人 口 知 能 84－4
（1992. 9. 10）

あいまい性を含んだ訓練事例からの学習

櫻井茂明, 荒木大
株式会社 東芝
システム・ソフトウェア技術研究所

J.R.Quinlan のID3アルゴリズムは、訓練事例から判断知識を表現する決定木を生成する有効なアルゴリズムである。しかしながらID3アルゴリズムには、事例を表現する属性の値として、互いに独立した離散的な属性値しか取り扱えないという適用上の制限があった。本研究ではあいまいな属性値を含む訓練事例から判断知識を生成する方法を提案する。この方法は、あいまいな値を含む属性に対して、分類クラスの判別の観点で有効な分類を行なえるファジイ集合を生成し、このファジイ集合により訓練事例集合をファジイ分割することにより、判断規則を生成することを特徴としている。このアルゴリズムの有効性を数値実験により検証する。

Learning from imprecise training samples

Shigeaki Sakurai, Dai Araki
TOSHIBA Coporation
Systems & Software Engineering Laboratory
70, Yanagi-cho, Saiwai-ku, Kawasaki, 210 Japan

ID3 algorithm can acquire an efficient decision tree from training samples. ID3 can deal with the training samples that have the distinct and independent attribute values, but it can't deal with the ones that have the imprecise attribute values. In this study, we propose an approach to acquire the classification knowledge from the imprecise training samples. This new algorithm creates some fuzzy sets, which are effective to discriminate the class values, for classifing the imprecise attribute values. This algorithm divides the training samples with these fuzzy sets. We examined the efficiency of this approach by some numerical experiments.

1 はじめに

近年の人工知能の研究において、判断知識を帰納的に獲得する機械学習の技術が広く研究されつつある。その一つであるID3アルゴリズム[1]は、事例を表現する属性とそれに対して与えられた分類クラスを訓練事例として、多数の訓練事例から属性と分類クラスの間の一般的な判断規則を生成する手法であり、すでに実用レベルのルール学習システムとしてツール化が行なわれている。ところが、ID3アルゴリズムには訓練事例の属性として、互いに独立した離散的な値で表現された属性しか取り扱えないという適用上の制限があった。現実世界で与えられる学習用の訓練事例には、数値やあいまい性を含んだ訓練事例が多数存在する。従って、数値やあいまい性を含んだ訓練事例を取り扱えるようにアルゴリズムを拡張する必要がある。

この問題に対する一つの解決策として、分岐ノードであいまいな判断を行なう決定木(ファジイ決定木)を提案し、数値やあいまい性を含んだ訓練事例を取り扱えるようにした。さらに、ファジイ決定木を学習する方法を提案した[2]。ファジイ決定木による推論では、入力された事例は、分岐ノードで確信度付きで複数のノードに伝播され、最終的に末端ノードに到達した事例の確信度の合計で分類クラスの判断を行なう。従って、従来の決定木のような択一的な判断は行なわれず、確信度が付いた結論が複数出力される。

しかしながら、[2]で提案したファジイ決定木の学習方法では、分岐ノードでの判断方法を表現するために必要なファジイ集合(ファジイ分岐判断項目)を外部から与える必要があった。一方、ファジイ決定木により表現される判断規則は、与えられたファジイ分岐判断項目に依存したものとなる。従って、有効な判断規則を生成するには、適切なファジイ分岐判断項目を与える必要があった。

本稿では、ファジイ分岐判断項目を訓練事例から生成するアルゴリズムを提案する。このアルゴリズムをファジイ決定木の学習方法に組み込むことにより、ファジイ分岐判断項目を外部から与えなくても、ファジイ決定木を生成することができる。

本稿を以下のように構成する。2章では、決定木とその学習方法であるID3アルゴリズムについて簡単に説明する。3章では、[2]で提案したファジイ決定木とその構成方法について簡単に説明する。4章では、ファジイ分岐判断項目を生成する方法について説明する。5章では、提案した方法の有効性を示すべく行なった数値実験の方法と実験結果について説明する。

2 決定木とID3アルゴリズム

本章では、従来の決定木とその代表的学習アルゴリズムであるID3アルゴリズムについて簡単に説明する。

2.1 決定木

決定木は、与えられた事例を分類するための手続きを表現したものであり、分類クラスがラベル付けされた末端ノードと、与えられた事例が持つ属性の値で事例を分類する分岐ノードから構成される。分岐ノードには判断に用いる属性がラベル付けされ、その属性値に対応して、それ以降の分類手続きを表現する部分木が連結される。決定木を用いた分類は、決定木の最上部にある分岐ノードから順次分岐ノードで判断を行ない、下位のノードに降りていくことにより行なわれる。最終的には、末端ノードで分類クラスが判定される。

2.2 ID3アルゴリズム

この節では、ID3アルゴリズムについて簡単に説明する。

ID3アルゴリズムに対して与える訓練事例は、事例を特徴付けるいくつかの属性と、分類クラスによって表現される。

すなわち、n_A 個の属性集合 $A = \{A_1, A_2, \cdots, A_{n_A}\}$ に対して、属性値の集合を $Range(A_i)$ とし、n_C 個の分類クラスの集合を $C = \{c_1, c_2, \cdots, c_{n_C}\}$ とすれば、訓練事例は、

$$(v_1, v_2, \cdots, v_{n_A}, c_k) \tag{2.1}$$

ここで

$$v_i \in Range(A_i)$$
$$c_k \in C$$

と表される。この訓練事例は「属性 A_1 の属性値が v_1, 属性 A_2 の属性値が v_2, $\cdots$, 属性 A_{n_A} の属性値が v_{n_A} ならば、分類結果は c_k である。」と解釈する。このような訓練事例を多数集めて、以下に示す手続きを順次適用することによって、決定木を生成することができる。ただし、ここで説明するID3アルゴリズムでは、訓練事例に含まれるノイズを考慮して、分岐ノードにおける判断属性の選択時に、

ノードに属する事例集合で同一の分類クラスを有する事例の占有率の最大値が設定値を越えた時に、ノードの分割を終了して、末端ノードとする枝刈りを行っている。

1. すべての訓練事例を対応付けたノード S_0 を生成する。
2. ノードに属する訓練事例のうちで、同一の分類クラスを持つ訓練事例の割合が設定値以上となれば、そのノードは末端ノードとし、設定値より小さければそのノードは分岐ノードとする。
3. 分岐ノードに対して、
   (a) 各属性 $A_i, (i = 1, 2, \cdots, n_A)$ の相互情報量を計算し、最大の値を取る属性 A_α を、その分岐ノードにおける判断属性とする。
   (b) ノードに属する訓練事例を A_α の属性値 $v_{\alpha j}, (j = 1, 2, \cdots, n_\alpha)$ によって、n_α 個の部分集合 $S_{v_{\alpha j}}$ に分割する。
   (c) 個々の $S_{v_{\alpha j}}$ に対して、新しいノードを生成する。このとき元のノードと新たに生成したノードを結ぶリンクには、対応する属性値をラベル付けする。
4. 新しく生成したノードに対して、ステップ 2. からの手続きを適用する。

上記の手続き中で用いられる相互情報量の計算方法について、簡単に説明する。

ノードに属する事例集合 S の中で、分類クラスが $c_k \in C$ となる事例集合を S_{c_k}、ノードに属する事例集合 S の中で、属性値が $v_{ij} \in Range(A_i)$ となる事例集合を $S_{v_{ij}}$ とする。このとき、属性 A_i で事例集合 S を分割したときの相互情報量 $gain(A_i, S)$ は

$$gain(A_i, S) = I(S) - E(A_i, S) \qquad (2.2)$$

ただし、

$$I(S) = -\sum_{c_k \in C} p_{S,c_k} \cdot \log_2 (p_{S,c_k})$$

$$p_{S,c_k} = \frac{|S_{c_k}|}{|S|}$$

$$E(A_i, S) = \sum_{v_{ij} \in Range(A_i)} p_{S,v_{ij}} \cdot I(S_{v_{ij}})$$

$$p_{S,v_{ij}} = \frac{|S_{v_{ij}}|}{|S|}$$

ここで、記号 $|\cdot|$ は集合の要素の数を表す。

となる。

3 ファジイ決定木とファジイ決定木生成アルゴリズム

本章では、ファジイ集合理論 [3][4] により、表現形式を拡張したファジイ決定木と、その学習アルゴリズムについて簡単に説明する。詳しくは [2] を参照されたい。

3.1 ファジイ決定木

ファジイ決定木は、分岐ノードにおいて、属性領域をあいまいに分類するファジイ集合 (ファジイ分岐判断項目) をラベル付けし、末端ノードにおいて、確信度の付いた分類クラスをラベル付けすることを特徴とする決定木である。

ファジイ決定木を用いた分類は、ファジイ決定木の最上部にある分岐ノードから順次分岐ノードであいまいな判断を行ない、確信度を持った事例を下位のノードに伝播していくことにより行なわれる。最終的には、末端ノードに到達した事例の確信度の合計で分類クラスの判断を行なう。

3.2 ファジイ決定木生成アルゴリズム

ファジイ決定木生成アルゴリズムに対して与える訓練事例は、訓練事例を特徴づけるいくつかの属性と、確信度の付いた分類クラスによって表現される。すなわち、(2.1) に訓練事例の確信度 p を付加した、訓練事例は、

$$(v_1, v_2, \cdots, v_{n_A}, c_k, p)$$

と表現される。また、v_i はメンバーシップ関数 $m_{v_i}(x)$ を持つファジイ集合を意味する。このメンバーシップ関数が事例の持つあいまい性の表現となる。

一方、各々の属性 A_i に対して、n_i 個のファジイ分岐判断項目 $F_i = \{f_{i1}, f_{i2}, \cdots, f_{in_i}\}$ を設定する。ただし、各訓練事例が持つ情報量を均一に取り扱うため、このファジイ分岐判断項目のメンバー

シップ関数 $m_{f_{il}}(x)$ が、F_i のすべての要素 x において、次の条件を満足するように設定する。

$$\sum_{f_{il} \in F_i} m_{f_{il}}(x) = 1 \qquad (3.1)$$

以下に示す手続きにより、ファジイ決定木を生成することができる。

1. すべての訓練事例を対応付けたノード S_0 を生成する。
2. ノードに属する訓練事例 S のうちで、同一の分類クラスを持つ訓練事例の確信度の和の割合が設定値以上となれば、そのノードは末端ノードとし、設定値より小さければそのノードを分岐ノードとする。
3. 分岐ノードに対して、
   (a) 各属性 A_i, $(i = 1, 2, \cdots, n_A)$ の相互情報量を計算し、最大の値を取る属性 A_α を、その分岐ノードにおける判断属性とする。ただし、A_i の相互情報量は次の手順で計算する。
      i. A_i に対して与えられた n_i 個のファジイ分岐判断項目 f_{il}, $(l = 1, 2, \cdots, n_i)$ に対する訓練事例 t_j の帰属度 $M_{f_{il}}(t_j)$ を計算する。
      ii. ファジイ分岐判断項目 f_{il} に対する訓練事例 t_j の確信度 $p_{f_{il}}(t_j)$ を計算する。
      iii. ファジイ分岐判断項目 f_{il} の出現確率 $p_{S,f_{il}}$ を計算する。
      iv. 事例集合 S に対する分類クラス c_k の出現確率 p_{S,c_k} を計算する。
      v. 属性 A_i の相互情報量を計算する。
   (b) 判断属性 A_α のファジイ分岐判断項目 $f_{\alpha l}$ によって、分岐ノードに属する訓練事例を n_α 個のファジイ部分集合 $S_{f_{\alpha l}}$ に分割する。
   (c) 個々の $S_{f_{\alpha l}}$ に対して、新しいノードを生成する。このとき元のノードと新たに生成したノードを結ぶリンクには、対応するファジイ分岐判断項目をラベル付けする。
4. 新しく生成したノードに対して、ステップ 2. からの手続きを適用する。

上記手続き中で用いられる属性の相互情報量の計算方法について詳しく説明する。

ステップ 3(a) i. では、ファジイ分岐判断項目 f_{il} に対する訓練事例 $t_j \in S$ の帰属度 $M_{f_{il}}(t_j)$ を計算する。$M_{f_{il}}(t_j)$ を、ファジイ集合 f_{il} とファジイ集合 v_{ij} の積集合の最大帰属度と訓練事例の確信度の最小値として与えることにする。すなわち、

$$M_{f_{il}}(t_j) = \vee_x \{ m_{f_{il}}(x) \wedge m_{v_{ij}}(x) \} \wedge p_j \qquad (3.2)$$

ここで、$\vee, \wedge$ は、それぞれ max, min 演算を表す。

と計算する。

ステップ 3(a) ii. では、t_j の f_{il} に対する確信度 $p_{f_{il}}(t_j)$ を計算する。t_j の f_{il} に対する確信度 $p_{f_{il}}(t_j)$ を、$M_{f_{il}}(t_j)$ を正規化した値 $d_{f_{il}}(t_j)$ に、t_j の確信度 p_j を掛けることにより与えることにする。すなわち、

$$p_{f_{il}}(t_j) = p_j \cdot d_{f_{il}}(t_j) \qquad (3.3)$$

ここで

$$d_{f_{il}}(t_j) = \frac{M_{f_{il}}(t_j)}{\sum_{f_{ik} \in F_i} M_{f_{ik}}(t_j)} \qquad (3.4)$$

と計算する。

ステップ 3(a) iii. では、事例集合 S における f_{il} の出現確率 $p_{S,f_{il}}$ を計算する。この出現確率 $p_{S,f_{il}}$ を、f_{il} に割り当てられる訓練事例の確信度の和を S の確信度の和で割ることにより与えることにする。すなわち、

$$p_{S,f_{il}} = \frac{1}{|S|} \cdot \sum_{t_j \in S} p_{f_{il}}(t_j) \qquad (3.5)$$

ここで、記号 $|\cdot|$ は事例集合の確信度の和を表す。

と計算する。

ステップ 3(a) iv. では、S における分類クラス c_k の出現確率 p_{S,c_k} を計算する。S_{c_k} を分類クラスが c_k となる S の部分事例集合とし、この出現確率 p_{S,c_k} を、S_{c_k} に割り当てられる訓練事例の確信度の和を S の確信度の和で割ることにより与えることにする。すなわち、

$$p_{S,c_k} = \frac{1}{|S|} \cdot \sum_{t_j \in S_{c_k}} p_j \qquad (3.6)$$

と計算する。

ステップ 3(a) v. では、式 (3.5), (3.6) で計算したそれぞれの値 $p_{S,f_{il}}$, p_{S,c_k} を、前章で説明した相互情報量の計算式 (2.2) の $p_{S,v_{ij}}$, p_{S,c_k} に代入することにより相互情報量を計算する。

4 ファジイ分岐判断項目の自動生成

前章で述べたファジイ決定木生成アルゴリズムでは、数値やあいまい性を含んだ訓練事例を取り扱うためには、専門家の持つ知識を利用して、ファジイ分岐判断項目を入力として与える必要があった。しかしながら、現実世界において、適切なファジイ分岐判断項目を与えることができるとは限らないので、ファジイ分岐判断項目を自動生成するメカニズムが必要になる。

以下では、ファジイ分岐判断項目を自動生成する方法を提案する。このアルゴリズムをファジイ決定木生成アルゴリズムに組み込むことにより、訓練事例を与えるだけで、ファジイ決定木を生成することができる。

まず分類クラスの判別という観点で有効なファジイ分岐判断項目を生成する問題を例を通して考える。例えば、入力データとして図 1の実線に示すようなメンバーシップ関数 $m_{v_1} \sim m_{v_6}$ を持つファジイ集合が属性値として与えられたとする。ここで、属性値が v_1, v_2, v_6 ならば分類クラスは c_1 となり、属性値が v_3, v_4, v_5 ならば分類クラスは c_2 となるとする。

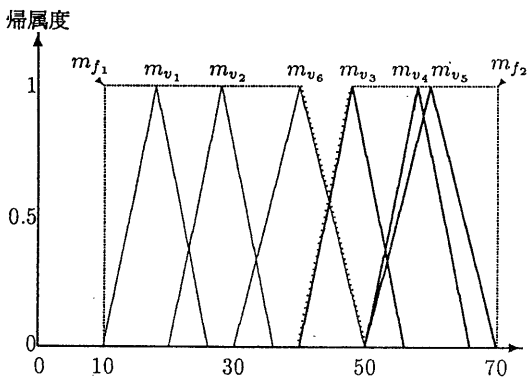

図 1: ファジイ分岐判断項目と属性値

この場合、分類クラス c_1 に対応するファジイ分岐判断項目 f_1 と、分類クラス c_2 に対応するファジイ分岐判断項目 f_2 を生成すれば良い。すなわち、図 1の破線で示すようなファジイ分岐判断項目のメンバーシップ関数 m_{f_1} と m_{f_2} を生成する。一般に、分類クラスが同一のサンプルごとに、属性値の代表値の平均を求めた場合、その平均値の周りに分類クラスが等しい属性値が集まっていると考えられる。また、分散の小さいものほど平均値の周りに属性値が分布し、大きなものほど平均値から離れて属性値が分布しているので、分散の小さい場合は作成したファジイ分岐判断項目のメンバーシップ関数の広がりを小さくし、分散の大きな場合はメンバーシップ関数の広がりを大きくすれば、分類クラスの判別を行なう観点でより有効なファジイ分岐判断項目が生成できると考えられる。

この考えを基に、訓練事例からファジイ分岐判断項目を生成するアルゴリズムを次のように構成する。ただし、このアルゴリズムでは、生成するファジイ分岐判断項目のメンバーシップ関数を、式 (4.1) 形式のメンバーシップ関数 $m_{f_{il}}(x)$ とする。

$$m_{f_{il}}(x) = (a,b,c,d) = \begin{cases} 0 & x \le a \\ \frac{1}{b-a} \cdot (x-a) & a < x < b \\ 1 & b \le x \le c \\ \frac{1}{c-d} \cdot (x-d) & c < x < d \\ 0 & d \le x \end{cases} \quad (4.1)$$

ここで、$a \le b \le c \le d$

また、式 (4.1) が表すメンバーシップ関数を図 2に示す。

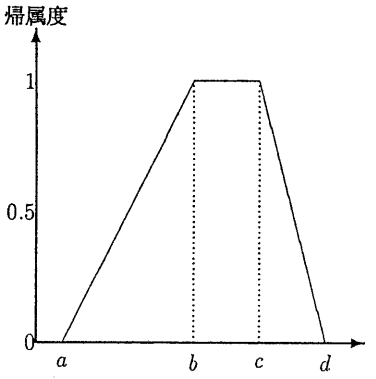

図 2: メンバーシップ関数

以下のアルゴリズムは、事例集合 S の属性 A_i の項に対するファジイ分岐判断項目を生成する。

1. 事例集合 S に含まれる各々の訓練事例 t_j に対して、 属性 A_i の属性値 v_{ij} のメンバーシップ関数の代表値 r_{ij} を計算する。

2. 分類クラス c_l が同一のサンプルごとに、 確信度を重みとした属性値の代表値 r_{ij} の平均値 E_{c_l} を計算する。

3. c_l を E_{c_l} の小さい順に並べる。

4. c_l ごとに、 属性 A_i のファジイ分岐判断項目 f_{il} を生成する。

5. f_{il} に基づいて事例集合 S を分割し、 各々を新たな部分事例集合 $S_{f_{il}}$ とする。

6. 確信度の和が設定するしきい値 T_1 より大きな $S_{f_{il}}$ に対して、 c_l に対する分離度 $u(S_{f_{il}})$ が最小となる $S_{f_{\alpha l}}$ を選択する。

7. 選択する $S_{f_{\alpha l}}$ がないか、 $u(S_{f_{\alpha l}})$ が設定するしきい値 T_2 より大きければ分割を終了する。 さもなければ $S_{f_{\alpha l}}$ に対して、 ステップ 1. からの処理を繰り返す。

このアルゴリズムでは、ファジイ分岐判断項目数が多くなりすぎないように、事例集合の確信度の和が一定値 T_1 以下になるか、事例集合の分離度が一定値 T_2 以上になった時に分割を終了している。

ファジイ分岐判断項目生成アルゴリズムで行なわれている、各種計算方法について説明する。

ステップ 1. では、属性値 v_{ij} のメンバーシップ関数 $m_{v_{ij}}(x)$ の代表値 r_{ij} を、式 (4.2) により計算する。

$$r_{ij} = \frac{\int_X x \cdot m_{v_{ij}}(x)dx}{\int_X m_{v_{ij}}(x)dx} \tag{4.2}$$

ここで、 $X = \{x | m_{v_{ij}}(x) > 0\}$

この値は、メンバーシップ関数 $m_{v_{ij}}(x)$ と x 軸により囲まれた領域の重心の x 座標に相当している。

ステップ 2. では、分類クラス c_l が同一のサンプルごとに、確信度を重みとした属性値の代表値の平均値を、式 (4.3) により計算する。

$$E_{c_l} = \frac{1}{|S_{c_l}|} \cdot \sum_{v_{ij} \in S_{c_l}} r_{ij} \cdot p_j \tag{4.3}$$

また、分類クラス c_l が同一のサンプルごとに、確信度を重みとした属性値の代表値の分散を、式 (4.4) により計算する。

$$V_{c_l} = \frac{1}{|S_{c_l}|} \cdot \sum_{v_{ij} \in S_{c_l}} \{(E_{c_i} - r_{ij})^2 \cdot p_j\}^{\frac{1}{2}} \tag{4.4}$$

次に、ステップ 4. のファジイ分岐判断項目のメンバーシップ関数の生成方法について説明する。分類クラスが同一のサンプルごとに計算した分類クラス c_β, c_γ に対する属性値の代表値の平均値を E_{c_β}, E_{c_γ} とする。ただし、 $E_{c_\beta} \leq E_{c_\gamma}$ とし、どの分類クラスに対する属性値の代表値の平均値も E_{c_β}, E_{c_γ} の間にないとする。また、分類クラスが同一のサンプルごとに計算した分類クラス c_β, c_γ に対する属性値の代表値の分散を V_{c_β}, V_{c_γ} とする。このとき、生成されるファジイ分岐判断項目 $f_{i\beta}$, $f_{i\gamma}$ のメンバーシップ関数の境界を式 (4.5) により計算する。

$$\begin{aligned} c_{i\beta} &= a_{i\gamma} = E_{c_\beta} + \frac{V_{c_\beta}}{2 \cdot (V_{c_\beta} + V_{c_\gamma})} \cdot (E_{c_\gamma} - E_{c_\beta}) \\ d_{i\beta} &= b_{i\gamma} = E_{c_\gamma} - \frac{V_{c_\gamma}}{2 \cdot (V_{c_\beta} + V_{c_\gamma})} \cdot (E_{c_\gamma} - E_{c_\beta}) \end{aligned} \tag{4.5}$$

ここで、

$$\begin{aligned} m_{f_{i\beta}}(x) &= (a_{i\beta}, b_{i\beta}, c_{i\beta}, d_{i\beta}), \\ m_{f_{i\gamma}}(x) &= (a_{i\gamma}, b_{i\gamma}, c_{i\gamma}, d_{i\gamma}) \end{aligned}$$

ステップ 6. では、事例集合 S の分離度 $u(S)$ を、式 (4.6) により計算する。

$$u(S) = \max_{c_l \in C} \frac{|S_{c_l}|}{|S|} \tag{4.6}$$

この値は、事例集合 S がひとつの分類クラスに対応する割合を表している。

5 数値実験と評価

この章では、提案したファジイ決定木生成アルゴリズム (IDF アルゴリズム) の性能評価のために行なった数値実験について説明し、アルゴリズムの性能評価を行う。

5.1 実験 1

IDF アルゴリズムは、あいまい性を取り扱うために ID3 アルゴリズムを拡張したアルゴリズムであり、直接 ID3 アルゴリズムと比較することはできない。

従って、初めに、数値データに対して、ID3 アルゴリズムと IDF アルゴリズムの比較を行なう。

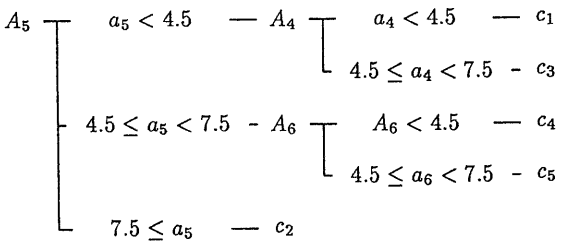

図 3: 性能評価サンプルに対する ID3 アルゴリズムによる決定木

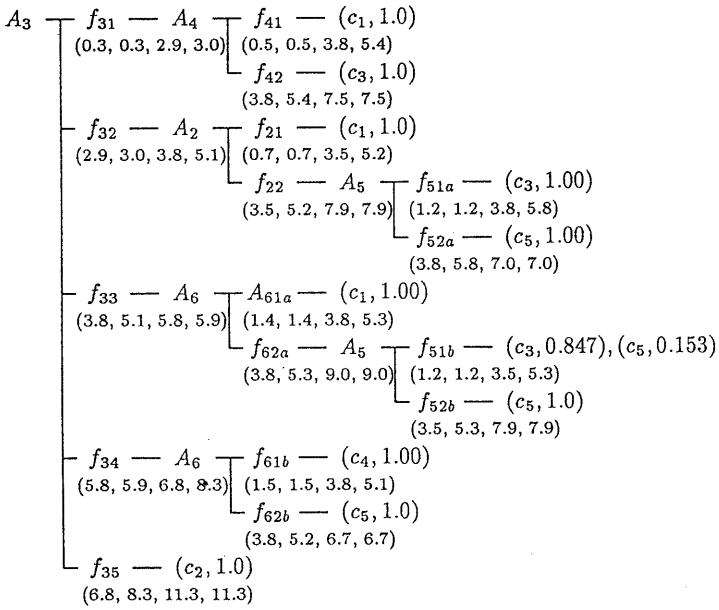

図 4: 性能評価サンプルに対する IDF アルゴリズムによるファジイ決定木

評価用のサンプルとして、表 1に示す事例を用いる。ただし、実際の個々の事例が持つ属性値は表 1に示した値を平均とし、分散 1.0 のばらつきを持った数値である。ID3 アルゴリズムでは、数値データをそのまま扱えない。従って表 1のそれぞれ値の中心を境界と設定し、数値を離散値に変換して扱う。例えば、属性 A_1 に対しては、$a_1 < 4.5$, $4.5 \leq a_1 < 7.5$, $7.5 \leq a_1$ という 3 つの区間を設定する。従って、属性 A_1 の属性値が 5.0 ならば、2 番目の区間 $4.5 \leq a_1 < 7.5$ に変換する。一方の IDF アルゴリズムでは、与えられたサンプルをそのまま使用する。

この評価サンプルは、A_1 と A_2、A_3 と A_4、A_5 と A_6 をそれぞれペアとして同一の分布を持つ属性が 3 通りあるので、本来ならば、2 つの属性を判定するだけで結論が得られる。しかしながら、データの持つノイズにより、3 つ以上の属性値を評価して分類性能の劣化を吸収する必要がある。

実験方法としては、最初に、正規乱数を用いて、各事例の出現頻度が同じになるように、全部で 150 個の訓練事例と 500 個の評価事例を発生させる。次に、150 個の訓練事例で決定木を生成し、決定木を用いて 500 個の評価事例の分類クラスを判定する。

表 1: 性能評価サンプル

A_1	A_2	A_3	A_4	A_5	A_6	class	確信度
3	3	3	3	3	3	c_1	1.0
9	6	9	6	9	6	c_2	1.0
3	6	3	6	3	6	c_3	1.0
6	3	6	3	6	3	c_4	1.0
6	6	6	6	6	6	c_5	1.0

この分類クラスと、本来与えられている事例の分類クラスとを比較し、決定木の性能評価を行なった。

ID3 と IDF により生成される決定木の一例をそれぞれ図 3と図 4に示す。図 4では、ファジイ分岐判断項目 f_{ij} のメンバーシップ関数が各ファジイ分岐判断項目の下に書かれている。ID3で生成した決定木の中で用いている区間と IDF によって生成されたファジイ集合のメンバーシップ関数を比較すると、概ね ID3 で生成された区間と同じになっていると分かる。ただし、図 3の区間 $a_3 < 4.5$ は、図 4のメンバーシップ関数 $m_{f_{31}}(x) = (0.3,\ 0.3,\ 2.9,\ 3.0)$ と $m_{f_{32}}(x) = (2.9,\ 3.0,\ 3.8,\ 5.1)$ を統合したものに対応し、図 3の区間 $4.5 \leq a_3 < 7.5$ は、図 4のメンバーシップ関数 $m_{f_{33}}(x) = (3.8,\ 5.1,\ 5.8,\ 5.9)$ と $m_{f_{34}}(x) = (5.8,\ 5.9,\ 6.8,\ 8.3)$ を統合したものに対応している。このとき、この区間も一致させようとするならば、ファジイ分岐判断項目を統合して、新たなファジイ分岐判断項目を生成する仕組みが必要である。

次に、ID3 と IDF の正解率を表 2に示す。ここで、表 2の正解率は、5 通りの学習サンプルと評価サンプルを組合せた実験で得られた正解率の平均値であり、第二候補までの正解率は、2 番目に大きい確信度を与える分類クラスまでに正解となる正解率である。また、各サンプルで学習に用いたしきい値は同じ値であり、ノードを末端ノードにするかを判定するしきい値については、ID3 と IDF で同じ値を使用した。

表 2: 実験結果

	正解率	第二候補までの正解率
ID3	85.4%	−
IDF	86.7%	96.4%

この表から分かるように、正解率は IDF アルゴリズムの方が高くなっている。従って、このような数値データに関して、IDF アルゴリズムはノイズを考慮した判断知識を生成したといえる。また、第二候補まで含んだ正解率は、ほとんど 100% に近くなっているので、確信度が妥当に付けられていたと分かる。

5.2 実験 2

ファジイ集合をデータに持つ訓練事例を考える。学習用の訓練事例の属性値として、表 1に示す値をおよそ表すファジイ集合を用いる。ここで、表 1に示す値を平均に持つ、分散 1.0 の 4 個の正規乱数を小さい順に並べた値を a, b, c, d とし、このファジイ集合のメンバーシップ関数を (a, b, c, d) となるように構成する。この訓練事例を正規乱数を用いて 150 個発生させ、ファジイ決定木を生成した。生成したファジイ決定木に対して、前の実験で比較の際に用いた 500 個の評価事例の分類クラスを判定し、本来評価事例が持つ分類クラスと比較して、性能評価を行なった。すると、生成されたファジイ決定木は、数値データに対して IDF アルゴリズムにより生成されたファジイ決定木と同形なものとなった。また、5 回の実験の正解率の平均は 86.6% となり、第二候補まで含んだ正解率は 97.7% となった。従って、この例に対して、ファジイ集合から有効な判断規則を生成することができたと分かる。

5.3 検討

5.3.1 ファジイ決定木の分類能力

ファジイ決定木で数値データを扱った場合に、境界値近辺の値に対して高い分類性能を示す傾向がある [5]。また、あいまい性を含んだデータに対しても、高い分類性能を示す傾向がある。この理由として、数値データやあいまい性を含んだデータが与えられた場合に、分岐ノードであいまいな判断しか行なわれないので、上位の分岐ノードで誤った判断がなされたとしても、下位の分岐ノードの判断でこれを回避する効果が得られたからと考えられる。

さらに、ファジイ決定木は確信度の付いた分類クラスを結果として出力するので、結果がどの程度信頼できるか数値により判断することができる。

5.3.2 ファジイ分岐判断項目の分割終了条件の影響

4 章のステップ 7. でファジイ分岐判断項目の分割終了を判定する最小事例数あるいは分離度を調整す

ることによって、生成されるファジイ分岐判断項目の数を調整することができる。この設定によっては、ファジイ決定木の上部にあるファジイ分岐判断項目の数が多くなり過ぎ、例題に含まれるノイズに影響された特殊な規則となる可能性がある。また、分割する属性が他にもある場合には、他の属性を使えばうまく分割できる可能性があるので、ファジイ決定木の上部では分離度があまり良くなくても分割を終了する必要がある。従って、現在は固定的に扱っている分割終了条件を可変的に取り扱う方法を検討する必要がある。

5.3.3 訓練事例の記述力

IDF では数値で与えられた属性値とファジイ集合で与えられた属性値を同時に取り扱うこともできる。すなわち、ある訓練事例の属性に対して、10, 8~12, およそ 7 といった属性値が与えられたとするならば、

$$\begin{aligned} m_{10}(x) &= (10,10,10,10) \\ m_{8\sim12}(x) &= (8,8,12,12) \\ m_{\text{およそ}7}(x) &= (6,7,7,8) \end{aligned}$$

といったメンバーシップ関数を定義することにより、各属性値をアルゴリズム内で取り扱うことができる。また、IDF では個々の訓練事例が持つあいまい性を初期確信度として与えて学習に反映することができる。すなわち、分類クラスが c_1 となる割合が 0.8、分類クラスが c_2 となる割合が 0.2 となる訓練事例が与えられたとするならば、確信度 0.8 で分類クラス c_1、確信度 0.2 で分類クラス c_2 とした 2 つの訓練事例を用いることにより学習を行なうことができる。このように訓練事例の属性値の表現方法の幅が広がったことが IDF の大きな利点である。

6 おわりに

あいまい性を含んだ訓練事例から、決定木の各分岐ノードであいまいな判断を行なうファジイ決定木を生成するアルゴリズムを提案し、提案したアルゴリズムの性能評価を行なった。

このアルゴリズムにより、訓練事例が数値やあいまい性を含んでいたとしても、あいまい性を加味した判断規則を生成することができるので、適用できる問題が広くなったと考えられる。特に、診断を行なうような問題で、不確実なデータを用いて推論を行なう必要がある場合に、提案したアルゴリズムが適用可能と考えられる。

今後の課題としては、現行のアルゴリズムでは、分類クラスが同一なサンプルごとに属性値の代表値の平均値を求め、その平均値が隣接するものに対し、分散の比を用いて、ファジイ分岐判断項目のメンバーシップ関数の境界を決定している。しかしながら、生成したファジイ分岐判断項目が、式 (2.2) の相互情報量のような評価値を最大にしているとは限らない。この解決策として、ジェネティックアルゴリズム [6]、アニーリングなどの最適化手法を適用した境界の調整が考えられる。しかしながら、各分岐ノードの決定ごとにファジイ分岐判断項目を生成し直さなければならないので、実行時間を考慮した最適化手法を考える必要がある。また、5.3.2 でも述べたが、ファジイ分岐判断項目の分割終了条件を可変的に取り扱う方法を考える必要がある。さらに、生成されたファジイ決定木の正当性の理論的評価として、PAC 学習理論の適用などを考えていきたい。

参考文献

[1] J.R. Quinlan, (1985), Induction of Decision Trees, *Machine Learning*, **1**, 71-99.

[2] 櫻井, 荒木, (1992), ファジイ理論を適用した知識獲得, 第 15 回知能システムシンポジウム, 169-173

[3] L.A. Zadeh, (1965), Fuzzy Sets, *Information Control*, **8**, 338-353.

[4] L.A. Zadeh, (1978), Fuzzy set as a basis for a theory of possibility, *Fuzzy Sets and Systems*, **1**, **1**, 3-28.

[5] 荒木, 櫻井, 小島, (1992), 数値データによるファジイ決定木の学習, 人工知能学会全国大会, 157-160.

[6] 野村, 荒木, 林, 若見, (1992), Genetic Algorithm によるファジイ推論ルールの最適化, システム工学部会研究会, 81-86.