ショートノート

Research Notes

決定木による階層属性を用いた概念の帰納学習

Learning Concepts in a Decision Tree Form Using Hierarchical Attributes

中島　誠*1　葉　玲如*2
Makoto Nakashima　Lin-Ju Yeh
伊藤　哲郎*1
Tetsuro Ito

*1 大分大学工学部知能情報システム工学科
Dept. of Computer Science and Intelligent Systems, Oita University, Oita 870-11, Japan.
*2 筑波大学 電子・情報工学系
Institute of Information Sciences and Electronics, University of Tsukuba, Tsukuba-shi 305, Japan.

1994年1月21日受理

Keywords: learning from examples, decision trees, generalization, hierarchical attributes, structured examples.

Summary

The ID3 algorithms provide robust inductive processes of learning concepts from examples by constructing decision trees. The standard ID3 algorithm, however, is restricted to utilize symbolic/numeric attributes, and receive non-structured examples. We here extend the algorithm so that it can treat hierarchical attributes, and can receive structured examples (A hierarchical attribute relates its values hierarchically, and a structured example is an example having more than one component). The first problem is solved by finding adaptively appropriate values for getting a target decision tree based on the formulated value generalization process. The second is by introducing a new type of attribute-based descriptions in which any attribute refers to some specified components. Computational experiments are also examined to show the validity of the proposed methods.

1. はじめに

機械による概念の学習は，一定の基準で区分けされた観測事例を属性・属性値を介した簡潔な記述で特徴づけていくこととみなせる．そして区分けが外部よりなされるか否かで，それぞれ例からの学習あるいは観察による学習と呼ばれる．従来から多くの概念学習モデルが提案されているが，なかでも前者に属するモデルの一つで，概念の記述を決定木の形で求めるID3 [Quinlan 83]は，観測データの人手による分類を支援するものとして，多くの場面で利用されてきている．

既存のID3では，事例の特徴づけに，値が相互関係を持たない記号属性あるいは値が数直線上に並べられる数値属性が用いられていた．ところが，記号属性に関しては，その値の間に概念的な関連があると考えるのが自然であるし，数値属性についても，並んだ値は上位の値としての区間を段階的に形づくっているとみられる．すなわち，記号属性ならびに数値属性は，値の関連が階層構造をした属性(以後，階層属性と呼ぶ)の一種と捉えられる．また，多次元ベクトルや図形データを学習に利用する場合にも階層属性とみなしたほうが柔軟な扱いが可能になる．この考察から本論文では，種々の属性を階層属性として統一的に扱いながら決定木を求めることができる学習モデルを，属性値の一般化方策に基づきデザインする．

ところで，ID3が入力として受け取る事例は，従来から部分を持たない非構造化事例とされていた．しかし，現実的な場面では，複数の部分からなる構造化事例を扱う必要性が出てくる[Michalski 80]．このとき学習を通常のように準備された属性だけに注目して行うと，決定木を成長させるための事例集合がうまく分割されなかったり，また得られた記述が複雑なものになったりする．これには，定められた部分（複数のときもある)が持つ属性を利用する方策[伊藤 94]を取り入れ，構造化事例をこれまでと同様の様式で扱えるようにする．

以下，2章ではID3を概観し，3章，4章ではそれぞれ階層属性ならびに構造化事例の扱い方を定式化する．5章ではここでの方法の有効性を計算機実験を通じて検討する．

2. 決　定　木

ID3は，いくつかの属性とそれらの値で特徴づけられた事例集合が与えられたとき，適当な属性を選択し，その値をもとに集合を分割しながら，最終的に学習すべき概念の事例（正事例という）とそうでない事例（負

表1 ID3手続き

:最初与えられた事例集合をC_0とする.

(S1) C_0に対応した根節点を作り,C_0をCと考えS2を行う.

(S2) 集合Cに対し次を行う.

(S2.1) もしC中の例がすべて正事例ならば対応する節点を正節点とし,すべて負事例ならば対応する節点を負節点とする.

(S2.2) S2.1以外ならば決定木を成長させるために最適な属性を一つ選ぶ.

(S2.3) 選んだ属性の値$v_1, v_2, \ldots, v_m (m > 1)$に従い,$C$を$C_1, C_2, \ldots, C_m$に分割する.これら部分集合に対応した節点を作りながら,各$C_i (1 \leq i \leq m)$を新しくCと考え,S2を繰り返す.

事例という)を区別できる決定木を構成していく.基本的な手続きは表1のようにまとめられる.

決定木が作られると,葉に位置する各正節点が学習すべき概念の一面を表しており,これは木の根からそこに至る道上で分割に使われた属性・属性値に関する記述の連言で規定される.正節点すべてに対する記述の選言が学習すべき概念を定めている.ステップS2.2での決定木を成長させる属性としては,節点総数のできるだけ少ない,言い換えればできるだけ簡潔な結果が得られるようなものが選ばれる.具体的には,ある属性の値$\{v_i\}$に従ってCを$\{C_i\}$に分割したとき,分割前のCが持つ正・負事例の分け方に関する情報量の期待値としてのエントロピー*1から,分割後の各C_iの持つエントロピーの平均*2を引いた情報獲得量を求める.そして各属性に関する獲得量を指標として,これを最大にする属性が選ばれる.

ID3手続きの現実的な利用に際しては,いくつかの配慮がなされる.まず,獲得量をそのまま使うと多くの種類の値を持つ属性が過大評価されてしまう.そのため,値の数を反映した係数*3で割って正規化しておく[Quinlan 86, 辻野 89].次に,事例ごとの属性値が正しいとは限らない(雑音が含まれている)場合,ステップS2をいくら繰り返しても適切な決定木を作ることはできない.これには,例えばS2.1で,正(あるいは負)事例に対する負(正)事例の比,すなわち誤り率,が定められた限度内になると,正(負)節点とみなす方法がとられる[Quinlan 87].また,属性値が数値の場合にはS2.3で非常に多くの部分集合が出てきてしまう.これには,数値をいくつかに区分けし,各区間を一つの値とみなす方法がとられる[荒木 92, Quinlan 86, 辻野 89].

*1 C中での正(負)事例の出現確率を$p(n)$としたとき,$-p\log_2 p - n\log_2 n$.

*2 C_i中での正(負)事例の出現確率を$p_i(n_i)$,値v_iの生起確率をO_iとしたとき,$\sum_i(-p_i\log_2 p_i - n_i\log_2 n_i)\cdot O_i$.

*3 例えば,各O_iをもとにした$\sum_i -O_i\log_2 O_i$.

3. 階層属性の扱い

3・1 階層属性

いま,表2でa, bを正事例,cを負事例とした場合を考えてみよう.属性「色」を用いた事例集合の分割により,「色が赤あるいは色が紫」が学習すべき概念の記述として求められる.「色が青」は負事例からの記述となる.ところで,図1の赤,紫や青を葉とした色の構造をうまく利用すると,求めるべき概念の記述はより簡潔な「色が暖色」とできる.正事例b, cと負事例aや,正事例a, cと負事例bからの学習を考えたときでも,それぞれ寒色を使ったり上位の値を使わないで済ませたりできる.すなわち,学習モデルが階層属性をうまく扱えるようデザインされていると,この例のように簡潔な決定木を作り出すのに大きく役立つ.

ここでの階層構造とは,根つき木あるいは図1のような根つき木に余分の辺が追加されたものを指す.構造中では各節点は属性値に対応しており,その子孫(先祖)に対しより概念的に広い(狭い)関係を有する.事例を特徴づける値は葉に現れる.構造が木の形をしているなら任意の値の集合に対する極小で共通の先祖lcga (least common generic ancestor)は一意に決まる.互いの間の関係を考慮しない記号属性の構造は,根に直接葉がついた形で,また数値属性は,隣りどうしの値が段階的により上位の属性値を作り出した形で表現できる(図2).数値属性からの階層構造は複雑に

表2 事例

事例＼属性	色	形
a	赤	丸
b	紫	丸
c	青	丸

図1 色の値の構造

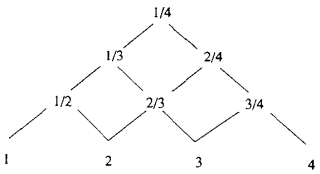

図2 数値属性の階層属性表現
*例えば1/3は区間[1, 3]を意味する.

見えるが，木の場合と同様，任意の値の集合に対するlcgaは一意に決まる．

3・2 属性値の選択

階層属性の扱いを可能とするには，決定木を成長させるとき，①複数ある属性のうちどれを用いるのか，また②選ばれた属性のどの値を使うのか，を明らかにすればよい．単純には，各属性から最適な分割を与えると思われる属性値を取り出し，これらのうちで最大の情報獲得量を与える属性を選ぶようにすればよい．問題は，属性の種類や属性ごとの値の数が多くなった場合，最適な分割を与える属性値を取り出すのに非常な手間がかかってしまうことである．ここでは次のように対処する．

①に対して：構造中の葉での値は，それを利用したとき，事例集合を分割できる潜在的な能力の限界を示している．それゆえ，葉を使ったときの情報獲得量を属性選択の一つの指標と考える．ただし，過大評価傾向を避けるため正規化しておく．

②に対して：構造中適当に定めたレベルでの値を用いると正と負の事例が混在した部分集合が多く出てきて，必要以上に複雑な決定木ができてしまう．それゆえ，

> 「現在利用しようとしている属性値のいくつかをまとめ一つの上位の値に一般化したものを使った分割でも，誤り率がそれほど増加しなければ，その分割はより簡潔な決定木を作り出せると期待できる」

ことを考慮し，分割段階での正・負事例の分布状況に適合させながら，できるだけ簡潔な結果が得られる属性値を選択していく方法を採用する．

一般化される属性値としては，**2章**での議論をもとに，それらの一般化値を用いた分割による情報獲得量が大きくなるようなものを取り上げる．この操作を表3に示す．例えば表2の a, b を正事例，c を負事例とした場合，図1を用いた一般化では，赤，紫と青からのlcgaは暖色，寒色と根になる．これらそれぞれの子孫になる値を持った事例集合は $\{a, b\}$, $\{b, c\}$ と $\{a, b, c\}$ であり，対する情報獲得量は0.92，0.25と0.0である．これより事例 a，b および c が持つ色についての値はそれぞれ暖色，暖色と青になる．全体的には，葉での値から始めて誤り率増加の割合が許容範囲内にある間一般化操作を繰り返し，最終的に必要な属性値を求めていく．

表4に階層属性を扱えるよう拡張した手続きを記す．表1では雑音のことを考えて，各節点での誤り率

表3 一般化操作

(S1) 一般化しようとしている属性値の任意の対の *lcga* を求め，これらそれぞれに対して次を行う．

(S1.1) 子孫になる値をもった事例を一つに集めた形での分割を行い，情報獲得量を計算する．

(S2) 最も大きな情報獲得量を与えた分割を取り上げ，これを作りだすもとになった事例それぞれがもつ属性値を *lcga* に置き換える．

表4 拡張手続き

：最初与えられた事例集合を C_0 とする．属性値は階層構造をしているとする．

(S1) C_0 に対応した根節点を作り，C_0 を C と考えS2を行う

(S2) 事例の集合に対し次を行う．

(S2.1) 全体的な誤り率が定められた限度内ならば各節点で正・負事例の分布に応じて正・負節点を作る*．そうでなければ局所的に誤り率の最も大きな部分集合 C についてS2.2,2.3を行う．

(S2.2) 決定木を成長させるために最適な属性を選び，そこでの分割のための属性値 $v_1, v_2, \ldots, v_m$ を表3の一般化操作を通じて定める．

(S2.3) 定めた属性値に従い C を $C_1, C_2, \ldots, C_m$ に分割する．そして，これら分割に対応した節点を作り，S2を繰り返す．

* 正(負)事例が多ければ正(負)節点とする [Quinlan 86].

が一定限度内になればそれを正あるいは負節点とみなす対処法がとられた．しかし，局所的に誤り率を見るだけではそこでの節点がサイズの大きな部分集合に対応づけられている場合，全体として見ると多くの事例が誤って判断されることになる．ここでは，誤り率といえば現時点までに作られている決定木全般にわたるものを考え，それが定められた値に収まるまで多くの正・負事例の混在した節点をさらに成長させる方法をとる．

4. 構造化事例の扱い

本章では構造化事例を扱うための方策を考える．非構造化事例から求めた決定木では，節点は事例の部分集合に，また辺は部分集合，すなわち節点，を特徴づける記述に対応していた．各記述は

$$DN_k(U) = Q(U, \delta_k^Q), \qquad (k \geq 1) \tag{1}$$

の形で書ける．ここで U は事例を指す変数，δ_k^Q は事例が持つ属性 Q の値である．兄弟節点を特徴づける記述の対については，値が階層構造中で共通の子孫を持たないよう互いに素となっている．複数の部分からなる構造化事例を扱う際には，与えられる属性が部分についてのものであることから，分割のための記述として

$$DN_k(U) = A(u_1, \delta_{k_1}^A) \wedge A(u_2, \delta_{k_2}^A) \wedge \cdots \wedge A(u_M, \delta_{k_M}^A), \quad (U = \langle u_1, u_2, \cdots, u_M \rangle) \tag{2}$$

が考えられる[Michalski 80]．ただし u_i $(1 \leq i \leq M)$ は部分を指す変数で，$\delta_{k_i}^A$ は部分が持つ属性 A の値である．

ところで形式(2)をそのまま用いると，兄弟節点特徴づけのための互いに素な記述を得ることや，上で述べた属性選択指標を利用することが困難になる．ここでは，属性 A のほかに部分特定化のための属性 B を組み入れた記述形式を採用する[伊藤 94]．具体的には，属性 B の値 δ^B_* を使って，「δ^B_* で特定化された部分全般が持つ属性 A（以後 δ^B_*-A と書く）」に関する記述

$$DN_k(U)=\delta^B_*\text{-}A(U,\delta_k^{\delta^B_*A}) \qquad (3)$$

を考える．ここで，$\delta_k^{\delta^B_*A}$ は特定化された部分(複数のときもある)が A に関して持つ値を羅列した組での値である．形式(3)は二つの属性に係わる記述と等価なため，これによれば互いに素な記述が得られやすくなる．また，δ^B_*-A を Q と見れば非構造化事例を特徴づける記述と同じ形になり，属性の選択に際しても上と同じものを利用できるようになる．

属性 B としては，これが各事例の部分を特定化するためのものであることから，値が各事例に均一に現れるようなものを選べばよい．具体的には part-of 階層がその一つと考えられる．B が定まれば値 δ^B_* をもとに，A の値の包含関係から階層属性としての δ^B_*-A を生成できる．

5. 計 算 機 実 験

(a) 実験 1

[荒木 92，表 6]のデータ（各正・負事例は五つのタイプの何れかに属し，四つの数値属性で特徴づけられている．各属性の値は決められた数値を平均とする分散 1.0 の正規分布をなす）を使い，数値属性を扱った場合の学習を試みた．実験では各タイプの事例を 20 個ずつ生成し，属性値は 0.5 きざみでの数値に変換した．決定木は全体的な誤り率が 15%になるまで成長させた．ただし，事例集合を分割していくとき，続けて同じ属性が利用されると，これらはまとめて同一レベルでの分割とみなせるため，値の種類の数が 4(事例集合の種類の数の 2 倍) 以下になるまでは一般化操作を繰り返した．

表 5 に誤り率増加の許容割合を変えた場合の学習結果 (5 回の試行に対する平均) を示す．表中の上段の値は，ここでの方法で得られた平均的な決定木の節点総数を，一次元配置された数値を扱うために開発された INDECTS[荒木 92]での最適と思われる決定木の節点総数で割ったものである．また下段の値は，ステップ S2 の繰返し数である．誤り率増加の許容割合を少し緩めた状況を考えると，INDECTS と同程度のものが得られていることがわかる（選ばれた属性の種類も同じであった）．ただし，緩めすぎると決定木を求めるコストが増える．属性値を 0.2 きざみでの数値に変換した場合も同様の傾向がみられた．結果として，ここでの方法によれば，数値属性も階層属性の一種として統一的に扱えると言える．

表 5 数値属性との比較

許容割合(%)	5	10	15	20
節点総数の比	1.20	1.16	1.16	1.08
繰返し数	3.0	3.0	3.0	3.4

(b) 実験 2

[Stepp 86]を参考にして作った図 3 に示すような図形を準備し，構造化事例についての学習効果を調べた（台車および荷物の形はそれぞれ 7 種と 6 種を用意した）．正・負事例集合は，同様の台車の形ならびに荷物の形をした部分を近くに含む図形対ほど大きな値を与える類似測度を用い，よく似たものの連鎖でつながれた図形のクラスタを二つ求めて得た[Zahn 71]．属性としては台車ならびに荷物の形に関するものとし，それらの構造は形の輪郭線が似ている割合をもとに階層的に関係づけて得た．図 4 に荷物の形の階層を示す．part-of 階層は部分の位置的な近さを参考にして作った (図 5)．これらをもとに，記述形式(3)に当てはまる属性として，おのおのの部分ならびに左側あるいは右側にある二つの部分が持つ台車の形と荷物の形を生成し*4，構造化事例が扱えるようにした．

実験では，部分の位置の近さの類似度への反映のさせ方やクラスタ間のギャップの大きさを変えた 15 の正・負事例集合(各集合は 50 の事例を含む)を用意し，決定木は全体的な誤り率が 10%になるまで成長させた．そして，扱えるのが記号属性（階層構造の葉に位

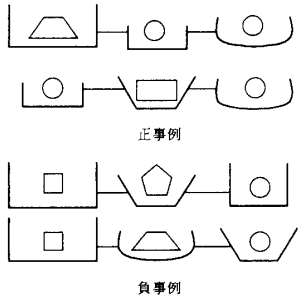

図 3 貨車の事例

＊4 三つ以上の部分が持つ属性の生成も考えられるが，その構造が複雑になるため，ここでは二つ以下の部分についてのものに限定した．

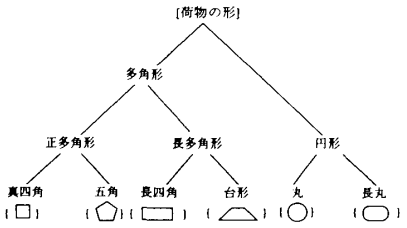

図 4 荷物の形の階層

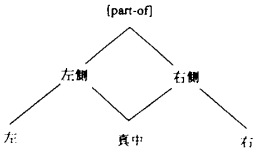

図 5 part of 階層

表 6 記号属性との比較

許容割合(%)	20	50
I	*0.64 (1.03)*	*0.61 (1.24)*
II	*0.52 (0.79)*	*0.49 (0.93)*

* 括弧内は繰返し数の比を示す

置する値だけを使う）でかつ一つの部分が持つ属性だけを取り上げた場合に対して，利用した属性の種類は同じであるが属性値間の関係を考慮した場合(I と書く）と，二つの部分が持つ属性も加えて利用した場合（II と書く）とを比較した．

表 6 に誤り率増加の許容割合を変えた状況で，得られた決定木の節点総数とステップ S2 の繰返し数の比の平均を示す．表より，誤り率増加の許容割合を少し緩めた状況では，階層属性を用いると，繰返し数を増やさず節点総数で 40%程度，また部分間の関連も考慮するとさらに 20%程度の改善ができたことが読み取れる．結果として，ここで定式化した方法が，利用できる属性ならびに属性値の幅を広げ，そのなかから場面に応じた適切なものをうまく選択していけることを示している．

6. ま と め

数値属性の扱いに関して[荒木 92, Quinlan 86, 辻野 89]では，実数値を，事例集合の分割に関し情報獲得量を最大にする区間にまとめた後，それぞれに名前をつけて記号属性とみなす方法が提案されている．ただしこれらの方法は，値の関係が複雑になる階層属性で特徴づけられた事例からの学習には利用できない．ここでは属性値の一般化操作を通じて，記号・数値属性を含めた統一的な処理を可能にする方法を定式化した．

また，従来のモデルでは十分な配慮がなされていなかった複数の部分を持つ構造化事例の扱いについても，いくつかの部分がまとまって持つ属性値を用いた記述形式を導入し，部分間の関連を保ちながら構造化事例を非構造化事例とみなせる方法を定式化した．節点を特徴づける記述形式として，複数の属性に係わる記述の連言を用いることも考えられる[Dietterich 83]．しかし，この種の記述は複数の属性と複数の部分とに言及した形をしており，これによれば，属性選択時にこれまでの指標が利用できなくなる欠点がある．また，たとえ簡潔な決定木が得られても，概念を規定する記述は複雑なものになってしまう．

提案した方法は事例集合が正と負の 2 種類からなる場合に有効となるものであった．今後の方向として，3 種以上の事例集合からの学習にも利用できる方向へ拡張することがあげられる．また，階層構造が複雑になった場合，分割に用いられる一般化値を効率良く求める探索法の定式化が必要となる．これに関連する事柄として，構造化事例を扱う際，複数の部分に係わる属性を生成すると，値が数値ならば多くの種類の組での値が出てきてしまう．これには，まず適当なきざみ幅で丸めた数値を用い，多くの正・負事例が混在する部分集合の分割時には，小さなきざみ幅で丸めたものを使うような対処法が必要となる．

◇ 参 考 文 献 ◇

[荒木 92] 荒木，小島：数値データによる決定木の帰納学習，人工知能学会誌，Vol. 7, No. 6, pp. 992-1000 (1992).

[Dietterich 83] Dietterich, T. G. and Michalski, R. S.: A comparative review of selected methods for learning from examples, R. S. Michalski, J. G. Carbonell and T. M. Mitchell (eds.), *Machine Learning*, Morgan Kaufmann, Los Altos, CA (1983).

[伊藤 94] 伊藤，葉，中島：構造化領域での生成的概念形成と記述の扱い，人工知能学会誌，Vol. 9, No. 6, pp. 917-926 (1994).

[Michalski 80] Michalski, R. S.: Pattern recognition as rule-guided inductive inference, *IEEE Trans. Pattern Anal. and Machine Intell.*, Vol. PAMI-2, No. 4, pp. 349-361 (1980).

[Quinlan 83] Quinlan, J. R.: Learning efficient classification procedures and their application to chess end games,

R. S. Michalski, J. G. Carbonell and T. M. Mitchell (eds.), *Machine Learning*, Morgan Kaufmann, Los Altos, CA (1983).

[Quinlan 86] Quinlan, J. R.: Induction of decision trees, *Machine Learning*, Vol. 1, pp. 81-106 (1986).

[Quinlan 87] Quinlan, J. R.: Symplifying decision trees, *Int. J. Man-Machine Studies*, Vol. 27, pp. 221-234 (1987).

[Stepp 86] Stepp, R. E. and Michalski, R. S.: Conceptual clustering: Inventing goal-oriented classifications of structured objects, R. S. Michalski, J. G. Carbonell and T. M. Mitchell (eds.), *Machine Learning*, Vol. II, Morgan Kaufmann, Los Altos, CA (1986).

[辻野 89] 辻野，竹之内，櫻井，千種，野村，溝口，角所：適応的ルールインダクションシステム：ARIS，信学論(D-II)，Vol. J72-D-II, No. 1, pp. 121-131 (1989).

[Zahn 71] Zahn, C. T.: Graph-theoretical methods for detecting and describing gestalt clusters, *IEEE Trans. Computers*, Vol. C-20, No. 1, pp. 68-86 (1971).

〔担当編集委員・査読者：有川節夫〕

著者紹介

中島　誠(正会員)

1987年図書館情報大学図書館情報学部卒業．1989年同大学院修士課程修了．(財)日本情報処理開発協会勤務を経て，現在，大分大学工学部知能情報システム工学科助手．認知モデル形成についての研究に従事．情報処理学会，AAAI, ACM, IEEE各会員．

葉　玲如(学生会員)

1991年図書館情報大学図書館情報学部卒業．1993年同大学院修士課程修了．現在，筑波大学大学院博士課程工学研究科在学中．研究対象は認知科学，データベース，情報ベース．情報処理学会，ACM, IEEE各会員．

伊藤　哲郎(正会員)

1970年大阪大学基礎工学部卒業．1972年同大学院修士課程修了．大阪大学助手，図書館情報大学助教授を経て，現在，大分大学工学部知能情報システム工学科教授．工学博士．研究対象はデータの連想検索，認知モデルの形成，機械学習．情報処理学会，電子情報通信学会，AAAI, ACM, IEEE各会員．