2000年度日本オペレーションズ・リサーチ学会
秋季研究発表会

データマイニングプロセスにおける属性の生成と選択について

03300790　東京海上研究所　*高倉 記行　TAKAKURA Noriyuki
01001374　京都大学　茨木 俊秀　IBARAKI Toshihide

1. はじめに

グローバル化や情報化によってもたらされる社会環境・市場環境の急激な変化に直面し、企業は、従来以上に客観的かつ定量的な意志決定を迅速に行うことを求められている。一方、企業に蓄積されるデータ量は飛躍的に増加しており、この大量のデータを効果的に分析することで意志決定に役立てようという試みが盛んに行われるようになっている。

大量のデータから意味のある傾向、因果関係などを効率よく抽出する手法としてデータマイニングが注目を集めている。データマイニングにはさまざまなデータ分析アルゴリズムが用いられているが、その中で最もよく使われているものの一つがCART[1]、C4.5[2]などに代表される決定木導出アルゴリズムである。

これらのアルゴリズムはいずれも、ある評価基準値に基づいて1つの属性を属性集合から抽出し、この属性によってベクトル集合を複数の部分集合に分割する、というプロセスを再帰的に繰り返す。このため、

(1)1つ1つではベクトル集合を分割する効果が小さい（すなわち、評価基準値が小さい）が、連続して用いられた場合には分割する効果が大きくなるような属性の組が存在しても抽出されない
(2)ルートに近いノードでどの属性が選ばれるかによって全く異なる決定木が導出され、その結果、決定木の分類精度も大きく異なってくる

という問題点を含んでいる。

本報告では、それぞれの問題点に対し、

① 2つの属性を結合した新しい属性を生成する（新しい属性の生成）
② データの属性集合の全てを候補として決定木を導出するのではなく、より精度の高い属性集合の部分集合をあらかじめ求めておく（属性の選択）

というアプローチを採用し、これらを組み合わせた手法を提案する。

この手法をC4.5に適用して計算実験を行った結果、多くのデータにおいて、C4.5単独の場合と比べて高い分類精度を得ることが出来た。

2. 新しい属性の生成

複数の属性の論理積（AND）結合によって生成される新しい属性が、元の属性1つ1つよりもベクトル集合を分割する効果が大きければ、この新しい属性を含む決定木は分類精度が高いと期待される。

したがって問題は、論理積によって生成される新しい属性の候補が多数ある中で、どのような基準値に従って“分割する効果の大きい”属性の組み合わせを特定するかということになる。

本研究では、

$$Up\_GR(A) = GR(A) - \max\{GR(A_1),\ GR(A_2)\}$$

によって計算される値を基準値とした。ここで、Aは2つの属性A_1とA_2が論理積によって結合された新しい属性、GR()は各属性の利得比（Gain Ratio[2]）を表している。

この論理積は、例えば、A_1=0 or 1 or 2、A_2=0 or 1のときA=0(if A_1=1 and A_2=0) or 1(Otherwise)のように定義される。あるデータに対し、A_1=1 or Othersの利得比GR(A_1)が0.059、A_2=0 or Othersの利得比GR(A_2)が0.044、A=0 or 1の利得比GR(A)が0.119であったとすると、Up_GR(A)は0.060となる。

実際の計算では全ての2属性の論理積結合についてUp_GRの値を計算し、Up_GRが正の値を取り、かつUp_GRの大きい順に上位50番に入っている新しい属性のみを以後の選択の候補とした。

なお、連続する値を取る属性については、欲張り法による2値化アルゴリズムを用いて2値化することができる[3]。このアルゴリズムでは、区間の内にある値を1に、外にある値を0に変換したときにベクトルの正例と負例のペアを最も多く区別できるような区間（インターバル）を、全てのペアが区別できるまでgreedyに選んでいくようになっている。2値化されたデータに上記の論理積の操作を適用するのである。

3. 属性の選択

オリジナルデータがもつ属性に2.の方法で新たに作成された属性を加えた全体の属性集合をSとする。これを、以下の局所探索法に従って、決定木の

導出に使用する属性の集合S_1と使用しない集合S_2に分ける。

STEP1：S_1:={オリジナルデータの属性}、S_2:={新しく生成された属性}

STEP2：S_1を用いてC4.5を実行する。

STEP3：$a \in S_2$の全てについて、S_1+{a}でC4.5を実行し、決定木の分類精度が最も高いaを選ぶ。

STEP4：S_1+{a}の決定木の分類精度＞S_1の決定木の分類精度のとき、S_1:=S_1+{a}、S_2:=S_2−{a}としてSTEP3へ進む。さもなければ、そのままSTEP5へ進む。

STEP5：$b \in S_1$の全てについて、S_1−{b}でC4.5を実行し、決定木の分類精度が最も高いbを選ぶ。

STEP6：S_1−{b}の決定木の分類精度＞S_1の決定木の分類精度のとき、S_1:=S_1−{b}、S_2:=S_2+{b}としてSTEP3へ進む。さもなければ、そのままSTEP7へ進む。

STEP7：$c \in S_2$および$d \in S_1$の全ての組み合わせについて、S_1+{c}−{d}でC4.5を実行し、決定木の分類精度が最も高いc、dを選ぶ。

STEP8：S_1+{c}−{d}の決定木の分類精度＞S_1の決定木の分類精度のとき、S_1:=S_1+{c}−{d}、S_2:=S_2−{c}+{d}としてSTEP3へ進む。さもなければ、そのままSTEP9へ進む。

STEP9：最も高い分類精度を記録したS_1およびS_1によって導出された決定木を出力して終了する。

なお、以上の計算では、

(1) 決定木の分類誤り率が小さいとき、もしくは、

(2) 分類誤り率が等しく、ノード数が小さいとき、

分類精度が高いと評価した。

4. 計算実験

2.で示した新しい属性の生成、および3.で示した属性の選択の有効性を確認するため、現実のデータを用いた計算実験を行った。

具体的には、UCIのMachine Learning Repository[4]から8つのデータをピックアップし、それぞれを、ベクトル数の比が2：1：1になるように3つのファイル（training、auxiliary、test）にランダムに分割した後、以下の4通りについてC4.5の実行結果を比較した。

(1) オリジナルデータでC4.5を実行（original）

(2) オリジナルデータのうち、連続値データを2値化したデータでC4.5を実行（binarized）

(3) 新しく生成された全ての属性を(2)のデータに加えてC4.5を実行（new att）

(4) (3)の属性集合に対し属性の選択を行った結果の出力（selected）

original、binarized、new attでは、trainingデータで決定木を導出し、testデータでその決定木の分類精度を評価する。またselectedでは、各ステップにおいてtrainingデータで決定木を導出し、auxiliaryデータでその決定木の分類精度を評価する。さらに、最終的に出力された決定木の分類精度をtestデータによって評価する。

【表1：計算実験結果の比較】

データ	original	binarized	new att	selected
breast-w	4.6(9)	7.4(13)	5.1(15)	**4.6(7)**
cleve	26.3(36)	21.1(34)	22.4(21)	**19.7(16)**
crx	17.3(29)	16.2(30)	**16.2(30)**	16.2(38)
heart	23.5(35)	8.8(22)	14.7(7)	19.1(14)
liver	39.5(49)	38.4(33)	50.0(31)	**36.0(35)**
monk2	41.2(31)	—	**28.7(33)**	32.4(39)
pima	25.5(77)	32.3(61)	34.9(75)	27.1(41)
vote	5.5(7)	—	5.5(7)	**3.7(6)**

数値は、testデータにおける分類誤り率（決定木のノード数）を表す

これによると、8つのデータのうち5つのデータにおいてoriginal、binarizedの良い方よりもnew att、selectedの良い方が分類精度が高くなっており、1つについては同じであることが分かる。

属性の選択はC4.5を繰り返し実行するため多くの計算時間を必要とする。大量のデータを対象にする場合には、ベクトルをランダムにサンプリングした小規模データに対して今回の手法を適用する、等の工夫を行う必要がある。

参考文献

[1] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, P. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth International Group.

[2] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufman.

[3] Boros, E., Hammer, P. L., Ibaraki, T., & Kogan, A. (1997). Logical analysis of numerical data. Math. Programming, 79) 163-190.

[4] http://www.icu.uci.edu/~mlearn/MLRepository.html