属性の属性の取捨選択での取り扱いであるが，これは特徴選択問題[8)~11)]と密接に関連する．特徴選択問題では，様々なアプローチにより節点の属性選択のための議論がなされてきた．しかしながら，非独立性属性群の扱いに関しては行われていない．

以降では，ID3を概観し，上述のデータ特性問題への一アプローチとしてGAを用いたID3-GAの提案を行う．最後に各手法に関する代表的問題に対する計算機実験を行い，提案手法の有効性および，特性を考察する．

2. ID3

本章では，決定木について説明を行い，その後ID3のアルゴリズムを詳論する．

2.1 決定木

決定木生成のためのトレーニングデータ E は，各事例のカテゴリー（クラス）が既知である事例集合で構成される．いま事例 E_i を表現するために用いる属性を A_i，その数を n，総事例 N をとした場合，トレーニングデータ，事例はそれぞれ，

$$E = \{E_1, E_2, \ldots, E_N\}, \tag{1}$$

$$E_i = \{A_1^i, A_2^i, \ldots, A_N^i\}, \tag{2}$$

となる．属性 A_j^i は，一般に個の属性値と呼ばれる要素を持つ集合として規定される．

$$A_j^i = \{a_1^i, a_2^i, \ldots, a_{mj}^i\}. \tag{3}$$

また，ID3が事例データの一般化表現に用いる決定木は，次のように定式化される[2)]．

- 端点である葉は，事例の属すカテゴリーであるクラスを規定する．
- 内部節点は属性を表し，その属性の有する複数の属性値のそれぞれに対してサブツリーに結合する枝を持つ．枝の先は端点か節点である．

2.2 ID3アルゴリズム

ID3で用いるアルゴリズムの概略をPROCEDURE ID3として**図1**に示す．

図1において事例データ集合の部分集合 S を，その副集合 T を，データ集合 S の中のクラス C_j である事例データ数 $freq(C_j, S)$ をとする．このとき，

(1) すべての事例データが1つのクラスに含まれている場合，その決定木，もしくはそのサブツリーはそのクラスを示す端点のみの木となる．

(2) (1)ではない場合，

(a) ゲイン法によって端点となる属性の選択を行う．各属性の情報量，$gain(A)$ を求め，$gain(A)$ が最も高い属性が端点として選択される．ただし，

$$gain(A) = info(T) - info_A(T), \tag{4}$$

$$info_A(T) = \sum_{k=1}^{N} \frac{|T_k|}{|T|} \times info(T_k), \tag{5}$$

```
PROCEDURE ID3

BEGIN
make tree(current node)
IF freq(C_j, S) = 0
    then label the leaf with the class
ELSE
    FOR(i; each attribute)
        info_X(T) = Σ_{i=1}^{N} (|T_i| / |T|) × info(T_i)
        info(S) = Σ_{j=1}^{m} (freq(C_j, S) / |S|) × log_2 ( freq(C_j, S) / |S| )
        gain(X) = info(T) − info_X(T)
    Selected attribute A = max_{k=1,2,...} (gain(k))
    FOR (i; each attribute value of attribute A)
        make branchof each i
        make tree(next node)
END
```

図1 ID3アルゴリズム
Fig. 1 ID3 algorithm.

$$info(S) = -\sum_{j=1}^{m} \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right). \quad (6)$$

(b) 選択された属性を節点とし，その属性の有するすべての属性値ごとに枝を生成する．

(c) 枝に対応する属性値によって分割された事例データを用いて (b) で生成された枝の先で，再帰的に (1)，(2) の手続きを行い，サブツリーを生成する．

2.3 属性選択基準

ID3 が用いている属性選択基準，式 (4)～(6) はゲイン法と呼ばれ[1),2)]，情報理論であるエントロピー[12)]に基づいて属性の持つクラスに関する情報量を計算する．この値に基づく属性選択は，情報量の高い属性の順にルートから並ぶため，この手続きによりサブツリーの冗長性が抑えられ，得られる決定木はコンパクトとなる．節点と端点の総点として定義されるサイズと，決定木の分類の正確さを表す分類精度の相関関係は密接であり，コンパクトな木は分類精度が高いことは理論的に証明されている[13)]．よってゲイン法による属性選択は，分類精度向上のために機能することが期待できる．ゲイン法のほかにも多くの選択基準が提案されている．代表的なものとして，Quinlan によるゲイン比法[1),2)]，Hart らによる χ^2 を用いる手法[14)]，Minger による G 統計[14)]を用いる手法などがある．しかしながら，Rendell らは，ID3 の分類精度はデータの構造に大きく依存することを明らかにしている[15)]．よって ID3 に対する拡張の多くは，汎用性の高い拡張よりも，特定のデータの構造，性質に依存した拡張である．

3. 非独立性属性の扱い

ここでは，単純 ID3 では分類困難なデータ特性である，非独立性属性を含むデータについて述べ，問題解決のアプローチを示す．

3.1 セマンティクス的に非独立な属性群を含むデータ

これまで有効な拡張が行われていないデータ特性として，事例データ中の属性がセマンティクス的に従属関係があるものがある．たとえば，これらの属性を含む事例は，

$$E_{dependance} = \{A_1, A_2, \ldots, A_{i-2}, (A^*_{i-1}, A^*_i, A^*_{i+1}), A_{i+2}, \ldots, A_n\}, \quad (7)$$

と表現される．$(A^*_{i-1}, A^*_i, A^*_{i+1})$ である属性群は，群単位で 1 つの意味を持つ．これに対し，式 (4)～(6) は，各属性がセマンティクス的に独立であるとき，正確な属性情報を出力する．現実的には，この非独立性属性群が対象データに含まれていることは十分に予想でき，また，データによっては互いの属性がクラス決定とは無関係に関連があると考えるのは自然である．よって，この特性を含むデータに対する ID3 の拡張を行うことは ID3 のデータへのバイアスを減少させる一手段であると考える．

3.2 非独立属性群に対しての問題解決のアプローチ

高い分類精度を示す決定木生成に際して中心的な役割を果たすのが，属性選択であり，データの性質に応じた柔軟な選択が望まれる．すでに述べたとおり機械学習の観点から様々な属性選択基準が提案され議論されている．この属性の選択は，一般的な見方をするならば，特徴選択問題 (Feature Selection Problem) と考えることができる．画像認識の分野などで広く認識されている特徴選択問題とは，入力された対象データに内包される様々な特徴のなかで，正確かつ能率的な対象データの認識を目的に，どの特徴に注目するかを取捨選択する問題である．

ここで，特徴選択問題における従来手法のアプローチを考察する．Vafaie & De Jong[10)]，John, Kohavi & Pfloger[11)]は，存在する属性を “適切”，“不適切”，さらに “弱い適切” と明示的に規定し，その後 (弱い) 適切と判断された属性のみが，帰納アルゴリズムに入力され，クラス決定が行われる．つまり，Vafaie & De Jong[10)]では，存在する属性の削除/追加を繰り返すことにより，クラス決定に有効な属性群の探索を行う．また，John, Kohavi & Pfloger[11)]は，有効な属性群の探索を GA を用いて行っている．

本論で問題としている，非独立性属性を含むデータは，Vafaie & De Jong[10)]での “適切”，“弱い適切” な属性によって構成されており，“適切”，“弱い適切” の関係を明確に定義することが，正確なクラス概念を表現するものと考える．したがって対象データの属性群のなかから有効な副集合を探索する従来の方策は有効ではないと考えられる．

上記の議論から，本論では以下の 2 つのアプローチを考察する．それらは，複合的に属性を評価する機構，あるいは従来とは異なる基準を追加することで属性を評価する機構の導入である．前者のアプローチとして，

トレーニングデータに存在する属性群によって対象概念の記述が困難である場合，システムが新たな属性を生成する手法が提案されている[16),17)]．このような手法は構成的帰納学習と呼ばれており，システムを通じて新たな概念の発見や部分概念の発見が期待できる．非独立性問題においても複合属性を1つの属性とみなすような属性の出現が期待できる．後者のアプローチにおいて，2節で述べたID3のゲイン法以外の選択基準もゲイン法と同様の理由で非独立属性に対して有効に機能しない．したがって，現在の課題に対して有効な選択基準とは，属性のクラスに関する統計的な情報量を用いるのではなく，属性が持つ概念的な情報を評価する基準であると考えられる．そこで本論では，この概念的な属性情報をEG2[6)]で用いられている背景知識（Background Knowledge）として利用し，特に属性が持つ相対的な重要度を背景知識として従来のID3への導入を試みる．

属性が持つ概念的な情報を評価することは，以下の点で有効性が期待される．まず，“適切”，“弱い適切”な属性が持つ概念的な情報を数値的に表現し，それらの間の関連性を規定することが可能であるならば，“適切”な属性と“弱い適切”な属性を明確に区別することが可能になる．さらに，概念的な情報を数値的に表現することによって，統計的にはクラスに無関係である非独立性属性に対して他の属性とのクラスに関する関連性，あるいは重要度を比較することが可能になる点である．

非独立性属性群 $(A^*_{i-1}, A^*_i, A^*_{i+1})$ において，各属性は独立してはクラスに関する情報を持たない．このとき，属性が持つ相対的な重要度 $\{\delta_1, \delta_2, \ldots, \delta_n\}$ をとし，ゲイン法による情報量 Λ_k は，属性がセマンティクス的に独立であるとき，多くの局面において，

$$\Lambda_i \cong \delta_i, \tag{8}$$

となる．しかし，非独立性属性の場合，

$$\Lambda^*_i \neq \delta_i, \tag{9}$$

であることが多い．したがって，ゲイン法による各非独立属性の情報量 $\{\Lambda^*_{i-1}, \Lambda^*_i, \Lambda^*_{i+1}\}$ は正確に属性の性質を反映しない．たとえば，属性 $\{A_k, A^*_l\}$ に対して，重要度が，

$$\delta_k < \delta_l, \tag{10}$$

であり，またゲイン法による情報量が，

$$\Lambda_k > \Lambda^*_l, \tag{11}$$

であるとき，選択される属性は，

$$\begin{aligned} A_s &= \max_{i=1,2,\ldots,n} (\Lambda_i, \Lambda^*_i) \\ &= A_k, \end{aligned} \tag{12}$$

は，節点の属性として適切ではない．ところで，本論で提案する概念的な重要度を表す重要度パラメータ I_i は，

$$I_j \cong \delta_j, \tag{13}$$

となるように設定される．そして，後述する式により，重要度パラメータによって各属性の情報量は，

$$M\Lambda^*_i = \Lambda^*_i + I_i \cong \Lambda^*_i + \delta_i, \tag{14}$$

と修正され，属性 A_l が選択されることになる．一方，前出の，“不適切”な属性の重要度が低く設定されるならば，式(14)から結果的にその属性は節点として選択されず，従来の特徴選択問題での方策でみられた属性の“削除”が再現されることになる．ただし，本論で問題としている非独立性属性群に関しては，“削除”よりも，概念的な重要度の規定が有効であると考えられるため，この重要度パラメータにより，従来の属性値とクラスの統計的関係を求めた情報量に加えて，概念的な重要度を評価に加えることにより，統計的にクラス決定に無関係な性質を持つ非独立性属性に対して有効に機能することが期待される．

4. 提案手法

本章では，非独立性属性を含むデータに対する分類精度向上のための拡張であるID3-GAについて述べる．前述の議論により，式(13)，(14)により，そのようなデータにおける決定木最適化のアプローチとして，属性の持つ重要度を生成プロセスに導入することを提案した．ここでは，各属性の重要度をパラメータという形で取り扱うことを提案する．具体的な手法は以下のとおりである．

4.1 ID3-GAの構成

各属性に重要度を表すパラメータを設定する．すなわち各属性を $\{A_1, A_2, \cdots, A_n\}$ とするとき，それぞれに対して重要度パラメータを設定し，$\{I_1, I_2, \ldots, I_n\}$ とする．また，このパラメータは属性選択の際に作用し，ゲイン法による *gain* は以下のように修正される．

$$\begin{aligned} \textit{modified gain}(A_i) = \textit{gain}(A_i) \\ + \frac{I_i}{\sum_{j=1}^{n} I_j} \times \log_2 k. \end{aligned} \tag{15}$$

ただし，k は対象データに存在するクラスの数である．重要度パラメータは正規化後，情報量 *gain* に加えられる．したがって属性選択は，従来のクラスに関する情報である *gain* に加えて，属性が有する相対的な重要度も評価に加えられることになる．

ここで問題となるのは，適切な重要度パラメータの設定である．対象データの属性情報が与えられること

は稀であるため，適切な重要度パラメータをトレーニングデータから獲得しなくてはならない．本論では，最適重要度パラメータの組み合わせをGAを用いて獲得する．具体的には，GAにおける探索ストリングと重要度パラメータとを対応させる．そして，ストリングに対応する重要度パラメータによって作用を受けた決定木の分類精度とサイズを用いてストリングの評価を行い，世代更新を行うことによって最適重要度パラメータの獲得を行う．以下では，GAによる重要度パラメータ獲得機構を述べる．

4.1.1 GAによる重要度パラメータ獲得機構

まず，GAによる重要度パラメータ獲得のための第一段階としてトレーニングデータの分割を行う．GAは，反復探索を行い最適値を得る．そのためには，ストリングを用いて決定木を生成するためのデータとその決定木を評価するデータが必要となる．分割比は，データに応じて決定するものとする．

次に，重要度パラメータの要素値を定義する．簡単のため，要素値はあらかじめ定められた範囲内の離散値とする．したがって重要度パラメータ I_n の要素値は $\{m_1, m_2, \ldots, m_l\}$ から選択されることになる．

GAストリングは，重要度パラメータをグレイコーディングしたものを使用する．要素値の数に応じて各属性のビット数を確保する．すなわち，属性数 $\{A_1, A_2, \cdots, A_n\}$，要素値を $\{m_1, m_2, \ldots, m_l\}$ としたとき，ストリングは，以下のようにして決定される．

$$[(P_{11}, P_{12}, \ldots, P_{1k})(P_{21}, P_{22}, \ldots, P_{2k}) \cdots \cdots (P_{n1}, P_{n2}, \ldots, P_{nk})]. \quad (16)$$

ただし，k は，$l - 2^k < 0$ となる最小の自然数，P_{ij} は，1または0が入る．

GAの評価関数は，以下の式で定義される．

$$fitness = (100.0 - percent\ correct) \times \alpha_c + size\ of\ tree \times \alpha_s, \quad (17)$$

percent correct とはストリングをデコードして得られた重要度パラメータで，式(4)を用いて生成された決定木の分類精度とする．また，*size of tree* は，節点と端点の合計とする．決定木のサイズを評価値のひとつとする理由は，前述したとおり分類精度と決定木には密接な関係が存在することが理論的に示されているためである[13]．よって，第1項，第2項ともに小さいほうがよい評価値が得られることになる．α_c, α_s は，それぞれを連結する修正定数である．

以上の記述をもとに非独立性属性を含むデータ向きのID3-GAのアルゴリズムを図**2**に示す．

前出の提案アルゴリズムを用いて構築した問題向きシステムを図**3**に示す．

```
PROCEDURE ID3-GA
BEGIN
initialize population P(t)
FOR(t;trial)
    devide training data into making-data and evaluating-data
    convert GAstring to importance parameter;I
    FOR(p;population size)
        make tree(current node) / *using making-data * /
        IF freq(C_j,S) = 0
            then label the leaf with the class, and RETURN
        ELSE
        FOR(i;each attribute)
            info_X(T) = Σ_{k=1}^{N} (|T_k|/|T|) × info(T_k)
            info(S) = Σ_{j=1}^{m} (freq(C_j,S)/|S|) × log_2 (freq(C_j,S)/|S|)
            gain(X) = info(T) − info_X(T)
            modified gain(X) = gain(X) + I_X / ΣI_X
            Selected attribute A = max_{k=1,2,...,n} (gain(k))
        FOR (i; each attribute value of attribute A)
            make branch of each i
            make tree(next node)
    FOR(p;population size)
        evaluate tree / *using evaluating-data * /
    reproduct P(t) from P(t−1)
    mutate P(t)
    crossover P(t)
END
```

図2 ID3-GAのアルゴリズム
Fig. 2 ID3-GA algorithm.

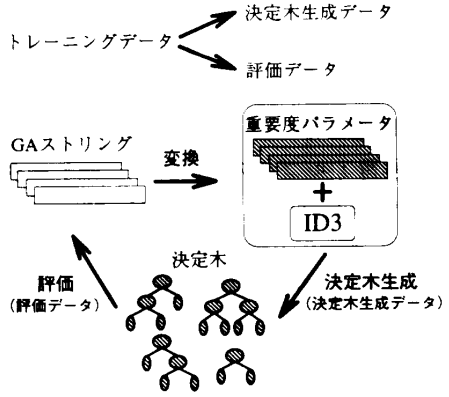

図3 ID3-GAの流れ
Fig. 3 A diagram of ID3-GA model.

5. 計算機実験

本論では，MONKS問題[18]，実データに基づく乳癌診断問題，6-Multiplexor問題[19]を用いて本提案の挙動の確認，考察を行う．MONKS問題は帰納学習システムの性能評価データとして知られており，これ

表1 パラメータ設定
Table 1 Used parameters in this experiment.

Population size	50
Trial	100
Crossover Rate	0.5
Mutation Rate	0.01
α_s	10.0
α_c	10.0
Patition rate of training data	50%

により提案手法の一般的な挙動を確かめる．乳癌診断問題は，実データに対する汎用性，適合性を確認する．また，6-Multiplexor 問題は，非独立性属性を含むデータであり，このデータを用いて非独立性属性を含むデータに対する性能評価を行う．

手法の評価の基準は，分類精度と木のサイズとする．実験は，データを2分割し，一方を決定木を生成するためのデータ（トレーニングデータ），一方を決定木を評価するデータとして用いる．よって，分類精度とサイズは後者のデータによって求められる．ただし，提案手法では，その生成のためのデータを各ストリングによる決定木生成のためのデータと，ストリング評価のためのデータとしてさらに2分割を行う．なお，各実験値は5回の試行の平均値である．また，各実験における最大世代数（Trial）は，予備実験において100世代で十分な収束がみられたため，また，α_c, α_s 修正定数は，同様に予備実験から，等比の値に設定することで有効な収束が観察されたため，それぞれを**表1**の値に設定した．

以下に，各実験の結果と考察を示す．

5.1 MONKS 問題

MONKS 問題とは，Thrun によるテストデータ群であり，帰納学習アルゴリズムの性能評価データとしてよく使用されている[2),18)]．MONKS 問題はそれぞれ6個の属性 $\{A_1, A_2, \ldots, A_6\}$ からなる3つの問題から構成され，それぞれ MONKS1，MONKS2，MONKS3 と呼ばれる．以下にそれぞれの MONKS 問題のクラス概念を示す．

•MONKS1
- class 1 if $(A_1 = A_2)$ or $(A_5 = 0)$
- class 0 otherwise.

•MONKS2
- class 1 if exactly two of the six attributes have their first value,
- class 0 otherwise.

表2 実験結果
Table 2 Result of experiment.

		ID3	Extended ID3
MONKS1	Accuracy	81.71%	100.0%
	Size of Tree	95.0	39.2
MONKS2	Accuracy	68.75%	69.44%
	Size of Tree	170.0	165.2
MONKS3	Accuracy	93.98%	93.97%
	Size of Tree	43.0	47.8
Breast Cancer	Accuracy	92.66%	92.93%
	Size of Tree	88.0	114.1
6-Multiplexor	Accuracy	62.63%	94.37%
	Size of Tree	27.1	18.4

•MONKS3
- class 1 if $(A_5 = 3$ and $A_4 = 1)$ or $(A_5 \neq 0$ and $A_2 \neq 3)$,
- class 0 otherwise.

ただし，MONKS3 には属性値に対して5%のノイズ（記述の誤り）が作為的に含まれている．

この実験での重要度パラメータのメンバーは $\{0, 1, \ldots, 6, 7\}$，とする．このとき，式(16)に従い各属性は，3ビット，ストリング長は18ビットとなる．集団に対して施される遺伝的オペレータとして，交叉と突然変異を用いる．交叉は単純一点交叉とし，突然変異は，ストリングの1ビットに対して行われるものとする．予備実験によって定められたその他の GA のパラメータは，表1に示す．

各 MONKS 問題における単純 ID3 と提案手法である ID3-GA の分類精度，サイズ，および獲得された重要度パラメータは**表2**，**表3**に示す．

各問題において ID3 と ID3-GA の分類精度，サイズの差にばらつきが見られる．このことは，各問題のクラス概念が深く関係している．MONKS1 では，クラスは，特定の属性で決定されることから，属性間に相対的な重要度の差が生じている．このとき，

$$\Lambda_i \propto \delta_i, \tag{18}$$

であり，また獲得された重要度パラメータは，

$$I_i \geq \Lambda_i, \tag{19}$$

であることから，双方を利用する ID3-GA は，高い分類精度を得た．また，トレーニングデータは，すべての属性値の組み合わせを含む完全な事例集合の副集合である．そこには必ず事例分布の偏差が存在する．したがって，事例を統計的な評価しか行わない ID3 は，この偏差によって属性選択がうまく行われないことがある．一方の提案手法ではデータの概念的な重要度を獲得するため事例分布の偏差によらないという特長を持

表3 獲得されたパラメータ
Table 3 Acquired parameters.

	A_1	A_2	A_3	A_4	A_5	A_6			
MONKS1	5.0	5.8	1.4	0.2	3.0	0.8			
MONKS2	0.8	3.6	3.4	0.6	3.2	1.8			
MONKS3	3.0	5.0	0.4	3.6	6.6	2.6			
6-Multiplexor	5.2	6.0	0.4	2.0	1.0	0.4			
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
Breast Cancer	1.4	5.0	3.6	3.4	4.0	6.6	4.0	2.6	2.7

つ．MONKS1とは対照的にMONKS2では，クラス概念は特定の属性を対象としておらず，よって属性間の相対的な重要度は存在しない．したがって，トレーニングデータの事例分布がそのまま決定木に反映されることになり，正しい概念は獲得できない．とりわけ，ID3-GAにおいては，扱うべき相対的な重要度が存在しないため十分には機能しない．したがって，図4からも読み取れるように探索が進行しても評価値の向上は見られなかった．最後にMONKS3においても両者に明確な差は見られなかった．MONKS3は，クラス決定条件はNONKS1と同様に容易であるにもかかわらず，両者の分類精度，サイズはほぼ似た値となった．その原因はノイズに起因する．ノイズを含む事例は，決定木生成データ，評価データの双方に含まれる．ノイズの影響を受けた決定木がノイズを含むデータでストリングの評価が行われるが，GAの反復探索を通じて，表3で読み取れるようにMONKS3のクラス概念を反映した重要度パラメータは獲得されている．しかしながら，獲得された重要度パラメータを用いて生成される決定木は，ノイズを含むデータで生成されるため，最終的な決定木は不完全なものとなり，ID3との差が生じなかった．ただし，その重要度パラメータを用いてノイズを含まないデータによって決定木を生成した場合は，分類精度が高くなることを確認している．

5.2 実データに基づく乳癌診断問題

実データに基づく乳癌診断問題は，Wolbergらによって採取されたデータを元に，乳癌の良性/悪性の事例を判別するデータである．9つの属性で特徴づけられる699の事例で構成される．それぞれの属性は，腫瘍の大きさ，癒着の具合などであり，属性値はすべて離散値である．

重要度パラメータの要素値は，MONKS問題と同様とする．よって，9つの属性からなるこのデータは，式(16)により各属性は3ビット，ストリング長は27ビットとなる．その他のパラメータは，表1に示す．

実験から得られた分類精度，サイズおよび重要度パラメータは表2，表3に示す．乳癌診断問題のクラス概念は複雑であり，クラス決定には，多くの属性が関係している．このようなデータにおいて，クラス決定情報を統計的に獲得するID3は，高い分類精度を示す．一方，ID3-GAでは，属性間の相対的な重要度は明確に存在しないため，分類精度向上のための重要度パラメータは獲得できなかった．このことは，図4からも読み取れる．しかしながら，得られた重要度パラメータは，トレーニングデータの事例分布偏差を解消するために働き，分類精度は，ID3と比較してやや高い結果となった．

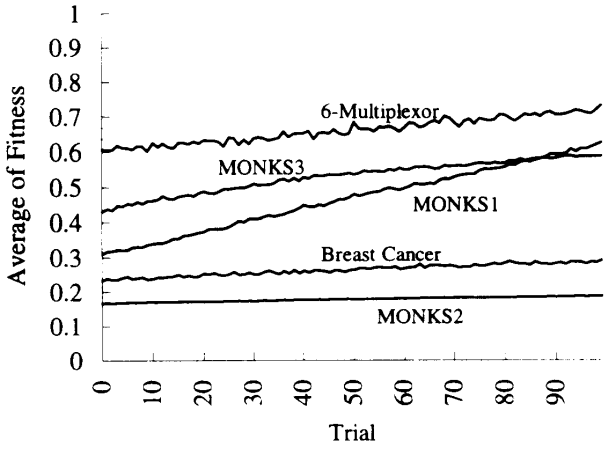

図4 評価値の推移
Fig. 4 Changing of fitness.

5.3 Multiplexor問題

Multiplexor問題とは，非独立性属性を含むデータの典型的な例として知られており，単純ID3では分類困難な問題とされている[2]．したがって本論では非独立な属性を持つデータの一例であるMultiplexor問題によって提案手法の考察を行う．Multiplexor問題は，長さ $L=k+2^k$（k は自然数）からなるビット列の分類問題である．すべての値は2値からなり，ビット列はアドレスビット a_i，データビット d_i に分けられる．そのビット列のクラスは，アドレスビットをバイナリーコードとしての値が示すデータビットの値によって一意に決定する．一般に L-Multiplexorは次のように与えられる．

$$a_0 a_1 \ldots a_k d_0 d_1 \ldots d_{2^k-1} c. \qquad (20)$$

特に6-Multiplexor（$L=6$，$k=2$）のクラスは次式

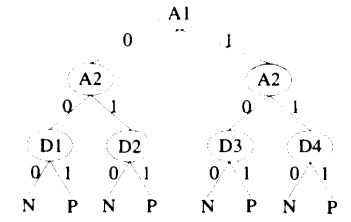

図5 6-Multiplexor の最適決定木
Fig. 5 Optimal tree of 6-Multiplexor.

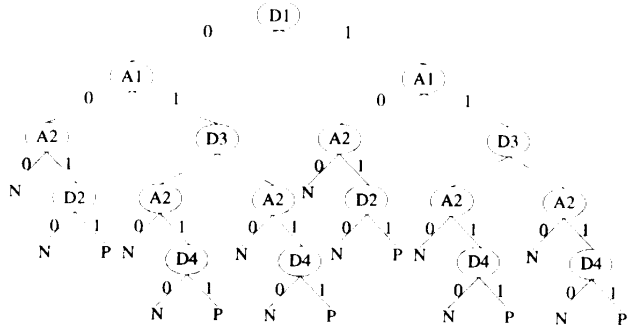

図6 ID3 による 6-Multiplexor の決定木
Fig. 6 ID3's tree of 6-Multiplexor.

で決定される.

$$c = \bar{a}_0\bar{a}_1d_0 + \bar{a}_0a_1d_1 + a_0\bar{a}_1d_2 + a_0a_1d_3. \quad (21)$$

6-Multiplexor 問題において，アドレスビットの二属性は，2つの属性で1つの情報を表すというセマンティクス的に従属な関係である．よって，6-Multiplexor 問題は，ID3-GA が最も有効に機能するデータであると考えられる．

6つの属性で構成されているため，ストリング長は18ビットとなる．その他のパラメータは，表1に示す．6-Multiplexor 問題における ID3，および ID3-GA の分類精度，サイズ，獲得されたパラメータは表2，表3となった（**図5，6**）．

以上の結果を元に，6-Multiplexor 問題における ID3，ID3-GA の決定木生成過程の検証を行う．アドレスビットおよびデータビットの分割表（Contingency Table）を，**表4，表5** に示す．

表4，表5の分割表から，アドレスビットにおいては，いずれのクラスに対しても属性値は平均的に分布していることが読み取れる．一方，データビットでは，属性値によって支持するクラスが異なる偏った分布であることがわかる．以上の事実をふまえ，式(4)～(6)を用いて各ビットの情報量を求める．前出の表記法からアドレスビットの情報量 Λ_i^* を，データビットは Λ_i とする．

$$\Lambda_i^* = 0.000, \quad (22)$$

$$\Lambda_i = 0.064. \quad (23)$$

これは，ルートの端点にクラス決定に重要な役割を果

表4 アドレスビットの分割表
Table 4 A contingency table of address bit.

		Class 1	Class 0	
Address bit	1	16	16	32
	0	16	16	32
		32	32	64

表5 アドレスビットの分割表
Table 5 A contingency table of data bit.

		Class 1	Class 0	
Data bit	1	20	12	32
	0	12	20	32
		32	32	64

たしているアドレスビットではなく，データビットが選択されることを示している．次に実験によって獲得されたパラメータによって情報量の修正を行う．

$$M\Lambda_i^* = \Lambda_i^* + I_i = 0.000 + 0.433 = 0.433, \quad (24)$$

$$M\Lambda_i = \Lambda_i + I_i = 0.046 + 0.034 = 0.080. \quad (25)$$

式(24)，(25)は，適切な重要度パラメータによって修正されることにより逆にアドレスビットが端点に選択されることが読み取れる．したがって上記の2式から獲得されたパラメータは，ID3 による情報獲得の是正を行い，非独立性属性に対して有効であることが示された．

5.4 考　　察

計算機実験で得た知見をまとめると以下のようになる．

(1) 属性間に相対的な重要度の偏りがあるデータに対して，その重要度をパラメータとして獲得を行い ID3 決定木の生成プロセスに導入することは有効である．つまり，この特性を持つデータにおいては，重要度パラメータはエントロピーによって獲得する情報を際立たせ，高い分類精度を示す決定木が得られた．

(2) 逆に重要度に差のない属性群から構成されるデータでは従来法に対して有効性は確認されなかった．しかし言い換えれば，冗長な計算は行われるものの，このようなデータにおいて提案手法は，最低限 ID3 の分類精度を得ることが確認された．

(3) ノイズを含むデータに対しても，提案手法は，

(存在するならば) 相対的な重要度の獲得を行うことが可能である. しかしながら, 最終的な決定木はノイズを含むデータで生成されるため ID3 と分類精度の差は生じなかった.

(4) 実データに見られる複雑なクラス概念を持つデータに対しては分類精度の改善は見られなかった. しかしながら, ここでも提案手法の分類精度は, ID3 の分類精度に変わらない値を示した.

(5) 非独立性属性を含むデータである Multiplexor 問題に対しては特に有効に機能した. すなわち, 属性が持つ概念的な重要度を考慮することで, 複合的な評価が行われ, 結果的に高い分類精度が得られた. この実験の結果から非独立性属性を含む他のデータに対しても有効に機能することが期待される.

6. おわりに

本論では, 従来の ID3 では分類が困難であった問題, 特に属性の非独立性問題をかかえる対象データに対して高い分類精度を得るための ID3 の拡張のアプローチとして ID3-GA を提案し, 計算機実験を通じて有効性, 特性を確認した.

また, 他の特性を含む様々なデータに対する本提案手法の汎用性の評価のための他の手法との比較は, 今後の課題とする.

参考文献

1) Quinlan, J.R.: *Induction of Decision Trees*, Machine Learning 1, pp.81–106 (1986).
2) Quinlan, J.R.: *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers 2929 Campus Drive, Suite 260 San Mateo, CA94403 (1993).
3) Mingers, J.: An Empirical Comparison of Pruning Methods for Decision Tree Induction, *Machine Learning 4*, pp.227–243 (1989).
4) Utogoff, P.E.: Incremental Induction of Decision Trees, *Machine Learning*, Vol.4, pp.161–186 (1989).
5) Tan, M.: Cost-Sensitive Learning of Classification Knowledge and Its Application in Robotics, *Machine Learning*, Vol.13, pp.7–33 (1993).
6) Núñez, M.: The Use of Background Knowledge in Decision Tree Induction, *Machine Learning*, Vol.6, pp.231–250 (1991).
7) Goldberg, D.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley (1989).
8) Devijver, P.A. and Kittler, J.: *Pattern Recognition: A Statistical Approach*, Prentice-Hall International (1982).
9) Ichino, M. and Sklansky, J.: Optimum Feature Selection by Zero-one Integer Programming, *IEEE Trans. Systems, Man, and Cybernetics*, Vol.14, No.5 (1984).
10) Vafaie, H. and De Jong, K.: Genetic Algorithms as a Tool for Feature Selection in Machine Learning, *Fourth International Conference on Tools with Artificial Intelligence*, IEEE Computer Society Press, pp.200–203 (1992).
11) John, G.H., Kohavi, R. and Pfloger, K.: Irrelevant Features and the Subset Selection Problem, *ICML-94*, pp.121–129 (1994).
12) 梅垣壽春：情報エントロピーと作用素解析, エントロピー—その様々な顔つき—, 別冊数理科学, サイエンス社, pp.34–41 (1992).
13) Mohammad, F.U.: *On the Induction of Decision Trees for Multiple Concept Learning*, UMI Dissertation Service (1991).
14) Mingers, J.: An Empirical Comparison of Selection Measures for Decision-Tree Induction, *Machine Learning*, Vol.3, pp.319–342 (1989).
15) Rendell, H. and Cho, H.: Empirical Learning as a Function of Concept Charactor, *Machine Learning*, Vol.5, pp.267–298 (1990).
16) 滝 寛和：構成的帰納学習とバイアス, 人工知能学会誌, Vol.9, No.6, pp.28–32 (1994).
17) Elio, R. and Watanabe, L.: An Incremental Deductive Strategy for Controlling Constructive Induction in Learning from Examples, *Machine Learning*, Vol.7, pp.7–44 (1991).
18) Thrun, S.B., et al.: The Monk's Problems: A Performance Comparison of Different Learning Algorithms. Technical Report CMU-CS-91-197, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA (1991).
19) Wilson, S.W.: Classifier Systems and the Animat Problem, *Machine Learning*, Vol.2, pp.199–228 (1987).

(平成 7 年 1 月 30 日受付)
(平成 7 年 12 月 8 日採録)

本堂　直浩（正会員）

昭和46年生．平成6年北海道大学工学部精密工学科卒業．現在，同大学大学院システム情報工学修士課程に在学中．推論，および自動プログラミングに興味を持ち，帰納学習，遺伝的アルゴリズム，遺伝的プログラミングの研究に取り組んでいる．

成瀬継太郎（正会員）

昭和43年生．平成2年北海道大学精密工学科卒業．平成4年同大学院精密工学専攻修士課程修了．平成7年同大学院博士課程修了．工学博士．ニュージャージー工科大学機械工学科研究官．現在に至る．ロボティクス，機械学習に興味を持つ．日本機械学会の会員．

嘉数　侑昇（正会員）

昭和16年生．昭和48年北海道大学大学院工学研究科精密工学専攻博士課程修了．工学博士．北海道大学工学部複雑系工学講座教授．創発的学習，自律系工学，ロボティクス，知識工学，ニューラルネットワーク等の研究に従事．日本機械学会，精密工学会，日本ロボット学会，計測自動制御学会等の会員．