DEWS2003 7-A-04

クラス階層を用いた決定木

高光　知哉† 　三浦　孝夫† 　塩谷　勇††

† 法政大学 工学研究科 電気工学専攻　〒184–8584 東京都小金井市梶野町 3-7-2
†† 産能大学 経営情報学部　〒259–1197 神奈川県伊勢原市上粕屋 1573
E-mail: †{i02r3230,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし　本論文では簡潔な信頼できる興味深い決定木の構築方法について述べる. そのような決定木を得るために, 我々はクラス階層と選言クラスを導入する. 興味深さの評価法としてパスエントロピを定義する. 我々はこの決定木に関する実験を行い, 有効性について評価する.

キーワード　決定木　知識発見　データマイニング

Decision Trees using Class Hierarchy

Tomoya TAKAMITSU†, Takao MIURA†, and Isamu SIOYA††

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
†† Department of Management and Information Science, SANNO University　1573, Kamikasuya, Isehara city, Kanagawa 259–1197 Japan
E-mail: †{i02r3230,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract　In this work, we propose decision tree to obtain simple, reliable and interesting decision trees. To do that we introduce class hierarchy and disjunctive class. We define path entropy to evaluate interests of decision trees. We discuss some experimental results and show how useful these trees are.

Key words　Decision Trees, Knowledge Discovery, Data Mining

1. 前　書　き

分類学習は入力された訓練データからデータのクラスを予測する分類モデルを構築することである. 決定木は分類学習の一種で, 有効な手法としてさまざまな分野で使用される. 決定木学習の利点として対象となるデータに対する背景知識なしで分類モデルを自動的に生成できる. また, 木構造で分類モデルを構築するため, ルールが理解しやすい. 決定木の欠点として, 複雑なデータを用いる場合に巨大な決定木が構築されることがあげられる.

巨大な決定木はパス数が多いため複雑で理解し難い. 巨大な決定木を短くする方法として枝刈りがある. 枝刈りは決定木構築後に χ^2 を用いて各パスの誤分類率が最小になるように決定木を刈り込む. しかし, データ数が過度に大きいと効果が薄い.

我々は過去に χ^2 検定を用いて決定木全体に対する各パスの重要性と信頼性を評価する方法 [4] を提案した.

本論文では簡潔な信頼できる興味深い決定木の構築法について述べる. 簡潔な決定木は解釈するのが簡単である. 信頼できる決定木はデータの分類に役立つ. 興味深い決定木はクラス分類に関する大量の情報量を含んでいる決定木である. このような決定木の構築のために, 背景知識としてクラス階層を導入し各ルールに複数のクラスを持たせる. データに対する背景知識を与えることで信頼できる興味深い決定木を構築する. また, 各パスに複数のクラスを持たせることにより各パスの情報量が低減するが, 決定木の構築の収束が早まり簡潔な決定木を構築する.

2 章では決定木について簡単に説明する.3 章で簡潔な決定木の構築方法を例をまじえて説明する.4 章では実験の内容とその結果について示し,5 章で結びとする.

2. 決　定　木

2.1 決定木の構造とデータ

決定木構築に使用されるデータ集合は以下のような形式を持つ.

$$T = \left(\begin{array}{cccc} A_1 & \ldots & A_k : \mathcal{C} \\ \hline a_1^1 & \ldots & a_n^1 : c_1 \\ .. & \ldots & .. : .. \\ a_1^n & \ldots & a_k^n : c_n \end{array}\right) = \left(\begin{array}{c} \mathcal{A} : \mathcal{C} \\ \hline t_1 : c_1 \\ .. \\ t_n : c_n \end{array}\right)$$

各行は一つのデータを示す. これをオブジェクトという. オブジェクトは複数の属性 $A_1, A_2, ..., A_k$ と一つのクラス C を持つ. 各属性はその値 $a_1, a_2, ..., a_k$ によって, オブジェクトの特徴を示す. クラスはその特徴を持つオブジェクトの分類を示す. 決定

木はこのような形で記述されたオブジェクト集合を用いる.

決定木はノードと葉によって構成される. ノードは 1 つの属性を持つ. 葉は 1 つのクラスを持つ. 入力データはノードの持つ属性によって分岐する. 入力データはノードでの分岐を繰り返し, 最終的にひとつの葉に達する. 葉が持つクラスが入力データに対して決定木が予想したクラスとなる.1 つの葉に向かう道筋は 1 つしかない. つまり,1 つの葉に向かうための属性条件は 1 つしかない. これをパスと呼ぶ. よって, 決定木はパス集合で構成される.

例 1：図 1 に決定木に用いるデータ集合の例を示す. このデータは Race Condition についての情報を表す. 各オブジェクトは属性 Weather(Fine, Cloud, Rainy), Temperature(Very High, High, Med, Low, Very Low), WindForce(Very Windy, Windy, Breeze, Windless) とクラス Race Condition(Held,Half,No) で構成される. 各オブジェクトは属性値によって天候の特徴を表し, クラスによってその天候の Race Condition の分類を示す. 例えば,Weather が Fine で,Temperature が Mid で,WindForce が Windy という天候である場合の Race Condition は Held となる. 図 2 に図 1 のデータを用いて構築した決定木を示す. 長方形は属性を表し, 楕円はクラスを表す. 属性と属性, 属性とクラスの間にある値は属性値を表す. 一番上のノードから葉までの経路をパスという. 決定木を見ると,Temperature の値が Very High のときクラスは No という経路がある. これが 1 つのパスである. 決定木はクラスが未知なデータの分類を予測する. 例えば,Weather が Rainy で,Temperature が Very Low で,WindForce が Windy のデータが入力された場合, まず,Temperature の属性を調べる.Temperature の値は Very Low なので Very Low に分岐する. その結果, 決定木はそのデータが分類されるクラスは NO と予測する.

属性			クラス
Wether	Temperature	windforce	Race condition
Fine	Mid	Windy	Held
Fine	High	VeryWindy	Half
Fine	Very High	Windless	No
Fine	Low	Breeze	Half
Fine	Low	Breeze	Held
Cloud	Low	Windy	Held
Cloud	High	Breeze	Held
Cloud	High	VeryWindy	Half
Cloud	Low	Windless	Held
Rainy	Low	Windy	No
Rainy	Very Low	Very Windy	No
Rainy	Med	Windless	Held
Rainy	Low	Windless	Held
Rainy	Low	Breeze	Half

図 1　教師データ

2.2　決定木の構築

決定木の構築に必要なデータはクラスが明記されたオブジェクトの集合である. これを教師データという. 教師データが必要な分類モデルを教師有り学習という. 決定木のノードは各ノードに分類される教師データが持つクラスと最も関連が深い属性を持つ.

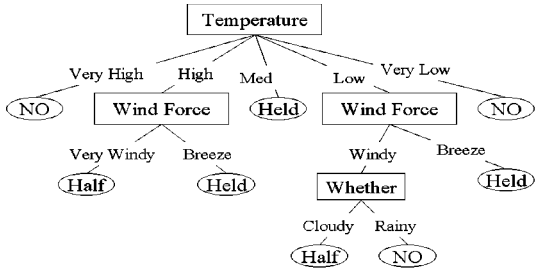

図 2　決定木

属性選択の方法としてエントロピがある. これは教師データ T を属性 A_k で分割したときのエントロピの変化を用いた基準である.n 行の入力データ T を仮定する. この基準は属性でデータを分割する前のエントロピ $E(T)$ と属性でデータを分割した後のエントロピ $E_A(T)$ の差を用いて属性を決める.$E(T)$ は次の式で計算する.

$$E(T) = \Sigma_j(-n_j/n)\log_2(n_j/n)$$

n はそのノードに分類されたデータの総数,n_j はクラス C_j に分類されるデータの数である.n_j/n はデータがクラス C_j に分類される確率である.$\log_2(n_j/n)$ は情報量と定義する. そして, すべての属性に関して, 属性 A でデータを分割したときのエントロピを計算する. 属性 A_k の持つ属性値 a_i のエントロピを次の式で計算する.

$$V_j = \Sigma_j(-n_{ij}/m_i)\log_2(n_{ij}/m_i)$$

m_i は属性値 a_i を持つデータの数,n_{ij} は属性値 a_i を持つデータがクラス C_j に分類されるデータの数である. 属性 A_k が持つすべての属性値に関するエントロピを次の式で計算する.

$$E_A(T) = \Sigma_i(m_i/n)V_j$$

これを用いてノードに分類されたデータをすべての属性で分割したときの $E_A(T)$ を求める. ノードが持つ属性を決める $gain$ を $G_A(T)$=$E(T)$-$E_A(T)$ と定義する.$G_A(T)$ が最大になる属性がそのノードが持つ属性として選択される. これを繰り返し, ノードに分類されたデータを属性で分割できないとき, そのノードは葉となる. このとき, そのノードに分類されたデータが最も多く持つクラスが葉のクラスとなる.

例 2：図 1 のような Race Condition を示す教師データから決定木を作る.$E(T)$ を計算すると,$E(T)$=$-(7/14)\log_2(7/14) - (4/14)\log_2(4/14) - (3/14)\log_2(3/14)$=1.49261. 各属性に関して $E_A(T)$ を計算すると,

$E_{Weather}(T)$=1.31889

$E_{Temperature}(T)$=0.886169
$E_{WindForce}(T)$=0.911063
よって, 一番効率良くデータを分類できる属性は Temperature となる. 同様に繰り返していくと, 最終的には図 2 のような決定木を構築する.

2.3 決定木の評価

一般的に, 構築した決定木は教師データとは別のオブジェクト集合を用いて評価する. これをテストデータという. テストデータを決定木を用いて分類し, 各オブジェクトが持つクラスと各オブジェクトを決定木で分類して得たクラスを比較することで評価する. 評価の方法としては誤分類率がある. これはオブジェクトが持つクラスと決定木で分類して得たクラスが不一致であるオブジェクトの割合で評価する. つまり, 決定木のクラス予測が間違った割合で評価する. 誤分類率が低ければ低いほど, その決定木はデータを正確に分類できるため, 予測精度が高い決定木となる.

我々は以前に決定木の評価法として χ^2 検定を用いた評価法 [4] を提案した. これは, 決定木構造に対して影響力が強いパスに対して注目する. 影響力の強いパスは入力データのクラス決定に対して重要な役割を持っている. また, 各パスを 1 つのクラスに断定せずにクラス分布を用いることによって, 分類予測が間違う可能性を考慮に入れて入力データを予測し, 決定木を評価する. この評価法は独立性の検定と累積 X^2 を用いて決定木全体に対するパスの影響力の評価を行い, 適合度検定でパスのクラス分布とテストデータの分布の一致性を評価する.

3. 簡潔な信頼できる興味深い決定木の構築

3.1 簡潔な信頼できる興味深い決定木

決定木は教師データを細かく分類しようとするために巨大な決定木を構築する場合がある. このような決定木は信頼できるが, パスが複雑すぎてわかり難い. しかし, 簡潔すぎる決定木は予測が正確でなく, 分類に関する情報量をあまり持っていないため興味深くない.

我々は簡潔な信頼できる興味深い決定木を構築するためにデータに対する背景知識の導入と各パスの葉に複数のクラスを持たせる. データに対する背景知識を得ることによって信頼できる興味深い決定木を構築する. 各パスの葉に複数のクラスを持たせることは, パスの分類能力と情報量が減少する. しかし, 決定木の収束は早くなるので簡潔な決定木を構築できる.

3.2 階層エントロピ

我々は決定木構築時に背景知識としてクラス間の関係を示す一つの階層構造を与える. これをクラス階層という. クラス階層はノードと枝で構成される. ノードはクラスを表す. 枝はクラス間の親子関係を表す. 親クラスを C_1, 子クラスを C_2 としたとき, この二クラスの関係は $C_1 \supseteq C_2$ となるので, 親クラスに分類されるデータの合計は子クラスに分類されるデータの合計を含む. 各クラスは一つの親クラスと複数の子クラスを持つ. ただし, 子クラスは持たなくてもよい. また, 最も上に位置するクラスには親クラスは存在しない. このクラスをルートクラスと定義する. ルートクラスは他のすべてのクラスを含むので, 分類されるすべてのデータを含む.

我々はクラス階層を利用することで簡潔な信頼できる興味深い決定木を構築するために, エントロピをクラス構造を考慮にいれて計算するように定義しなおす. これを階層エントロピ $E(T)$ と定義する.n 行の入力データ T と一つのクラス階層を仮定する. クラス階層のルートクラス c_0 はすべてのデータ T を含むので,$E(T)$=$E(c_0)$ である. 階層エントロピ $E(T)$ は以下のように定義する.

$$E(T)=E_1(T)+E_2(T)$$

$E_1(T)$ と $E_2(T)$ は次のように定義する.

$$E_1(T)=\Sigma_j(-l_j/l)\log_2(l_j/l)$$

l はそのクラスに分類された全データ数,l_j はそのクラスの子クラス c_j に分類されたデータ数である.

$$E_2(T)=\Sigma_j(l_j/l)V_j$$

V_j は $E(T_j)$ を意味し, 子クラス c_j に関する階層エントロピである. 子クラス c_j を持たないクラスの場合, そのクラスに関するエントロピは 0 である [3]. 属性 A でデータを分割したときの階層エントロピ $E_A(T)$ は次のように定義する.

$$E_A(T)=\Sigma_i(m_i/n)E(T|a_i))$$

n は全入力データ数,m_i は属性 A の属性値 a_i を持つデータ数,$E(T|a_i)$ は属性 A の属性値 a_i を持つデータ数 m_i を用いた階層エントロピである. これを用いてすべての属性の $E_A(T)$ を求める. ノードが持つ属性を決める $gain$ を $G_A(T)$=$E(T)$-$E_A(T)$ と定義する.$G_A(T)$ が最大になる属性がそのノードが持つ属性として選択する.

階層エントロピは次のような性質を持つ. 1 つの親クラスといくつかの子クラスからなる 1 レベルのクラス階層を仮定する. すべてのデータがこのクラス階層の子クラスのみに所属していた場合, 階層エントロピはエントロピと等しい. いくつかのクラスからなる 2 レベル以上のクラス階層を仮定する. 一番上の親クラスと一番下にいくつかの子クラス以外のクラスに所属するデータがない場合, そのクラスを取り除いて計算した階層エントロピと取り除かないで計算した階層エントロピは等しい.

例 3：図 3 のクラス階層を持つデータがある. あるノードに分類されたデータが持つクラスを () 内に示す.() 外の数字は子クラスのデータ数を含んだデータの数である. 例えば,C_1 は子クラス C_2 と C_3 を持つデータの数を含む. このときのクラス階層エントロピを求めると,

$E(C_4)$=$E(C_5)$=$E(C_6)$=$E(C_7)$=0
$E(C_3)$=$-(1/3)\log_2(1/3)-(1/3)\log_2(1/3)$=1.05664
$E(C_2)$=$-(2/4)\log_2(2/4)-(1/4)\log_2(1/4)$=1

$E(C_1)$=$-(4/8)\log_2(4/8) - (3/8)\log_2(3/8) + ((4/8)E(C_2) + (3/8)E(C_3)$=1.92688
となり,$E(T)$=1.92688 となる.

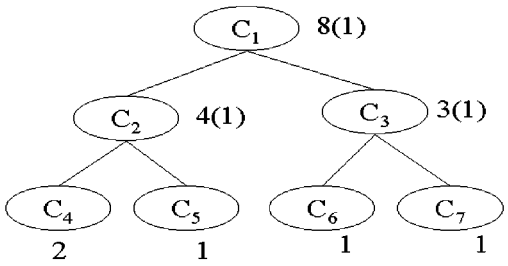

図 3　クラス階層

例 4：図 4 のようなクラス階層の各クラスに最も下位に位置する子クラスのみを持つデータが分類されているとする. これはクラス階層がない場合と等価でなければならない. 階層エントロピとエントロピの値を比較する. 階層エントロピは
$E(C_3)$=$E(C_2)$=$-(3/4)\log_2(3/4) - (1/4)\log_2(1/4)$=0.81128
$E(C_1)$=$-(4/8)\log_2(4/8) - (4/8)\log_2(4/8) + ((4/8)E(C_2) + (4/8)E(C_3)$=1.81128
エントロピは
$C4$,$C6$:$(4/8)(3/4)\log_2(3/4) = -0.53064$
$C5$,$C7$:$(4/8)(1/4)\log_2(1/4) = -0.375$
$E(T)$=-C_4-C_5-C_6-C_7=1.81128
となり, 階層エントロピはエントロピと等しい.

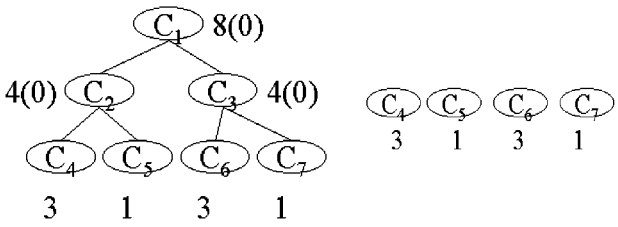

図 4　階層エントロピとエントロピ

例 5：図 1 のデータに図 5 のようなクラス階層を与えて決定木を作成する. ルートノードに分類されるデータの数は 14 件で, データが持つクラスの分布を見ると DontCare=0,Held=7,Half=4,No=3 である. クラス階層を考慮したクラスの分布に直すと DontCare=14,Held=11,Half=4,No=3 である. クラス階層エントロピ $E(T)$ を計算すると,
$E(DontCare) = 1.16658$
となる. 次に各属性でデータを分割した場合の階層エントロピ $E_A(T)$ を計算すると,
$E_{Weather}(T)$=1.003525
$E_{Temperature}(T)$=0.605826
$E_{WindForce}(T)$=0.768206
$E(T)$-$E_A(T)$ が最大になる属性は Temperature となり, ノードの属性として選択する.

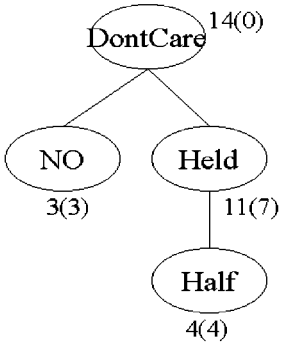

図 5　Race Condition のクラス階層

3.3　情報量の低減

従来の決定木では各葉が持つクラスの数は一つである. そのため, データ数が多い場合, もしくはクラス数が多い場合, もしくは属性数が多い場合, 決定木が巨大になる傾向がある.

本論文では, 各葉が持つクラスを複数にする. その結果, 各パスの情報量が低減するが, 簡潔な決定木を構築することができる. 葉が持つ複数のクラスを選言クラスと定義する.

クラスを複数持つための条件として, 選言基準 β, 最大クラス数 γ を定義する. 選言基準 β はノードを葉に置き換える基準となるデータの割合を示す. そのため,β は 0 から 1 の間の値をとる. ある選言クラスを持つデータのノード内の全データに対する割合が β 以上になったとき, そのノードは葉に置き換わる. 最大クラス数 γ は, 選言クラスに用いるクラスの最大数を示す.β,γ を超えない範囲で各クラスを組み合わせて選言クラスを構築する.β の値が低い場合,β を上回る選言クラスが二つ以上できる可能性がある. その場合,β を上回った選言クラスの用いたクラス数が最小である選言クラスを葉として選択する. それでも選言クラスの候補が複数残る場合, クラス数が最小である選言クラスの中で最も大きな β を持つ選言クラスを葉として選択する.

葉が持つクラスはクラス階層の影響を受ける. 上の階層に位置するクラスはそのクラスの子以下の階層に位置するクラスを含むので, あるパスに葉が持つクラスより子以下の階層に位置するクラスが分類された場合, 正しい分類であると判断する.

選言クラスは新たなクラスの生成を意味する. 新たなクラスを階層に組み込むため, クラス階層を書き換える必要がある. 選言クラスに用いたクラス間で最も近い共通の上の階層に位置するクラスを探す. そのクラスの子クラスとして選言クラスをクラス階層に追加する選言クラスに用いたクラスは階層から削除し, 削除したクラスの子以下の階層は選言クラスの子クラスとして追加する. 各パスによって葉が持つクラスは違うため、各パス固有のクラス階層を持つ.

例 6：選言基準 $\beta = 0.9$, 最大クラス数 $\gamma = 2$ として, 図 1 のデータに図 5 のようなクラス階層を与えて決定木を作成する. ルートノードに分類されるデータは 14 件で, データが持

つクラスの分布は Held=7,Half=4,No=3 である. データが多い Held,Half を選言クラスとすると, ノードに分類されたデータに対する割合は 11/14 となり,$\beta = 0.9$ 以下となる. よって, ルートノードは葉に置き換わらない. そのため, 分岐するための属性を選択する. ルートノードの属性は例 5 で計算したように Temperature を選択する.Temperature で教師データを分割すると, 図 6 のようになる.Very High, Mid, Very Low はデータが持つクラスが 1 つしかないのでクラスが確定する.High は Held と Half を用いた選言クラスが β,γ の条件を満たすのでクラスが確定する. 選言クラスを構築したのでクラス階層を書き換える.Half は Held の最も近い共通の上の階層に位置するクラスは DontCare である. その子クラスとして選言クラスを配置し,Held と Half の子以下のクラスを選言ノードの子以下として置き換え,Held と Half を削除する. その結果, 図 8-2 のような階層構造になる.

葉が確定しなかったノードに対して同様に階層エントロピと選言クラスの適用を繰り返す. その結果, 図 7 のような木が構築される. 各パスが持つクラス階層は図 8 に示す. 図 7 の各パスの下に書いてある数字が図 8 の各クラス階層に対応する.

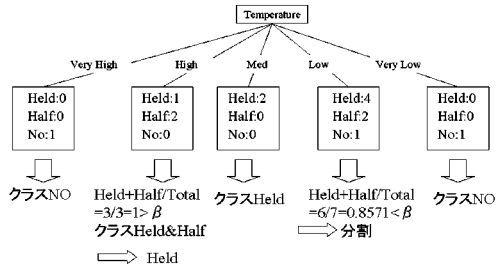

図 6　属性選択後のデータ分布

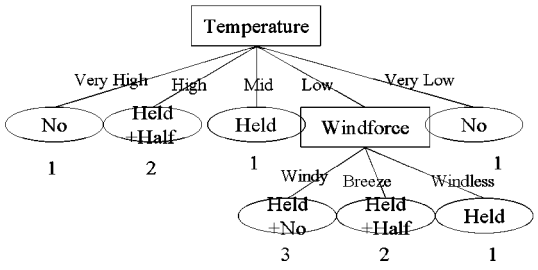

図 7　　決定木

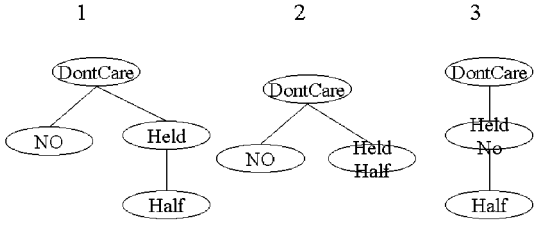

図 8　クラス階層

3.4　決定木の評価

選言基準 β, 最大クラス数 γ の定義によって簡潔な決定木を構築できる. しかし, あまりに簡潔すぎる決定木は選言クラスに属するという意味では信頼できるが, 情報量が減りすぎているため興味深い決定木ではない. そのため, 興味深さという尺度で決定木を評価する方法を定義する.

我々は決定木の評価としてパスエントロピと誤分類率を定義する. これらはテストデータを用いて計算する. パスエントロピは興味深い決定木かどうかを表す. これは各パスに分類されたテストデータを用いて各パスのパスエントロピを計算し, それを合計した値で評価する. この値が低い場合, その決定木が持つ各パスに分類されたテストデータは特定の 1 つのクラスに偏る. そのため, その決定木はクラス分類に関する情報量が大きいので興味深い. パスエントロピ $P(T)$ は次の式を用いて計算する.

$$P(T)=\Sigma_j(n_j/n)E(P_j)$$

n はテストデータの総数,n_j はパス P_j に分類されたテストデータの数,$E(P_j)$ は各パスに分類されたテストデータを用いて計算された階層エントロピである. このとき用いられるクラス階層は最初に与えられたクラス階層である.

誤分類率は決定木がどれだけデータを間違って分類したかを表す. この値が低ければ, その決定木は正確な分類ができている. クラス階層を用いた決定木の場合, 誤分類率も階層の影響を受ける. クラス階層がある場合, 親クラスは子クラスを含むため, 子クラス以下をもつデータを正しいと判断する.

例 7：選言基準 $\beta = 0.9$, 最大クラス数 $\gamma = 2$ として, 図 1 のデータに図 5 のようなクラス階層を与えて構築した図 7 の決定木にテストデータ図 9 を分類させる. テストデータの分類結果を図 10 に示す. 図 5 のクラス階層を考慮した誤分類データの数は 4 件である. よって, 誤分類率は 4/22=0.181818 となる. この決定木のパスエントロピを計算する. 例として,Temperature-VeryHigh → NO のパスエントロピを計算する. このパスに分類されるテストデータの数は 4 件で, データが持つクラス分布は Held=1,Half=1,No=2 である. クラス階層を考慮したクラス分布に直すと Held=2,Half=1,No=2 である. このパスのエントロピは $E(P) = -(4/4)(2/4)\log_2(2/4) - (2/4)(1/2)\log_2(1/2) - (4/4)(2/4)\log_2(2/4) = 1.25$ となる. このパスのパスエントロピは 1.254/22=0.227273 となる. 他のパスについても同様に計算してすべてのパスのパスエントロピを合計する. その結果, この決定木が持つパスエントロピは 0.682368 となる.

4.　実　　験

4.1　実験手順と実験結果

気象データを用いて階層エントロピを用いた決定木の実験を行う. 教師データとして 1997 年 1 月,1998 年 1 月,1999 年 1 月の気象データ 5297 件を用いる. テストデータは 2000 年 1 月の気象データ 1772 件を用いる. 気象データは属性として平均気温, 最高気温, 最低気温, 平均湿度, 最低湿度, 瞬間最大風速, 日射量,

属性			クラス
Weather	Temperature	windforce	Race Condition
Fine	Low	Windy	No
Cloud	High	Breeze	Half
Fine	Mid	Windy	Held
Cloud	Very High	Windy	No
Rainy	Very Low	Windless	No
Cloud	Mid	Windy	Half
Fine	Very High	Very Windy	Held
Fine	High	Breeze	Half
Cloud	Very High	Windless	No
Rainy	Low	Breeze	No
Fine	Very Low	Breeze	No
Fine	High	Very Windy	Held
Fine	Mid	Windy	Held
Cloud	High	Windless	Half
Cloud	High	Breeze	Half
Cloud	Very High	Breeze	Half
Rainy	High	Windy	Half
Rainy	Low	Windy	Half
Rainy	Mid	Breeze	No
Fine	Very Low	Windy	No
Fine	High	Windy	Held
Fine	Low	Windy	Half

図 9　テストデータ

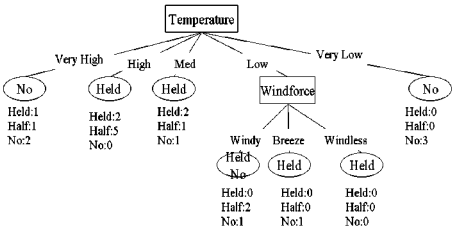

図 10　テストデータの分類結果

全天日射量, 降水量, 降雪の深さ, 最深積雪の 11 の属性を持ち, クラスとして地方名を持つ. 属性値は一定の範囲でグループ化することで数値から非数値に直す. また, 各データの属性値には不明な値は存在しない. クラスは図 11 の階層構造を持つ. 気象データ例を図 12 に示す. このデータを用いて $\beta = 0.51, 0.7, 0.9$ と $\gamma = 1, 2, 3, 4, 5$ と値を変えて決定木を構築する. また,C4.5 を同じデータで実行し, その結果と比較する.C4.5 は階層構造を用いず設定をディフォルトで実行する.

図 13 に β と γ の値を変えて決定木を構築したときのパスの数と誤分類率とパスエントロピの値を示す. また,C4.5 の実行結果も同様に示す. 決定木のサンプルとして,$\beta = 0.7$, $\gamma = 5$ のときの決定木の構築結果を図 14 に示す.

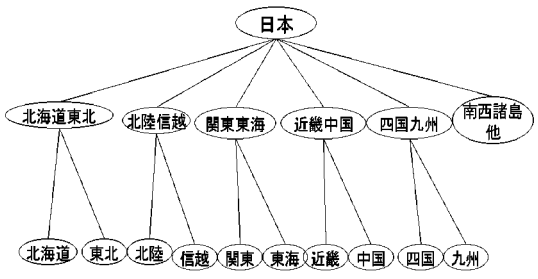

図 11　クラス階層

4.2　考　　察

図 13 を見ると,γ 値が増加すると決定木が持つパスの数が減少し, パスエントロピが増加し, 誤分類率が減少する. また,β が増加すると決定木が持つパスの数は増加し, パスエントロピは減少する. しかし, 誤分類率はほとんど変化しない.

図 15 に選言クラスの最大クラス数 γ に対する誤分類率とパスエントロピの変化を示す. これを見ると,γ が増加すると誤分類率は減少する. これはクラス階層と選言クラスを導入したことにより, 正しいと判定されるクラスが増加したためだと考えられる. また,β が増加するとパスエントロピが減少する. これは決定木が正しくデータを分類したときに少ないパスエントロピを得るためだと考えられる. そのため、パスエントロピが小さい決定木は興味深い木である. これより, パスエントロピの大きさの要因になるのは γ といえる.

図 16 に選言クラスの最大クラス数 γ に対するパスの数とパスエントロピの変化を示す. これを見ると,γ が増加するとパス

平均気温	最高気温	最低気温	平均湿度	最低湿度	瞬間最低風速	日射量	全天日射量	降水量	降雪の深さ	最深積雪	地区
-5~0	-5~0	-5~0	55~70	40~55	10~15	0~1	3~6	0~5	0~5	40~60	北海道
10~15	15~20	5~10	55~70	0~40	25~	4~6	6~9	0~5	0	0	近畿中国
5~10	10~15	0~5	70~85	40~55	5~10	6~8	6~9	0~5	0	0	東海
15~20	20~	10~15	70~85	40~55	10~15	0~1	3~6	20~25	0	0	南西諸島
5~10	15~20	0~5	70~85	40~55	5~10	6~8	6~9	0~5	0	0	東北
5~10	10~15	-5~0	55~70	0~40	15~20	6~8	9~12	0~5	0	0	信越
10~15	20~	0~5	70~85	40~55	25~	2~4	6~9	5~10	0	0	九州
0~5	5~10	0~5	55~70	40~55	15~20	2~4	6~9	0~5	0	0	関東東海
10~15	15~20	5~10	70~85	40~55	15~20	2~4	6~9	0	0	0	関東

図 12　気象データ例

β	γ	パスの数	パスエントロピ	誤分類率
0.51	1	1333	0.975129	0.604433
0.51	2	249	1.81999	0.474644
0.51	3	19	2.33476	0.38544
0.51	4	8	2.48689	0.380926
0.51	5	1	3.37011	0.37754
0.7	1	2010	0.657139	0.608597
0.7	2	944	1.13728	0.487273
0.7	3	314	1.67327	0.388477
0.7	4	79	2.16907	0.309819
0.7	5	15	2.37429	0.25
0.9	1	2387	0.525469	0.613765
0.9	2	1518	0.804537	0.513585
0.9	3	1034	1.04057	0.426129
0.9	4	645	1.27115	0.35277
0.9	5	347	1.60641	0.26504
C4.5		1769	1.36877	0.629797

図 13　実験結果

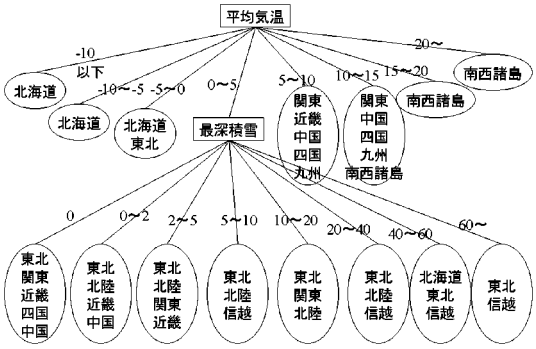

図 14　構築した決定木

の数が大幅に減少する. 図 15 の結果より, 簡潔な決定木はパスエントロピの量が大きい. つまり, 簡潔な決定木は興味深くない木といえる.

C4.5 と比較すると,$\gamma=1$ ときの木は C4.5 とほとんど同一である. しかし, パスエントロピを見てみると,C4.5 で構築した決定木はパスの数が最も多いわけではないにもかかわらず, 最も大きいパスエントロピを持つ.γ¿ 1 の場合,C4.5 より少ないパスを持つ決定木を構築するが, パスエントロピの量は増加する. しかし,γ¿ 1 β=0.9 は C4.5 より少ないパスエントロピを最も長く維持する.

これらの結果から, 高い信頼性 β を必要とするとき, 我々の手法で構築した決定木は簡潔な（パスの数）興味深い面（小さいパスエントロピ）も維持することができる.

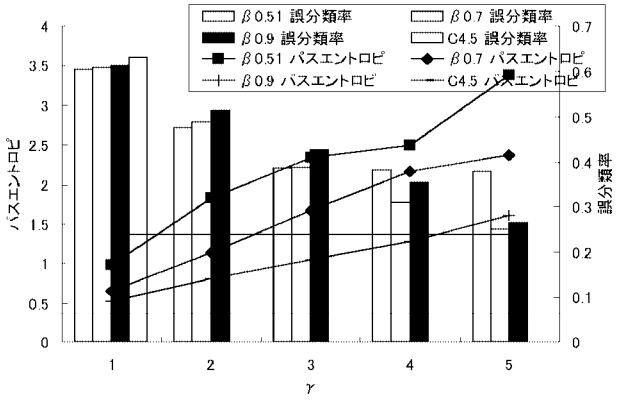

図 15　誤分類率とパス情報量

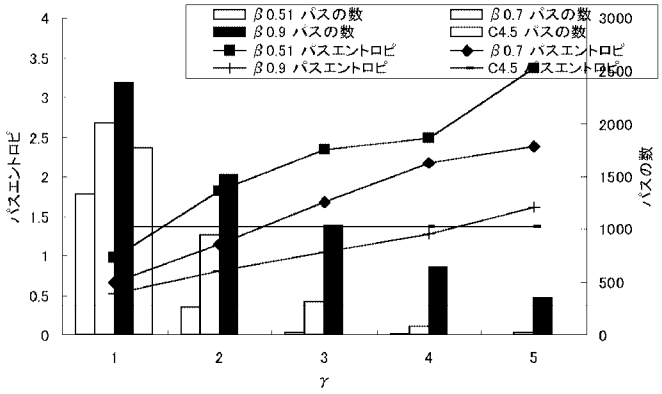

図 16　パスの数とパス情報量

5. 結　　び

本論文ではクラス階層と選言クラスを用いた簡潔な信頼できる興味深い決定木の構築法を提案した. また、決定木の評価法としてパスエントロピーを提案した.

今後の課題として, 数字と値が不明な属性値に対しての対応について考えている.

謝　　辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 14580392) の支援による.

文　　献

[1] 古川康一: AI によるデータ解析, 株式会社トッパン (1995)
[2] 古川康一, 尾崎知伸, 植野研: 帰納論理プログラミング, 共立出版 (2001)
[3] I. Shioya, T. Miura: Knowledge Pruning in Decision Trees , proc. IEEE *ICTAI* Conference 2000
[4] T. Takamitsu, I. Shioya, T. Miura: Testing Structure of Decision Trees , proc. SCS *ISE* Conference 2002