相互情報量基準を用いた連続値クラスに対する仮説の構築

3W－4

後藤 匡史　鈴木 英之進

横浜国立大学工学部電子情報工学科

1 序論

分類問題とは，各データが有限個の値を持つクラスのうち一つに属するデータ集合から，データのクラスを予測する仮説を推定する問題である．分類アルゴリズムを連続値を持つクラスで構成されたデータに適用する場合，あらかじめクラスを離散化する必要がある．しかし，このような前処理としての離散化は，領域知識を持たない者にとって困難であり，また，分類アルゴリズムの特性を考慮しないために不適切な仮説生成の原因となってしまう．したがって，適切な仮説を容易に構築するためには，領域知識を用いずに，構築される仮説を用いてクラスを離散化する必要がある．本稿では，仮説からクラスを評価する新基準を用いて，領域知識を用いずに分類アルゴリズムに適した離散値クラスを生成する手法を提案する．

2 分類アルゴリズムの連続値クラスデータへの適用

分類アルゴリズムを連続値クラスを持つデータ集合に適用する場合，複数の属性と一つのクラスを持つデータ集合 $D'=\{E_1,E_2,\cdots,E_n\}$ は，クラスの離散化によりデータ集合 $D=\{e_1,e_2,\cdots,e_n\}$ へと変換される必要がある．このとき，例 E_i の連続的に表わされるクラス値 A_i は，有限個の値 $c_1,c_2,\cdots,c_m$ の一つを持つクラス値 $a_i\in\{c_1,c_2,\cdots,c_m\}$ となる．D は，訓練例集合 D_L とテスト例集合 D_T に分けられ，D_L を用いて例 e_i の属性値からクラス値を b_i と予測する仮説を学習し，D_T を用いて構築された仮説を評価する．

有用な仮説は，未知の例に対して，より多くの例のクラス値を正しく推定出来ることが必要である．また，学習問題は，クラスを離散化することにより単純なものとなるが，クラス値数 m を過度に少なくしたり，多くの例が特定のクラスに属してしまうと，仮説から得られる情報が少なくなってしまうため，適度の複雑さを保つ必要がある．したがって，仮説の正確さと学習問題の複雑さを同時に評価する必要があり，それぞれの基準として，テスト例のクラス値を正しく推定する割合である正答率と，テスト例のクラス値の情報量を用いるのが妥当であると考えられる．

3 相互情報量基準による仮説駆動型クラス離散化

正答率とクラスの情報量はトレードオフの関係にあり，クラスの情報量を低くすると学習問題が単純となり正答率は上昇するが，クラスの情報量を高くすると，仮説が複雑となり正答率は低下してしまう．したがって，両者の値を同時に高く保つ単一基準を設ける必要があり，本手法では，データのクラス値と仮説によるクラス値の相互情報量 (*Mutual Information: MI*) を用いる．これは，データのクラス値と仮説によるクラス値が等しくなるときの情報量であり，データの情報量が高くなり，同時に正答率が高くなるときにだけ高い値を取る．訓練例集合 D_L に対する MI は，未知の例に対しても高い正答率，情報量を保つように D_L を r 個の例集合 $D_{L1}\cdots D_{Lr}$ に分割し，D_{Lj} を除いた訓練例集合を用いて構築された仮説に対して D_{Lj} を用いて計算する r-fold-crossvaridation を用い，r 回計算した平均値をとる．

$$
\begin{aligned}
MI = \frac{1}{r}\sum_{j=1}^{r}\sum_{i\in D_{Lj}} \Big\{ & p(a_i,b_i)\log_2\frac{p(a_i,b_i)}{p(a_i)\cdot p(b_i)} \\
& + p(\bar{a_i},b_i)\log_2\frac{p(\bar{a_i},b_i)}{p(\bar{a_i})\cdot p(b_i)} \\
& + p(a_i,\bar{b_i})\log_2\frac{p(a_i,\bar{b_i})}{p(a_i)\cdot p(\bar{b_i})} \\
& + p(\bar{a_i},\bar{b_i})\log_2\frac{p(\bar{a_i},\bar{b_i})}{p(\bar{a_i})\cdot p(\bar{b_i})} \Big\}
\end{aligned}
$$

ただし，$p(x)$ は x の生起確率を表し，$p(x,y)$ は x と y の結合確率を表わす．

本稿で提案するアルゴリズムは，データ集合 D' の各クラス値 A_i を分割候補点として，それぞれの値をクラスの分割点として D_L を離散化し，MI を計算して，上位 N 通りの分割を選択する．分割点の決定は，0 個から始まり，N 通りについて選ばれた分割点を残した状態で，さらに各候補点を用いて，分割点を順次増やすことにより行う．ここで，候補点数が多い場合には，分類される未知の例が全く無い無用なクラスを生成する分割候補点を選ぶ可能性があり，また，実行時間が過度に長くなってしまう．したがって，ある分割候補点によって，テ

Concept Formation from Continuous Class Data based on Mutual Information Criteria

Masafumi Gotoh and Einoshin Suzuki
Division of Electrical and Computer Engineering, Faculty of Engineering, Yokohama National University
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan

スト例を一つも分類しないクラス値が生成されたとき，その候補点は，分割候補点集合から削除される．全ての分割候補点において MI を高くする分割が出来なくなったとき，離散化を終了し，生成された離散値を持った訓練例集合 D_L を用いて仮説を構築する．

4 データベースへの適用実験

本手法の有用性を評価するため，1994 年の JR 東日本キヨスク東京支店 243 店舗における 52 種類の商品に対する年間棚卸し数をまとめたデータベースを用いて実験を行った．243 の例は,200 の訓練例と 43 のテスト例に分け,MI の計算は,200 の訓練例に対し 5-fold-crossvaridation を用いて行い、分割の選択数 N は 5 とした。クラス属性には,52 品目のそれぞれを順に割り当て 52 回の実験を行ない，仮説の構築には，決定木構築アルゴリズムである C4.5 [1] を使用した．実験には PentiumII266MHz 搭載のパーソナルコンピューターを用い，その実行時間は一度の実験に対し最大で約 30 分程度であった．また，比較手法に，クラス数を本手法で作られた仮説のものと同数として，クラスを等間隔方式と等頻度方式[2]を用いて離散化し，同様に訓練例 200，テスト例 43 として,C4.5 を用いて仮説を構築した．

実験における正答率と情報量について，本手法の正答率の低い順にクラス属性を並べ替え，等間隔方式と比較したものを図 1，等頻度方式と比較したものを図 2 に示す． 各手法における正答率と情報量の平均

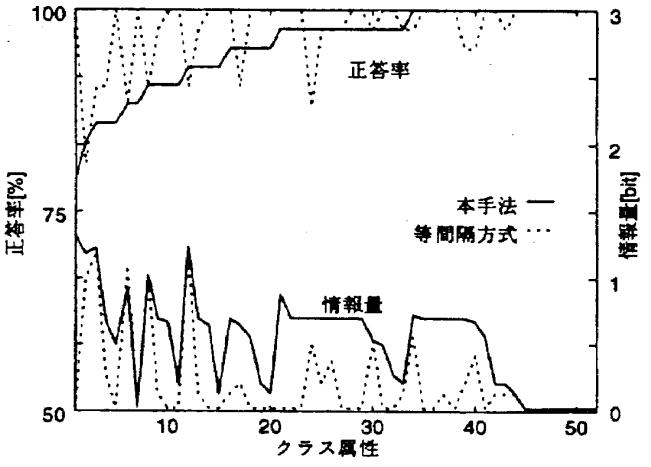

図 1: 等間隔方式との比較

値はそれぞれ，本手法が 95.75%,0.53bit，等間隔方式が 97.66%,0.20bit，等頻度方式が 94.05%,0.43bit であった．等間隔方式は、その仮説の多くが単一クラスだけで構成され，その他の仮説も多くの例が分類されたクラスだけを考慮することが多く，正答率は本手法を上回ったが，情報量が低く，過度に問題を単純化してしまう離散化方式であった．また，等頻度方式は，訓練例について最も情報量の高くなる離散化方式であるが，テスト例を考慮しないため，情報量の平均値は本手法が上回った．したがって，本実験においては，仮説の正確さと，学習問題の複雑さを共に評価する本手法が最も有用であることが示された．

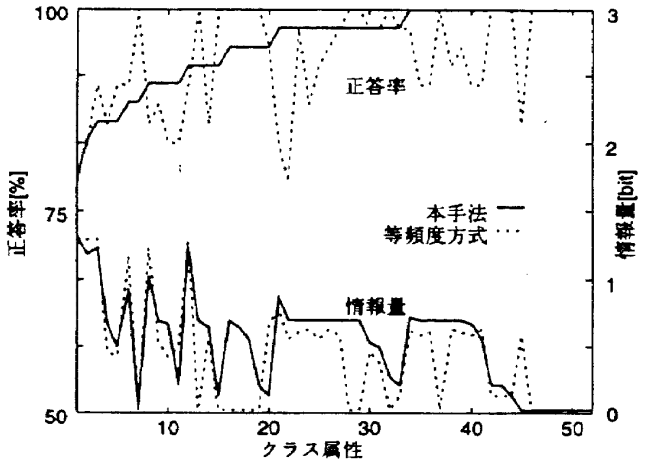

図 2: 等頻度方式との比較

なお，本手法によって生成された仮説は，クラスが国内煙草であれば，外国煙草や喫煙用具など，クラスがビールであれば，清酒やブランデーなど，クラスの属性の関連商品で構成され，妥当であると考えられる．

5 結論

本稿では，相互情報量に基づいて，仮説とクラスの両方を考慮した離散化方式を導入する事によって，連続値クラスを持つデータ集合に分類アルゴリズムを適用する手法を提案した．これは，従来の手法とは異なり，クラス数をあらかじめ決定する必要が無く，また領域知識も必要としない．この手法を，商品棚卸しデータベースに適用して，その有用性を確認し，連続値クラスに対して適切な仮説が構築されることを示した．

参考文献

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

[2] Dougherty, J., Kohavi, R., and Sahami, M.: "Supervised and Unsupervised Discretization of Continuous Features", Proc. of the 12th International Conference on Machine Learning, pp.194-202 (1995).