平成13年度卒業研究		
入学年度　平成10年度	学籍番号　10117950	氏名　長谷川 友治
論文題目　属性間のクラス条件付き相関に強い決定木生成手法		犬塚研究室

1　はじめに

決定木学習は、決定木という表現形式が人間にも理解しやすく、大規模データにも適用できるアルゴリズムが多数存在することから、データマイニングの手法としても利用されている。しかし、クラスに対する属性間の相関によっては、C4.5のような一つの属性のみで分割した時に得られる重要度を用いた属性選択では、本当に重要な属性が選択できない問題がある。本論文では属性間のクラス条件付き相関に強い属性選択基準を提案する。

2　属性間のクラス条件付き相関

一般的に、全事例集合$\mathcal{D}$上で成り立つ属性間の相関は冗長な属性が含まれていることになる。それとは別に、あるクラスcに所属する事例の部分集合$\mathcal{D}_c$のみに現れる相関を属性間の**クラス条件付き相関**と呼び、その共起パターンを以下のように表現する。

$$(a_i = v_i) \overset{\mathcal{D}_c}{\Longrightarrow} (a_j = v_j)$$

v_i, v_jはそれぞれa_i, a_jの取りうる属性値のひとつである。このクラス条件付き相関は、クラスcに所属する事例の中だけで発生している相関であり、この相関がクラスcについての性質を説明していると考えられる。

しかし、複数の属性の相関によってクラスを決定付ける場合、ID3やC4.5では単一の属性の評価しかしないため見逃してしまうことがある。

3　相関に強い属性選択基準の提案

いくつかの属性の間に何らかの相関がある問題を考える。ここで、ある属性a_i以外のすべての属性の値を一定にしたまま、属性a_iの値の変化によるクラスの変化があれば、a_iがクラスに影響を与えていることが分かる。このとき、a_iとa_i以外の一定にした値との間には何らかの相関がある。このようなa_iを選ぶことで、相関を考慮した属性選択ができると考えられる。そこでa_iに関係する相関がどの程度あるかを、a_i以外の属性の値の組合せV_{a_i}が生起したときのクラスの条件付きエントロピーとして量的に計算し、重要度として用いる。

$$\underset{a_i \in \mathcal{A}}{\operatorname{argmax}} - \sum_{V_{a_i}} \frac{|\mathcal{X}_{V_{a_i}}|}{|\mathcal{X}|} \sum_{j=1}^{l} \frac{|\mathcal{X}^{c_j}_{V_{a_i}}|}{|\mathcal{X}_{V_{a_i}}|} \log_2 \frac{|\mathcal{X}^{c_j}_{V_{a_i}}|}{|\mathcal{X}_{V_{a_i}}|}$$

ここで、$\mathcal{X}_{V_{a_i}}$はV_{a_i}と同じ値（a_iの値を除く）を取る事例の集合を表し、$\mathcal{X}^c_{V_{a_i}} (\subseteq \mathcal{X}_{V_{a_i}})$は$\mathcal{X}_{V_{a_i}}$内の事例でクラスが$c$である事例の集合を意味している。

4　実装と実験結果

上の基準式を用いた実装にはハッシュ表を用いることで、漸近的な計算量がC4.5と同じになる実装を行なった。

実験として、相関性のある11-multiplexor問題とUCIのデータベースを用いてC4.5, MINITREE, 提案手法で比較実験を行なった。評価方法は、テストデータが提供されているmonk1, monk3はそれを用いて評価し、それ以外の問題は10-fold cross validationを行ない評価した。

結果を表1に、得られた木の例を図1, 2に示す。

表 1: 実験結果 Size:木のサイズ,E:誤分類率

	C4.5		MINITREE		提案手法	
DB	Size	E	Size	E	Size	E
mux11-0	204.8	0.2	**35.8**	**0.0**	46.6	**0.0**
mux11-5	246.2	15.8	78.4	14.9	**47.0**	**14.7**
monk1	**18.0**	24.3	38.0	**5.6**	28.0	8.3
monk3	**12.0**	2.8	**12.0**	2.8	17.0	**0.0**
mux6	22.6	36.2	**17.4**	2.9	21.8	**0.0**
lenses	**6.4**	**18.3**	6.9	21.7	**6.4**	**18.3**
car	**168.6**	7.4	189.1	13.5	182.0	**6.8**
nursery	503.6	**2.8**	634.8	5.2	**495.6**	**2.8**
zoo	18.2	**0.6**	**1.0**	59.8	27.3	16.9
vote	**13.5**	**4.4**	18.7	5.3	19.6	14.0
tic-tac-toe	125.5	**14.6**	113.5	24.4	**28.6**	29.6
mushroom	**33.1**	**0.0**	675.5	0.1	134.7	7.3

mux11-xは5000事例の11-multiplexor問題にx%noiseを加えたもの

C4.5と比較して、相関のあるmux11, mux6, monk1などでは決定木の性能の向上が見られた。muxの問題では、提案手法によって生成された決定木の方が非常にコンパクトになって知識としてより分かりやすい表現になっている。monk1の問題では、分類精度が非常に向上した。それ以外の問題でも、事例数が十分あるcar, nursery, lensesなどはC4.5とほぼ同等な結果が得られた。

ノイズがあるmux11-5では、MINITREEと比較して木が小さく精度も勝っており、ノイズに強いと言える。

しかし、事例空間に対して事例数が少ない問題や、zooのような各事例にIDを付けている属性および冗長性のある属性を含む問題では、$\mathcal{X}_{V_{a_i}}$の要素数が少なく重要度が十分に計算できない。そのため、zoo, vote, tic-tac-toe, muchroomではC4.5と比べて決定木の性能が悪くなった。

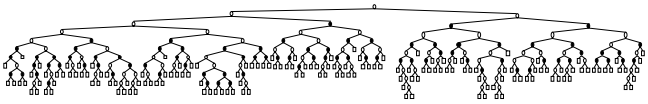

図 1: C4.5によって生成された木(mux11-5)

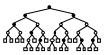

図 2: 提案手法によって生成された木(mux11-5)

5　まとめ

本論文では、決定木生成における相関に強い属性選択基準の提案を行なった。この基準を用いた実験では、属性間に相関のある問題について性能の高い決定木が生成でき、ノイズが含まれている場合にも耐性があった。また、属性間に明確な相関がない問題でも訓練事例の数が十分ある問題においては、C4.5と同等か良い結果が得られた。ただし、事例数が少ない場合や、冗長な属性、事例にIDを与えるような属性が含まれる問題については、重要度が満足に計算できない。

今後の課題は、学習を阻害する属性の削除、および欠損値、連続値の対応が挙げられる。

参考文献

[1] 長谷川, 松井, 犬塚, 世木：決定木生成における属性間の相関に強い属性選択基準, 2002年度人工知能学会全国大会（発表予定）