ファジイ決定木生成アルゴリズムによる未知データの取り扱い

5D−6

櫻井茂明, 荒木大
(株) 東芝 研究開発センター

1　はじめに

数値やあいまい性を含んだ訓練事例から、判断規則を学習するIDFアルゴリズムを提案し、その有効性を確認してきた[1][2]。本論文では、IDFにおいて未知データを含んだ訓練事例を取り扱う方法を提案する。この方法は、未知データにファジイ集合「*unknown*」を割り当てることを特徴とする。

2　あいまいな事例からの学習

ファジイ決定木は従来の決定木を拡張したものであり、あいまい性を含んだ判断規則を表現できる。ファジイ決定木の分岐ノードはテストを行なうラベル付けされたファジイ集合(ファジイ分岐判断項目)により、あいまいな判断を行なって、下位ノードに事例を伝播させる働きを持つ。

ファジイ決定木を生成するIDFアルゴリズムは、devide & conquer 戦略で与えられた訓練事例集合を分割しながら決定木を成長させる ID3-like な手法である。しかしながら、数値やファジイ集合を属性値として取り扱うために、分岐ノードを選択するステップにおいて、ファジイ分岐判断項目を訓練事例が持つ属性値から自動的に生成する。また、このファジイ分岐判断項目により、訓練事例集合はファジイ部分集合に分割される。すなわち、各訓練事例はファジイ分岐判断項目により、確信度を更新して、下位ノードに伝播する。例えば、n 個の属性値 $v_1, \cdots, v_n$ と分類クラス c、確信度 p_0 を持つ訓練事例がファジイ分岐判断項目 f_{i1}, f_{i2}, f_{i3} をラベル付けされた分岐ノードに入力されたとする。このとき、訓練事例は確信度 p_j $(j = 1,2,3)$ を更新して図1に示すように伝播する。ここで、確信度 p_j は v_i の f_{ij} に対する帰属度を max − min 演算により求め、その値を p_0 に正規化することにより求められる。

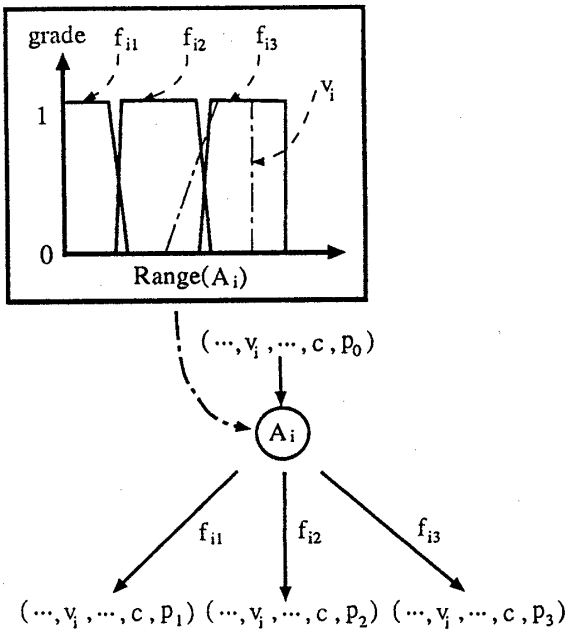

図 1: ファジイ決定木

3　未知データの取り扱い

決定木において、未知データを含む訓練事例を取り扱う手法として、1. 未知データを含む訓練事例を訓練事例集合から取り除く。2. 既知の属性値から、未知属性値を推定する。3. 未知データを一つの属性値と見なし、未知という分類を行なう。といった方法が考えられる。しかしながら、各手法にはそれぞれ、1. 未知データを含む訓練事例に与えられている情報が無視される。2. 妥当な推定値が推定されないと未知データを含む属性での分類が誤ったものとなる。3. 未知という分類がなされるノードが生成され決定木が大きくなる。といった問題点がある。そこで、これらの各手法の問題点を考慮しつつ、ファジイ決定木の性質を活用した手法を提案する。

属性値が未知ということを、すべての属性値を取る可能性があると解釈する。すなわち、式(3.1)のように定義されるメンバーシップ関数を持つファジイ集合

Dealing with unknown values in fuzzy inductive learning
Shigeaki Sakurai, Dai Araki
Research and Development Center, TOSHIBA Corp.

「$unknown$」を未知データに割り当てると解釈する。

$$m_{unknown}(x) = 1, \quad x \in Range(A) \qquad (3.1)$$

ここで、$Range(A)$ は属性 A の定義域を表す。

このように未知データを取り扱うことにより、分岐ノードにおいて判断属性の属性値が未知な訓練事例は、その分岐ノード以下の全てのノードに伝播する。すなわち、決定木全体で見れば、属性値が既知の属性だけを使用して判断がなされるという効果が得られる。その一方で、ファジイ分岐判断項目を作るプロセスでは未知データを無視するので、「$unknown$」という未知データのみを取り扱う分岐ノードは生成されない。従って、ファジイ決定木が未知データによって大きくならない。さらに、未知データをファジイ集合「$unknown$」として取り扱うので、IDF アルゴリズムにおいて、他のファジイデータと同様に取り扱うことができる。

4 数値実験と考察

実験の方法としては、乱数を用いて生成した未知データを含んだ訓練事例と、未知データを含まない評価事例に対して、「削除」、「平均値」、「$unknown$」の各手法を適用し、IDF アルゴリズムによりファジイ決定木を生成する。そして、生成されたファジイ決定木の性能評価を行なう。ここで、「削除」とは未知データを含んだ事例を取り除いて学習する手法であり、「平均値」とは未知データに既知データから求めた平均値を推定値として割り当てて学習する手法であり、「$unknown$」とは未知データにファジイ集合「$unknown$」を割り当てて学習する手法である。

実験 1 では、4 つの数値属性 A_i $(i = 1, \cdots, 4)$ からなる事例を取り扱う。また、この事例には属性 A_1, A_3 と属性 A_2, A_4 の属性値の分布が同じになる、ひとつの分類クラスに対応する属性値の分布が単峰性の分布になるといった特徴がある。

一方、実験 2 では、2 つの数値属性 A_i $(i = 1, 2)$ からなる事例を取り扱う。また、この事例にはひとつの分類クラスに対応する属性値の分布が多峰性の分布になるといった特徴がある。

それでは、実験結果を表 1、表 2 に示す。ここで、正解率とは、ファジイ決定木によって判定した最大確信度を持つ分類クラスと評価事例の分類クラスが一致した評価事例数の割合である。

実験 1 の結果より、「削除」の場合の正解率は下がるが、「$unknown$」の正解率はあまり変わらないことが分かる。これは、「削除」では未知データを含んだ事例を取り除いてしまうため、訓練事例の数が少なくなり過ぎてしまったためと考えられる。また、実験 2 の結果より、「平均値」の場合の正解率は下がるが、「$unknown$」の場合の正解率はあまり変わらないことが分かる。これは、実験 2 で取り扱っている事例ではひとつの分類クラスに対応する属性値の分布が多峰性の分布になっているため、平均値が未知データの良い推定値になっていないためと考えられる。

表 1: 実験結果 1(訓練事例数 100, 評価事例数 500)

未知データ数	削除	平均値	$unknown$
0	84.40	84.40	84.40
100	81.93	84.40	86.47
150	74.33	82.80	85.20
200	63.87	85.60	85.53

表 2: 実験結果 2(訓練事例数 200, 評価事例数 300)

未知データ数	削除	平均値	$unknown$
0	88.00	88.00	88.00
30	86.22	79.33	86.78
60	87.56	86.45	87.56
90	86.89	75.00	86.67

以上により、「$unknown$」と未知データを取り扱う手法は、事例の持つ情報を有効に活用すると分かる。

5 まとめと今後の課題

ファジイ決定木において、未知データを含む事例を取り扱う方法を提案し、その有効性を数値実験により確認した。これにより、訓練事例として取り扱える事例が多くなり、IDF の適用分野がさらに広くなった。

今後の課題として、分類クラスの値として、離散値だけでなく数値やファジイ集合を与えた場合の取り扱い方法について検討していきたい。また、一度生成したファジイ決定木に対して、新たに事例が与えられたとき、ファジイ決定木をリファインする方法を検討していきたい。

参考文献

[1] 櫻井, 荒木, 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, 1992.

[2] 櫻井, 荒木, あいまい性を含んだ訓練事例からの学習, 情処研報, 92-AI-84, 31-39, 1992.