人　工　知　能 99−24
（1995．3．7）

ファジイ帰納学習アルゴリズム IDF の評価

櫻井 茂明, 荒木 大

(株) 東芝 研究開発センター

数値属性を扱えるように ID3 を改良した C4.5 は、有効な決定木生成アルゴリズムである。しかしながら、C4.5 は、再帰的に 2 分割することにより、数値属性からクリスプな領域を生成するため、分類クラス数が多かったり、事例集合内にノイズがある場合に、適切な領域を生成することができない。そこで、数値属性及びファジイ属性に対して、分類クラス数に基づいたファジイ分割を行ない、*AIC* を用いて分割数を決定するアルゴリズムを提案し、このアルゴリズムを組み込んだ IDF を提案する。また、サンプルデータに対して、C4.5 との比較を行ない、正解率に関して t-検定を適用して、IDF の有効性を検証する。

The evaluation for a fuzzy inductive learning algorithm IDF

Shigeaki Sakurai, Dai Araki

Research & Development Center TOSHIBA Corporation

C4.5, which deals with numerical attributes, is an effective inductive learning algorithm. However, C4.5 can not generate appropriate ranges for a numerical attribute when there are many kind of classes or noises in a training example set. Thus the authors propose IDF, which expresses ambiguous ranges using fuzzy sets and determines the appropriate number of ranges using Akaike's information criterion. The authors also carried out numerical experiments on well-known data sets and applied the t-test to the experimental results. The authors showed that IDF is more efficient than C4.5 in terms of classification accuracy.

1 初めに

エキスパートシステムの成否は、専門家の持つ知識をどれだけシステムに反映できたかにかかっていると言っても過言ではない。しかしながら、専門家は知識を体系的に持っているとは限らないので、知識の獲得は試行錯誤を伴わざるを得ず、知識獲得ボトルネックと呼ばれる問題が生じていた。一方専門家は、対象領域に関する問題が与えられれば、その問題に対する解答を比較的容易に提示することができる。従って、この問題とその解からなるデータを多数集めたデータ集合から知識を自動的に獲得できれば、知識獲得ボトルネックをある程度解消することができる。

Quinlan によって提案された ID3[3] は、データの性質を表す属性とデータを識別する分類クラスからなる訓練事例から決定木形式の知識を生成する有効な帰納学習アルゴリズムである。この ID3 に数値属性を扱うためのアルゴリズムを組み込んだ C4.5[4] も Quinlan によって開発されている。C4.5 は数値属性に対して 2 分割を基本としたクリスプな領域を生成する。従って、属性領域を 3 分割以上にした方がよい場合には、下位のノードで再度その属性により分割する必要があり、決定木の表現が冗長になる。また、ノイズやあいまい性の影響を考慮して、あいまいな判断を下すべきところでも、択一的な判断を行なってしまう問題もあった。以上の問題点を解決するべく、決定木の概念をファジイ集合理論 [9] を用いて拡張し、境界付近であいまいな判断を行なうファジイ決定木を提案した。また、数値属性及びファジイ属性に対して、ファジイ分岐判断項目と名付けた属性領域を分割したファジイ集合を生成しながら、ファジイ決定木を成長させる IDF[5][6] を提案した。しかしながら、IDF において、ファジイ分岐判断項目数はパラメータによって制御されており、パラメータの変化がファジイ分岐判断項目数に影響していた。このため、ファジイ分岐判断項目数を制御するパラメータを決定する問題が新たに発生していた。そこで、モデルの良さを計る測度である赤池情報基準 (AIC)[1] を導入して、パラメータを調整することなしに、ファジイ分岐判断項目数を決定できるように IDF を改良した。

今回の論文では、改良した IDF の性能評価として、C4.5 との比較実験を行ない。IDF の有効性を検証する。

2 ファジイ決定木

本章では、ファジィ集合理論により表現形式を拡張した決定木 (ファジイ決定木)[5][6] とファジイ決定木を用いた推論方法について説明する。

2.1 構成

ファジィ決定木はあいまい性を含んだ判断規則を表現できるように、従来の決定木を拡張した決定木であり、分岐ノード、末端ノード、ノード同士を結ぶ枝から構成されている。分岐ノードには離散属性、数値属性あるいはファジィ属性のいずれかがラベル付けされ、事例を下位のノードに伝播させる判断を行なう。末端ノードには確信度の付いた分類クラスがラベル付けされ、到達した事例の分類クラスを判断する。また、枝には対応するファジイ分岐判断項目あるいは属性値がラベル付けされている。

すなわち、ファジイ決定木は図 1 に示す形式で与えられる。ここで、属性「室温」と「湿度」はファジィ属性で、それぞれファジィ分岐判断項目「寒い」、「暑い」と「低い」、「高い」を持ち、対応するメンバーシップ関数が与えられている。また、属性「電源」は離散属性で、属性値「ON」、「OFF」を持つ。

2.2 推論

ファジイ決定木を用いた推論は、分岐ノードにおいて、ラベル付けされている属性に対応する事例の属性値を評価して、下位のノードに事例を伝播させる。このとき、属性値が複数のファジイ分岐判断項目に跨る場合には、ファジイ分岐判断項

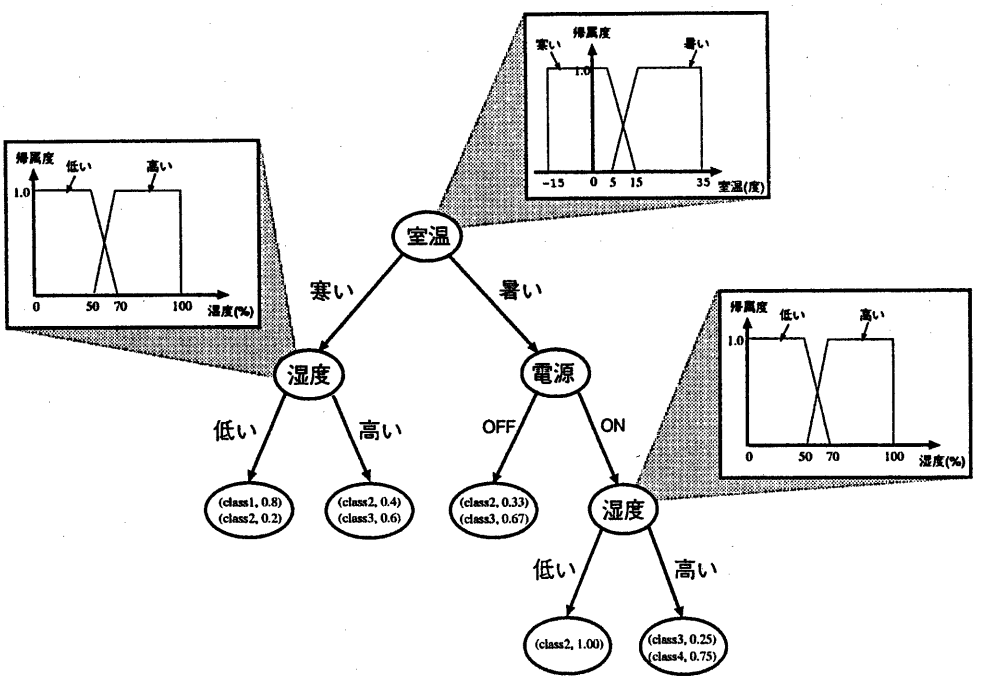

図 1: ファジイ決定木

目ごとの事例の確信度を計算して、複数の下位ノードに事例を伝播させる。このように確信度の付いた事例を分岐ノードで、逐次評価していくことにより、末端ノードまで事例を伝播させる。末端ノードでは、到達した事例が持つ確信度と末端ノードが持つ分類クラスの確信度とを掛け、事例が到達したすべての末端ノードに対して、分類クラスごとに確信度を合計して、確信度の付いた分類クラスを結果として出力する。

3 ファジイ決定木の生成

この章では、[5][6] で提案した、IDF の概略を説明するとともに、改良を加えたファジイ分岐判断項目の生成方法について説明する。

3.1 IDF アルゴリズム

IDF 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、Quinlan によって提案された ID3[3] を改良したアルゴリズムである。基本的な処理の流れは ID3 と同じであるが、ファジィ分岐判断項目の生成、属性選択基準の計算方法、事例集合の分割方法等が改良されている。

IDF は確信度の付いた事例を多数集めた事例集合から開始して、相互情報量が最も大きくなる属性を用いて、事例集合をファジイ分割することにより、ファジイ決定木形式の判断規則を成長させていく。ファジイ決定木の成長は、最小占有率、最小確信度濃度と名付けた 2 つのしきい値により制御され、しきい値が示す条件によりファジイ決

定木の枝刈りが行なわれる。ここで、最小占有率は同一の分類クラスを有する事例の占有率に関するしきい値を表し、最小確信度濃度はノードに属する事例の確信度の和に関するしきい値を表す。以上で説明したIDFを表1に示す。

1. 訓練事例をルートノードに割り当てる。
2. ノードに対して、最小占有率を越えるか最小確信度濃度を下回ったならば、ノードを末端ノードとし、確信度の付いた分類クラスをラベル付けする。
3. さもなければ、ノードを分岐ノードとする。
   (a) 非離散属性に対して、ファジイ分岐判断項目を生成する。また、離散属性に対しては、属性値をファジイ分岐判断項目として設定する。
   (b) 各属性ごとに相互情報量を計算する。
   (c) 相互情報量が最大となる属性を選択する。
   (d) 選択した属性に対応するファジイ分岐判断項目に従って、ノードに割り当てられている事例をファジイ分割する。
   (e) 各ファジイ部分集合に対して、ノードを生成し、もとのノードと生成したノードを枝で結ぶ。このとき、各枝には対応するファジイ分岐判断項目を割り当てる。
   (f) 生成した各ノードに対して、ステップ2からの処理を実行する。

表1: IDFアルゴリズム

3.2 ファジイ分岐判断項目の生成

ファジイ分岐判断項目生成アルゴリズムは、数値属性あるいはファジイ属性に対して、ノードに割り当てられている事例の属性値と分類クラスから属性領域を適切な数にファジイ分割するアルゴリズムである。このアルゴリズムは、分割フェーズと統合フェーズから構成されており、分割フェーズにおいては、分類クラスの観点で十分な分割がなされるまで、分類クラスに基づいたファジイ分割を行なう。このとき、十分な分割が行なわれたかどうかの判定には、前節で説明した最小占有率、最小確信度濃度を用いる。この分割フェーズの終了後に、過度の分割を避けるために、隣接するファジイ分岐判断項目を統合する統合フェーズを実行する。今までのIDFでは、統合フェーズにおいて、相互情報量を分割数のα乗で割った値が増加し続ける間、統合を実施するといった戦略を導入していたが、この方法では、αの値の変化が分割数の変化に大きく影響するため、適切なαの値を決定する必要があった。そこで、このようなパラメータを持ち込まずに統合フェーズを実現するために、AIC(赤池情報基準)[1]を導入し、統合を継続するかどうかの判定に利用する。

AICはモデルの良さを測る一つの基準であり、値が小さいほど良いモデルとなる。このAICにおいて、データを分割したモデルを比較する場合には、AIC[8]の値は式1で与えられる。

$$AIC = -2\sum_{i,j} n_{ij} \log_e \frac{n n_{ij}}{n_{i.} n_{.j}} + 2(C_0 - 1)(C - 1) \quad (1)$$

ここで、nはデータ数、n_{ij}はi番目のクラスを持ち、j番目の領域に含まれるデータ数、$n_{i.}$はi番目のクラスを持つデータ数、$n_{.j}$はj番目の領域に含まれるデータ数、C_0は領域の分割数、Cはクラス数を表す。

このAICを用いて統合フェーズを次のように改良する。初めに、現状のモデルにおけるAICの値を計算する。ただし、ファジイ分岐判断項目の生成においては、事例はファジイ分割されるので、

データ数の代わりに事例の確信度を用いて AIC の値を計算する。次に、隣接するファジイ分岐判断項目を統合した場合の AIC を各々計算し、最も AIC の値が小さくなるモデルを選択する。このモデルと統合前のモデルの AIC の値を比較し、統合後のモデルの AIC の値が小さい場合に、ファジイ分岐判断項目の統合を実施する。このファジイ分岐判断項目の統合を AIC の値が減少しなくなるかファジイ分岐判断項目数が2になるまで繰り返し、適切な分割数を持つファジイ分岐判断項目を生成する。

以上のように統合フェーズを改良したファジイ分岐判断項目生成アルゴリズムを表2に示す。

4 IDFの評価

この章では改良したIDFの評価として、サンプルデータを用いて数値実験を行なう。

4.1 実験方法

実験データの中から一様乱数を用いて、50個の評価事例を選択し、残りのデータを訓練事例とする。この訓練事例からIDF及びC4.5を用いて、決定木形式の判断規則を生成し、評価事例の評価を行なう。ここで、C4.5に設定するパラメータはデフォルトの値を使用し、IDFに設定する最小占有率、最小確信度濃度はそれぞれ98.0%, 1.0%と設定する。また、AIC を導入しないIDFにおける、ファジイ分岐判断項目の統合を制御するパラメータ α を0.05, 1.0, 2.0とする。この実験を各20回行ない、評価事例の正解率の比較を行なった。このとき、個々の事例集合は表3に示すように与えられ、すべての属性が数値属性からなるという特徴を持っている。

4.2 実験結果

AIC を導入したIDFと AIC を導入しないIDFを用いた場合の、Liver-disordersに関する実験の正解率の変化を図2に示す。

---

1. 非離散属性に対して、属性領域を一つのファジイ分岐判断項目とみなす。
2. 全てのファジイ分岐判断項目に対して、最小占有率を越えるか最小確信度濃度ノードを下回ったならば、ファジイ分岐判断項目の分割を終了し、ステップ4に進む。
3. さもなければ、条件を満たさないファジイ分岐判断項目を一つ選択する。
   (a) 分類クラスごとの属性値の平均値を計算する。
   (b) 平均値を中心に持つファジイ分岐判断を生成する。
   (c) 生成したファジイ分岐判断項目に対して、ステップ2からの処理を実行する。
4. 分割フェーズが生成したモデルに対して、AIC の値を計算する。
5. 隣接するファジイ分岐判断項目を統合したとして、AIC の値を計算し、AIC の値が最小になるモデルを選択する。
6. 統合前のモデルの AIC の値が 統合後のモデルの AIC の値より小さくなるか、統合前のモデルのファジイ分岐判断項目数が2になるならば、統合フェーズを終了する。
7. さもなければ、選択したモデルに対応する、隣接するファジイ分岐判断項目を統合し、AIC の値を更新する。
8. ステップ5に戻る。

表 2: ファジイ分岐判断項目生成アルゴリズム

出典	データ名	事例数	属性数	分類クラス数
The liver disorders data from BUPA Medical Research Ltd.	Liver-disorders	345	6	2
The famous iris classification data used by R.A. Fisher (1936)	Iris	150	4	3
The glass types data from USA Forensic Science Service	Glass	214	9	6

表 3: 実験データ概要

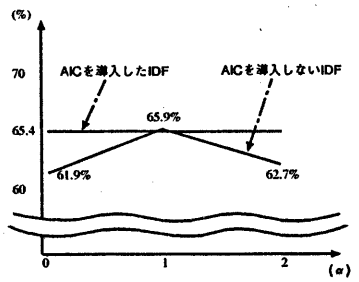

図 2: Liver-disorders における正解率の変化

また、*AIC* を導入した IDF と C4.5 を比較した場合の結果を表 4に示す。表 4において、第 1 正解率は確信度の最も高い分類クラスと評価事例の分類クラスが一致する割合の平均値を表す。また、分類クラス数が 3 以上となる Iris と Glass の実験の場合には、第 2 候補の分類クラスを考えることに意味があるので、確信度の最も高い分類クラスあるいは確信度が 2 番目に高い分類クラスと評価事例の分類クラスが一致する割合の平均値を表す第 2 正解率も実験結果として記載する。ただし、C4.5 が生成する決定木においては、クリスプな評価しか行なわれず、第 2 候補の分類クラスが何になるか分からないので、第 2 正解率は記載しない。また、この表において、標本分散は第 1 正解率に関する標本分散を表している。

4.3 実験の考察

図 2から分かるように、*AIC* を導入した IDF の正解率が、α の値を調整した正解率と同様な結果を与えている。Glass、Iris の場合についても同様な実験結果が得られており、*AIC* を導入することにより、α の値を調整しなくても、優れたファジイ決定木を生成できると分かる。

次に、*AIC* を導入した IDF と C4.5 を比較することを考える。表 4から分かるように、IDF の第 1 正解率が C4.5 の第 1 正解率より大きくなっている。しかしながら、この結果が統計的に意味のある結果かどうかは分からないので、平均値の差に関する t-検定 を実施し、統計的な有意性を検証する。すなわち、IDF が生成するファジイ決定木の第 1 正解率を μ_x、C4.5 が生成する決定木の第 1 正解率を μ_y とし、

$$\text{仮説:} \quad \mu_x = \mu_y$$
$$\text{対立仮説:} \quad \mu_x > \mu_y$$

といった仮説を立てて、第 1 正解率の差に関する t-検定 を行なってみる。ただし、t-検定 に必要な統計量 T は式 2を用いて計算する [2]。

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{n_x S_x^2 + n_y S_y^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}} \quad (2)$$

ここで、$\bar{x}$ は IDF による評価事例の第 1 正解率、$\bar{y}$ は C4.5 による評価事

Liver-disorders		第1正解率	第2正解率	標本分散 $((\frac{1}{100})^2)$	決定木サイズ
	IDF	65.4%	-	60.84	69.05
	C4.5 (枝刈り前)	62.4%	-	36.24	102.70
	C4.5 (枝刈り後)	63.3%	-	59.71	72.90
Iris		第1正解率	第2正解率	標本分散 $((\frac{1}{100})^2)$	決定木サイズ
	IDF	95.7%	98.9%	5.31	10.45
	C4.5 (枝刈り前)	93.6%	-	11.84	7.40
	C4.5 (枝刈り後)	93.7%	-	11.71	6.60
Glass		第1正解率	第2正解率	標本分散 $((\frac{1}{100})^2)$	決定木サイズ
	IDF	70.0%	86.0%	37.60	43.95
	C4.5 (枝刈り前)	65.9%	-	39.79	48.60
	C4.5 (枝刈り後)	66.0%	-	42.80	44.80

表 4: 実験結果

例の第1正解率、n_x は IDF による実験回数、n_y は C4.5 による実験回数、S_x^2 は IDF による評価事例の第1正解率に関する標本分散、S_y^2 は C4.5 による評価事例の第1正解率に関する標本分散を表す。

表5は第1正解率の差に関する統計量 T の値を示している。

	IDF2 vs. C4.5 (枝刈り前)
Liver-disorders	1.327
Iris	2.210
Glass	2.032
	IDF2 vs. C4.5 (枝刈り後)
Liver-disorders	0.873
Iris	2.113
Glass	1.945

表 5: 第1正解率の差に関する統計量

この実験において、自由度は 38 となるので、5% 水準における T の値は 1.687 と与えられる。従って、Liver-disorders に対する実験に対しては、仮説を棄却できるほど T の値は大きくならなかった。この理由としては、Liver-disorders が2種類の分類クラスからなるといった特徴によるものと考えられる。すなわち、この事例集合の場合、IDF のファジイ分岐判断項目生成アルゴリズムも、C4.5 の数値分割アルゴリズムと同様に、2分割を基本にした分割を行なっているので、どちらの手法を用いても分割の仕方にそれほど大きな違いが表れなかったためと考えられる。

一方、Iris 及び Glass に関する実験に対しては、仮説を棄却できるほど T の値が大きくなった。すなわち、これらの事例集合に対して、IDF は第1正解率の点で優れた決定木を与えると統計的にいうことができる。この理由としては、分類クラス数に基づいた分割を行なうとともに、AIC を利用して適切な数にファジイ分岐判断項目数を調整するため冗長な領域の生成を回避したためと考えられる。また、IDF の生成したファジイ決定木の場合、境界付近において、あいまいな判断を行なうので、上位ノードにおける誤った判断を下位ノードで回避する効果が表れたためと考えられる。

次に、Iris 及び Glass の場合における第2正解

率について考察してみる。Irisの場合は、第1正解率が100%に近いため、わずかしか正解率が向上していない。一方、Glassの場合には、かなりの正解率の上昇が観測されており、識別の難しい分類クラスがあり、クラス数が多いい、Glassのような場合に、確信度を利用した第2候補の選定に、大きな効果が期待できる。

最後に、決定木の形状について考察してみる。IDFで生成したファジイ決定木は、各属性を*AIC*を基準とした適切な数のファジイ分岐判断項目で、1度しか分割しないので、水平方向に広がる傾向にある。これに対して、C4.5で生成した決定木は属性を下位のノードで再分割するので、垂直方向に広がる傾向にある。決定木サイズを比較しただけでは、どちらの形状が優れているとは一概には言えないが、C4.5の生成する決定木では、同じ属性を複数回評価しなければならないので、IDFで生成するファジイ決定木に比べてルールとしての解釈がしにくい。従って、この点に関して、IDFで生成したファジイ決定木の方が優れていると言うことができる。

5 まとめと今後の課題

今回の論文では、ファジイ分岐判断項目の統合基準として、*AIC*を導入し、パラメータに依存しないファジイ分岐判断項目を生成するIDFを提案した。また、この改良したIDFとC4.5をいくつかのサンプルデータを用いて比較し、正解率に関する考察をt-検定を用いて行なった。分類クラスが2種類しかない場合には、統計的に有意となるほどには、正解率の差は観測できなかったが、分類クラスが3種類以上になる場合には、統計的に有意となるほどに正解率の差を観測でき、改良したIDFの有効性を確認した。

決定木の形状については、*AIC*を基準とすることにより、適切な数のファジイ分岐判断項目を一度に生成し、上位ノードで分割した属性を用いて下位ノードで再分割することにより生じる決定木の冗長性を排除できることを確認した。

今後の課題としては、数値基準を扱うIDF[7]においても、初めに生成する分類クラスの数を制御するパラメータの調整が必要であるので、*AIC*の導入を検討し、パラメータを調整する問題を排除していきたい。また、数値実験における属性の数は一桁と少ないため、それほど問題となっていないが、属性数に対して計算量は指数関数的に増大していくので、大規模な問題への適用に向けて、学習の高速化方法を検討していきたい。

参考文献

[1] H.Akaike, (1974), A New Look at the Statical Model Identification, *IEEE Trans. Automat. Contr.*, **19**, 716-723.

[2] P.G.Hoel, 訳者 浅井, 村上, (1978), 入門数理統計学, 培風館.

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

[4] J.R. Quinlan, (1992), C4.5:PROGRAMS FOR MACHINE LEARNING, San Mateo, CA: Morgan Kaufmann.

[5] 櫻井, 荒木, (1993), ファジィ決定木生成アルゴリズムにおける未知データの取り扱い, 情報処理学会全国大会, **93**, **70**, 31-39.

[6] 櫻井, 荒木, (1993), 帰納学習によるファジイ決定木の生成, 電気学会論文誌C分冊, **113-C**, **7**, 488-494.

[7] 櫻井, 荒木, (1993), ファジイ決定木を用いた数値予測, 知識のリフォーメーションシンポジウム, 71-80.

[8] 坂本, (1985), カテゴリカルデータのモデル分析, 共立出版.

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