数値属性にも対処可能な決定木のインクリメンタルな学習

4E−3

秋山文人　黄瀬浩一　日下浩次
大阪府立大学

1. はじめに

機械学習において，決定木を用いた学習方法は，アルゴリズムが簡潔で，大量のデータを高速に扱えることから非常に有効性が高い．しかしながら，決定木を実際の問題に対して適用するためには，数値のように連続的な値をとる属性に対処可能であること，生成された決定木をインクリメンタルに修正可能であることが，同時に考慮されなければならない．

そこで本稿では，以上の点を考慮した決定木学習法を提案する．また，文書画像処理での，領域分割オペレータの決定問題について実験を行ない，本手法の有効性を示す．

2. 決定木学習

2.1. 決定木

決定木とは，ノードとアークにより表される有向グラフであり，葉以外のノードはテストする属性，アークはその属性の値，葉ノードはクラスに対応する．決定木への入力は，$<(A_1,V_{A_1}),(A_2,V_{A_2}),\cdots C_k>$の形式で表された事例の集合である．ここで，$A_i,V_{A_i},C_k$は，各々，事例を特徴付ける属性，属性$A_i$の値，事例が属するクラスである．この事例集合に対して，属性についてのテストを行ない，集合が単一クラスになるまで部分集合に分割していき，決定木を生成する．決定木学習の中心は，ノードでテストする属性をどの順序で選択するかということである．一般に決定木学習法では，属性でテストした時に得られる情報量を評価基準とし，その期待値が最大になるものを事例集合を最も良く分類する属性として選択する．こうして，最適な属性から順にテストすることで，事例を効率良く分類する木を生成する．

2.2. 従来の決定木学習手法

決定木の学習方法の中でも，ID3[1]が有力な方法として知られている．しかし，ID3には，数値属性を扱えない，木を生成した後は修正するのが容易ではないという欠点がある．前者に対しては，荒木ら[2]がそれに対処する方法を提案している．また，後者に対しては，Utgoff[3]がID5を提案している．ところが，これらの欠点に同時に対処する手法は，まだ提案されていない．そこで，本稿では，両者を統合し，数値属性を扱うことが可能で，かつ修正可能な決定木を生成することを試みる．

3. システムの概要

本システムは，逐次入力される事例集合を対象に，インクリメンタルに決定木を生成・修正する．

なお，生成される決定木のノードとアークには，そこに分類された事例集合のクラス別の数を記録するカウント，ノードには更に，区間分割の状態を記録する数直線を設け，修正時に使用する．以下，決定木の生成・修正について説明する．

3.1. 決定木の生成

最初に与えられた事例集合から決定木を生成する．その方法として，テスト属性の選択，区間分割，終了条件の判定を行なう．

1)テスト属性の選択
事例集合を最も良く分類する属性を選択し，テストを行なう．選択の方法は，ID3と同様に，テストによって得られる情報量を基準とする．

2)区間分割
決定木は，数値のような連続的な値を扱うことができない．そこで，区間分割を行なって，記号に変換して扱う．これは，荒木らの手法[2]を導入して行なう．まず，各事例の属するクラスを一つのクラスタとみなし，各クラスタにそれぞれ区間を割り当てる．この時，割り当てられた区間外に，誤って分類された事例の割合が設定値よりも大きければ，クラスタリングの手法を用いてクラスタを分割する．各事例には，属する区間の区間値をラベル付けし，記号属性に変換する．これにより，実際の事例の分布状態に基づいた有効な分類が可能である．

3)終了条件の判定
事例集合に，ノイズや少数の特殊な事例が含まれている場合，集合を完全に単一のクラスになるまで分類してしまうと，両者の影響により，決定木が過度に特殊化されてしまう．そこで，事例集合をどこまで分類するかの基準として，分類誤り率を設定する．分類誤り率とは，ノードに分類された事例のうち，同一のクラスに属するものの割合である．事例集合が，設定値を満たせば，分類を終了し葉ノードとする．これによって，一定の割合までは誤りを許容できる．

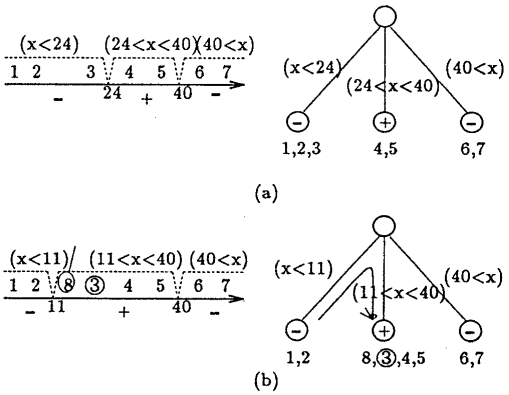

図1　分割区間の修正

Incremental Learning of Decision Tree with Symbolic and Numerical Attribute

Fumihito AKIYAMA Koichi KISE Hiroji KUSAKA
University of Osaka Prefecture

3.2. 決定木の修正

決定木を生成した後は，新たな事例集合に対して決定木を修正する．まず，新たな事例を，各ノード，アークのカウント，数直線を更新しながら分類する．次に，区間値の修正，テスト属性の変更，終了条件の判定によって木を修正する．

1)区間値の修正

区間値の修正例を図1に示す．ここで，図中の＋，ーは，事例の属するクラスを表す．(a)は事例1〜7までが与えられた状態，(b)は，さらに事例8を加えた状態である．まず，新たな事例8を加えた状態で，ノードの数直線に元の区間分割の状態を記録する．次に，新たに区間分割を行ない区間値を修正する．そのとき，事例3のように，修正後，以前と異なったクラスに分類されるものに対しては，元の状態と比較して，新しく分類されるクラスに移す．こうして元の状態と，修正後の状態の差分を修正する．

2)テスト属性の変更

新たな事例により，現在テストしている属性よりも，他の属性の方がより良く集合を分類する場合は，事例集合を最も良く分類するようにテスト属性の順序を変更する．これは，UtgoffのID5[3]の手法に基づいて行なう．まず，各ノード，アークのカウントを用いて，各属性について情報量を調べ，現在の最適属性を決定する．属性の変更が必要な場合は，カウントの値を増減することによって部分木へ分割・ノードの入れ換え・再統合を行ない，属性の順序を入れ換える．

3)終了条件の判定

木の修正によって，更に分類が必要なノードができたり，逆にそれ以上の分類は必要ないノードができたりする．そこで各ノードの分類誤り率を調べ，設定値を満たしているノードは，葉ノードに置き換える．そうでなければ，最適属性を選択し，更に事例集合を分類して木を成長させる．

これらの操作により，決定木を部分的かつインクリメンタルに修正する．

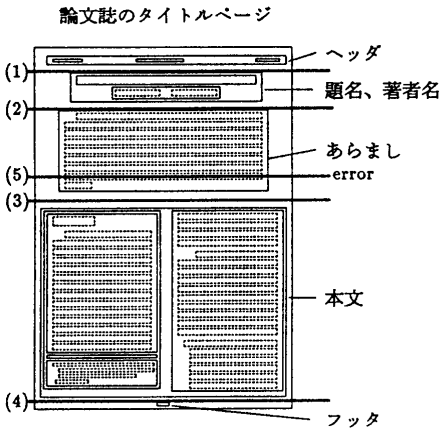

図2 領域分割オペレータ

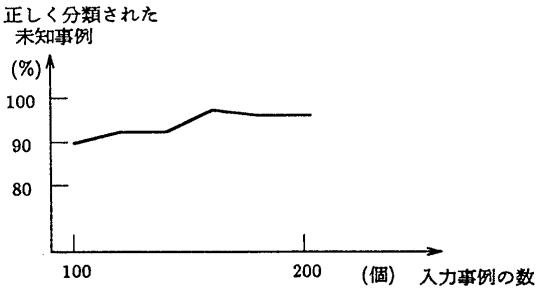

図3 実験結果

4. 実験と検討

4.1. 学習対象

文書画像の領域分割オペレータを学習対象とする．領域分割オペレータとは，図2の(1)〜(4)のように題名，本文など，画像を意味のある領域に分割するオペレータのことである．事例を分類するクラスは，(1)〜(4)の4クラス，および(5)のような errorを合わせた5クラスである．また，事例を記述する属性は，距離，横幅，面積など9種類を使用する．

4.2. 実験

実験対象は，電子情報通信学会論文誌，情報処理学会論文誌のタイトルページ各30枚から抽出した，領域分割オペレータの事例320個である．事例320個のうち，210個を既知事例，残りを未知事例とした．システムには，事例210個のうち，100個をまず与えて，決定木を生成し，その後，残り110個を22個ずつ5回に分けて与え決定木を修正した．結果を，図3に示す．図3から分かるように，決定木を修正するに従って正しく分類される事例がほぼ増加している．このことから，決定木をインクリメンタルに修正可能であることが分かる．

5. おわりに

本稿では，決定木の学習方法を提案した．本手法の特徴は，数値属性を扱えること，決定木をインクリメンタルに修正可能なことである．また，文書画像処理に対する実験から，本手法の有効性を示した．

参考文献

[1] Quinlan：“Induction of Decision Trees”，Machine Learning，pp. 57-69 (1986).

[2] 荒木他：“決定木学習における数値データの区間分割”，1991年度人工知能学会全国大会論文集，pp. 157-160 (1991).

[3] Utogoff, P. E：“ID5 : An Incrimental ID3”，Proc. of the 5th ICML, pp. 107-120 (1988).