確率分布による概念表現を用いた事例ベースの最適圧縮*

4M－3

大杉仁隆†　上原邦昭†
神戸大学 工学部 情報知能工学科‡

1 はじめに

事例に基づく学習の分野において，事例の削減は計算コストを下げるために有益な技術である．そのため，多くの研究者によって研究されている．しかし，ほとんどの研究ではインクリメンタルに事例を削減する手法が用いられている．このようなインクリメンタルな手法では，訓練事例の入力順序が異なると削減される事例も異なってくるという問題点が分かっている．本稿では，ノンインクリメンタルかつ事例を最適に削減することが可能な OPT-CBL アルゴリズムを提案する．

2 OPT-CBL

2.1 基本的な考え方

本稿では，事例の削減問題を，訓練事例中のどの事例を事例ベースに保持するかを決定する問題として考える．従来の研究では，典型的な事例から例外的な事例までむらなく保持する手法，典型的な事例だけを保持する手法，事例の平均だけを保持する手法が提案されている．これに対して，本稿では，平均的な事例と例外的な事例を保持する手法を採用している．これは，平均的な事例により大部分の事例をカバーすることができるが，その事例だけではカバーすることができない領域が存在するため，例外的と思われる事例を保持して，より広い領域をカバーするといった考え方である．なお，平均的な事例とは，事例の平均から生成される各カテゴリーの象徴となるような唯一の事例である．

また，保持する例外的な事例は，最適化問題によって決定される．すなわち，初期の事例ベースを CB とし，事例削減後の事例ベースを CB' とする．また，$error(CB')$ を CB 内の事例を CB' によって分類をした際の誤分類数とする．事例の保持率を $s(CB') = \frac{|CB'|}{|CB|}$ とし，C を圧縮率（目標の事例保持率）とすると，最適化問題は以下のように表現できる．ただし，C はユーザが自由に設定できるものとする．

[最適化問題]
$s(CB') \leq C$ の条件のもとで $error(CB')$ が最小となるような CB' を CB から求めよ

2.2 プロトタイプの構築

平均的な事例の生成，及び例外的な事例の発見のために事例の平均であるプロトタイプ [1] を構築する．本稿では，事例は属性－値対とそのカテゴリーからなる組で表されているものとする．プロトタイプはベクトルで表現された属性の集合として定義される．各属性のベクトルの要素は，その属性の取り得る値の頻度である．つまり，各属性は取り得る値の確率分布によって表現されていることになる．カテゴリー C において，属性値 c_{jk} の頻度 $P_C(c_{jk})$ は以下のように表現できる．

$$P_C(c_{jk}) = \frac{\sum_{i=1}^{N} f(a_{ij}, c_{jk})}{N}$$
$$f(a_{ij}, c_{jk}) = \begin{cases} 1 & a_{ij} = c_{jk} \\ 0 & a_{ij} \neq c_{jk} \end{cases} \quad (1)$$

ただし，i は C に属する事例，j は属性，k は属性値の番号を示している．a_{ij} は i 番目の事例における j 番目の属性の値を示している．N は C に属する事例数である．このプロトタイプから，各属性において最も頻度の高い属性値を選択し，それらの属性値で事例を構成してカテゴリーごとの平均的な事例を生成している．

次に，プロトタイプ C と事例 I_i 間の距離を定義する．以下の定義は，$Distance$ の値が大きくなればなるほど，事例 I_i が例外的な事例であることを示して

*Optimal Storage Reduction Using Prototypes and Exceptional Cases
†Yoshitaka Oosugi and Kuniaki Uehara
‡Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University

いる．

$$Distance(C, I_i) = \sum_j (1 - P_C(a_{ij})) \qquad (2)$$

2.3 アルゴリズム

事例を保持する際に，個々の事例ごとに保持するかどうかを逐次的に決定する手法は効率がいいとは言えない．そこで，カテゴリーごとにプロトタイプと事例との距離に対して閾値を設定し，その閾値以上の事例のみを保持するようにしている．そのために，まずカテゴリーごとに各閾値で事例を保持した際の事例数を示した表を作成する（表1）．この表から最適化問題の条件である $s(CB') \leq C$ を満たすように，カテゴリーごとに閾値を変動させ，最適な事例ベースの候補を探索する．例えば，表1において，圧縮率 C を 10% とすると，訓練事例 90 個（閾値が 0 の時の事例数の合計）の中から総事例数が 9 個以下となる事例ベースの候補が探索される．その結果，事例ベースの候補として以下の 32 通りがあげられる．

総事例数	事例数の組み合わせ	
7	(1 1 1 1 1 1 1)	1 通り
8	(1 1 1 1 1 1 2)	7 通り
9	(1 1 1 1 1 2 2)	21 通り
9	(1 1 1 1 1 1 3)	3 通り

それぞれの事例ベース候補において，1-Nearest Neighbor アルゴリズムを用いて訓練事例の分類を行い，誤分類数を計算する．最終的に，最も誤分類数の少ない事例ベースが最適な事例ベースとなる．

3 実験と結果

本実験では，代表的なインクリメンタル手法である CBL3 [2] との分類精度の比較を行った．使用したデータベースは UCI machine learning repository から引用したものである．結果はすべて 30 回の試行による平均値である．また，OPT-CBL は圧縮率をユーザが自由に指定できるが，本実験では，CBL3 との比較のために，事例保持率が等しくなるように圧縮率の調節を行っている．

実験結果を表 2 に示す．太文字は検定によって差があったものである．結果より，votes を除くデータベースに関しては CBL3 と同等，もしくはそれ以上の精度を示していることが分かる．また，訓練事例をすべて保持する CBL1 [2] と比較しても，分類精度の低下を押えつつ，事例の削減が行われていることが分かる．votes の精度が悪いのは，データベース中にノイズを含む事例が含まれていることが原因だと思われる．この結果より，OPT-CBL ではノイズを含む事例が誤って保持されてしまうことがあるという問題点が明らかとなった．

表 1: 事例ベース内に保持される事例数

距離（閾値）	カテゴリー						
	1	2	3	4	5	6	7
平均的な事例	1	1	1	1	1	1	1
1.0	2	2	2	2	2	2	2
0.9	2	4	3	3	4	5	3
0.8	4	10	6	5	5	5	8
:	:	:	:	:	:	:	:
0	10	20	10	10	15	10	15

表 2: 分類精度と事例の保持率 (%)

Database	CBL1	CBL3	OPT-CBL2
breast-cancer	70.85	65.01 (20.66)	65.28 (21.50)
zoo	95.70	91.75 (22.38)	**94.29 (19.86)**
soybean	100.00	88.18 (27.88)	**99.09 (22.22)**
votes	87.90	91.02 (9.00)	**87.75 (8.27)**

4 おわりに

本稿では，ノンインクリメンタルな事例削減手法を提案した．ノイズを含まないようなデータベースに関しては，従来のインクリメンタルな手法よりも性能が良かった．今後は，ノイズを含む事例の識別を行い，予め削除するアルゴリズムの開発が必要である．また，属性の重みの考慮，事例削減の際の計算コストの軽減についても検討する予定である．

参考文献

[1] 谷澤正幸，上原邦昭，前川禎男: 典型性に基づく概念学習アルゴリズム，情報処理学会論文誌, Vol.35, No.10, pp.1988-1997 (1994).

[2] Aha, D. W.: Case-Based Learning Algorithms, *Proc. of Case-Based Reasoning Workshop*, pp.147-158 (1991).