知能と複雑系 110－4
(1998. 1. 22)

事例の選択的な記憶による極小事例ベースの獲得

大杉 仁隆† 上原 邦昭†‡

†神戸大学 工学部 情報知能工学科

‡神戸大学 都市安全研究センター

〒657 兵庫県神戸市灘区六甲台町 1-1

Email: sugi@jedi.cs.kobe-u.ac.jp uehara@kobe-u.ac.jp

あらまし 事例ベース推論では，次々と事例を事例ベースに蓄えていくため，事例を削除して記憶量と類似事例の検索にかかる計算コストを削減することが重要となってくる．本稿では，事例ベース推論を応用した事例の分類手法に着目し，分類に有効な事例のみを選択的に記憶して，極小の事例ベースを構成するアルゴリズムを提案する．本アルゴリズムは，まず典型的な特徴からなり，大部分の事例を正分類することができる仮想的な事例を生成する．さらに，仮想的な事例に加えて，カテゴリ間の境界付近の事例のみを選択的に記憶している．このため，分類に必要のない事例を選択することはなく，非常に少ない事例で分類精度が維持できるという特徴がある．さらに，例外的な事例は事例ベースに含まれないため，ノイズにも強固であるという特徴がある．

キーワード 事例ベース推論，概念学習，事例の選択，探索，バックトラック法

Acquisition of a Minimal Instance-Base by Storing Most Prototypical Instances

Yoshitaka Oosugi† Kuniaki Uehara†‡

† Department of Computer and Systems Engineering,
Faculty of Engineering, Kobe University
‡ Research Center for Urban Safety and Security, Kobe University
Rokkodai-cho 1-1, Nada-ku, Kobe-shi, Hyogo, 657 Japan
Email: sugi@jedi.cs.kobe-u.ac.jp uehara@kobe-u.ac.jp

Abstract In the field of instance-based reasoning, reducing storage cost and computational cost is an important problem. In this paper, we will introduce a new algorithm that constructs a minimal instance-base by storing the most prototypical instance for the classification task. In addition, a small number of near-boundary instances are stored into the instance-base.We empirically show that storage requirements are sharply reduced with small sacrifices in classification accuracy.

key words case-based reasoning, concept learning, selecting instances, searching, backtracking

1 はじめに

事例ベース推論 (Case-Based Reasoning: CBR) [1] は，過去に行った解の導出結果（事例）を事例ベースに記憶しておき，類似した事例を用いて新たな問題の解を導く推論手法である．基本的に，CBR は過去の事例をすべて格納しているため，事例の記憶量と検索にかかる計算コストが問題となってくる．このため，冗長な事例やノイズとなる事例を削除し，記憶すべき事例を減らすことは，重要かつ有益な技術であると考えられている．

本稿では，CBR を応用し事例を分類する手法に着目し，分類に必要な事例のみを選択的に記憶して，極小の事例ベースを獲得するアルゴリズム SABI を提案する．Rosch ら [3] のプロトタイプ理論によれば，あるカテゴリを表象しているのは，プロトタイプと呼ばれる典型的な特徴を多く持つ事例である．この考え方に基づいて，SABI ではプロトタイプとして典型的な特徴からなる仮想的な事例を新たに生成しており，この事例を平均事例と呼んでいる．この平均事例を用いれば大部分の事例は正分類できると考えられる．

一方，Riesbeck ら [2] の心理学的な観点によると，人はまず一般概念を用いて推論を行い，一般概念に当てはまらなければ過去に経験した特殊な事例を用いて推論を行う．この考え方を本研究に適用すると，「一般概念」は「平均事例」とみなすことができる．また，「一般概念に当てはまらない事例」は，分類に悪影響を与える事例やノイズとなる事例（例外的な事例）ということになる．このため，本稿では新たに境界事例という概念を導入し，これを「一般概念に当てはまらない事例」とみなしている．境界事例とは，カテゴリ間の境界の近くに存在する事例であり，典型的な事例と例外的な事例の間に属するような事例である．このような事例は，平均事例で正分類できない，かつ分類に悪影響を与えない事例であるため，非常に重要な事例と考えられる．最終的に，平均事例と境界事例を記憶した事例ベースが極小の事例ベースとなる．

一般に，分類手法における事例の選択は，一種の概念学習として扱われており，機械学習の分野で多く研究されている．例えば，Nearest Neighbor 法 [4] は分類手法の一つであり，CBR と同様に，与えられる事例をすべて記憶するため，事前に事例の選択を行う必要がある．これらの研究では，カテゴリの種となる事例（例えば，最初に与えられた訓練事例や最も典型的な事例）を選択した後に，この事例に類似している順に分類に必要な事例を選択する．このため，後で選択した事例がすでに選択している事例をカバーしていることも起こりうる．つまり，極小の事例ベースが得られるとは限らないという問題があった．

以上の問題を解決するために，本稿では平均事例に加えて，カテゴリ間の境界に近い順に分類に必要な事例のみを選択するようにしている．このため，後で選択する事例がすでに選択している事例をカバーすることはないという特徴がある．また，例外的な事例が含まれる場合，これらの事例を選択しないようにしているため，ノイズの影響を受けにくいという特徴がある．

2 SABI アルゴリズム

2.1 平均事例

まず，本稿で用いる事例の記述形式を定義する．事例は，属性とその値からなる特徴とその事例が属するカテゴリからなる．たとえば，n 個の特徴を持つ事例 I は，以下のように表わすものとする．

$$I = (c, a_1, a_2, \ldots, a_n) \tag{1}$$

ただし，c はカテゴリ，a_i は i 番目の属性の値を表している．

上原ら [5] の典型性に基づく概念学習アルゴリズム (Prototype-Based Learning Algorithm) では，特徴の出現頻度を典型性の度合として，事例の分類を行っている．つまり，あるカテゴリにおいて，ある属性が取り得る値のなかで，出現頻度が最も高い値とその属性の組が，そのカテゴリの典型的な特徴だという考え方である．本稿でも同様に，カテゴリ c に属する事例の i 番目の属性が取り得る値のうち，出現頻度が最大の値を c の平均事例の i 番目の属性値としている．また，属性値が数値の場合は，すべての属性値と類似するように平均値を採用している．

2.2 境界事例

一般に，ほとんどの問題領域では真の境界が未知であり，境界を発見することは非常に困難である．本稿では，カテゴリと事例間の相対距離 Dis を定義して，統計的な手法によりカテゴリ間の境界を推定し，境界事例の選択を行うようにしている．

2.2.1 カテゴリと事例間の相対距離

事例と属するカテゴリの類似性の度合を表したものに典型度がある．Zhang [6] の典型度は，同じカテゴリに属する事例との類似度の平均 (intra-concept similarity) を分子，別のカテゴリに属する事例との類似度の平均 (inter-concept similarity) を分母とした比で定義されている．このため，intra-concept similarity が大きくても，inter-concept similarity が大きければ典型度は小さくなる．逆に，intra-concept similarity が小さくても，inter-concept similarity が非常に小さければ典型度は大きくなる．このようにすれば，同じカテゴリに属する多くの事例と類似しているからといって典型度が高くなったり，ほとんど類似していないからといって典型度が低くなることはなく，別のカテゴリとの相関を考慮した典型度が計算できる．つまり，この典型度は同じカテゴリの事例のみを考慮した尺度（intra-concept similarity と inter-concept similarity）を用いて計算される相対的な尺度だと言える．

この考え方を基に，SABI では同じカテゴリの事例のみを考慮して計算されるカテゴリとの距離を用いて相対距離を定義している．カテゴリ c と事例 I の相対距離では，intra-concept similarity と inter-concept similarity に，それぞれ c と I の距離と c 以外のカテゴリ $\overline{c}$ と I の距離を対応させている．なお，カテゴリ c と事例 I の距離を以下のように定義している．

$$distance(c, I) = \sum_{i=1}^{n} impossibility(c, a_i)$$

$$impossibility(c, a_i) = \begin{cases} 1 - P_c(a_i) & \text{非数値属性} \\ \sigma_{ci}^2 + (a_i - \mu_{ci})^2 & \text{数値属性} \end{cases} \quad (2)$$

ここで，n は属性数を表している．$impossibility(c, a_i)$ は，属性値 a_i がカテゴリ c 自身の属性値である可能性を表す値であり，値が小さいほどその可能性が高いことを表している．$P_c(a_i)$ はカテゴリ c における属性値 a_i の出現頻度を表し，μ_{ci}，σ_{ci} はカテゴリ c における i 番目の属性値の平均と標準偏差を表している．なお，数値属性の値は最小値が 0，最大値が 1 となるように正規化し，$distance$ が 0 と 1 の間の値をとるようにしている．

カテゴリ c と事例 I 間の相対距離 Dis は，$distance$ を用いて以下のように表される．

$$Dis(c, I) = \frac{distance(c, I)}{\min_{\overline{c}}[distance(\overline{c}, I)]} \quad (3)$$

ここで，分母は c を除くすべてのカテゴリの中で最小となる $distance$ としている [1]．

式 (2) の距離の定義も踏まえると，対象のカテゴリに属する事例の多くが持つ特徴（平均値に近い特徴）を多く持っていても，別のカテゴリに属する事例が多く持つ特徴（平均値から遠い特徴）を多く持っていれば相対距離は大きくなる．逆に，対象のカテゴリに属する事例の多く持つ特徴（平均値に近い特徴）をあまり持っていなくても，別のカテゴリに属する事例が多く持つ特徴（平均値から遠い特徴）を全く持っていなければ相対距離は小さくなる．

2.2.2 カテゴリ間の境界の推定

前節で定義した相対距離 Dis を用いれば，m（事例の持つ特徴数）次元の事例空間で表現されている事例を 1 次元の Dis 軸上で表現できるようになる．さらに，二つのカテゴリに属する事例を Dis 軸上に表現すれば，この軸上でカテゴリを区別する点（値）を統計的手法により推定することができる．このカテゴリ間の境界となる値を境界値と呼ぶ．

まず，カテゴリが 2 種類の場合の境界の推定について説明する．図 1 (a) は，カテゴリが 2 種類，属性が 2 種類の事例空間を表している．○ はカテゴリ c_i に，● はカテゴリ c_j に属している事例を表しているものとする．これらすべての事例に対して，どちらか一方のカテゴリとの Dis を計算すると，図 1 (b) のように 1 次元の Dis 軸上に事例を表現できる．この例では，c_j と事例間の Dis を計算して軸上に表現している．このため，Dis が小さいほど事例は c_j に近く，逆に大きくなるほど c_i に近いことを示している．

カテゴリを区別する境界値は，1 次元の Dis 軸上のある一点で表される．本稿では，境界値を推定するために，c_i に属する事例の Dis が正規分布 $N(\overline{X}_{c_i}, s_{c_i}^2)$，$c_j$ に属する事例の Dis が $N(\overline{X}_{c_j}, s_{c_j}^2)$ に従っていると仮定している．ここで，$\overline{X}_{c_i}$，$\overline{X}_{c_j}$ はそれぞれのカテゴリに属する事例間の Dis の平均を表している．また，s_{c_i}，s_{c_j} はそれぞれのカテゴリに属する事例の Dis の

[1] Zhang の典型度では，カテゴリが 3 種類以上ある問題領域を対象としていない．

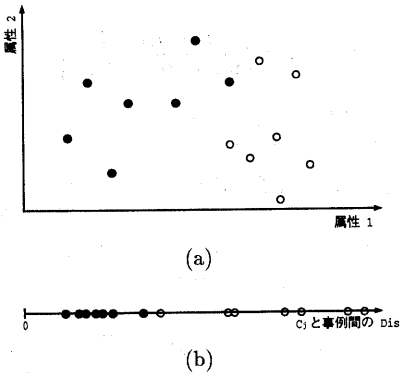

図 1: 2 次元空間から 1 次元空間への変換

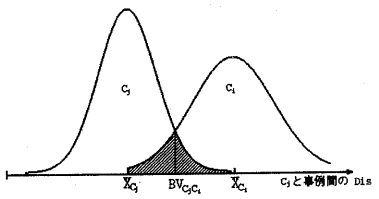

図 2: c_i と c_j 間の境界値

標準偏差を表している．このため，密度関数 は図 2 のように表すことができる．

境界値は，事例が別のカテゴリに属している確率（図 2 の斜線部分の面積）を最小にする Dis の値（図 2 の $BV_{c_jc_i}$）としている．この境界値は，二つの密度関数が交わる Dis の値であり，次式によって求められる．

$$BV_{c_jc_i} = \frac{(\overline{X}_{c_i} \times s_{c_j}) + (s_{c_i} \times \overline{X}_{c_j})}{s_{c_i} + s_{c_j}} \quad (4)$$

カテゴリが n 種類の場合には，カテゴリが 2 種類の場合の境界の推定方法を拡張し，隣接するカテゴリ間の境界値をすべて求めて，互いを区別できるようにしている．例えば，訓練事例中に n 種類のカテゴリ $c_1, c_2, \ldots, c_n$ があるとする．また，カテゴリ c_i に属しながら，距離 $distance$ が最小である別のカテゴリが c_j である事例の集合を $S_{c_ic_j}$ とする．この時，任意の 2 つのカテゴリ c_i と c_j の境界値は，$S_{c_ic_j}$ と $S_{c_jc_i}$ に含まれる事例の分布を 1 次元で表現して求めている．このため，すべてのカテゴリを互いに区別するために，全部で ${}_nC_2$ 個の境界値を求めている．

2.2.3 境界事例の選択と最適な事例ベース

境界事例は，カテゴリ間の真の境界付近に存在する事例であるが，前節までの定義を踏まえると，Dis 軸上で推定できるカテゴリ間の境界（境界値）に近い Dis を持つ事例ということになる．このため，SABI では境界値と Dis の差が，ある値（以後，選択値と記す）に存在する事例を境界事例として選択するようにしている．

SABI では，訓練事例に対して境界事例の占める割合がそれほど大きくないことから，事例の選択率の上限を指定し，この選択率が上限を満たすように選択範囲を設定するようにする．ここで，選択率が $N\%$ とは，境界事例数が訓練事例数の $N\%$ であることを意味している．そして，選択される境界事例と平均事例を記憶して事例ベースを獲得するようにする．

ここで，以下で用いる記号の定義をしておく．カテゴリ数を n，訓練事例の集合を TI，そして平均事例の集合を AI で表すものとする．また，選択率の上限を UL，選択値を ϵ で表すものとする．さらに，境界値と Dis の差が ϵ 以下である境界事例の集合を $BI(\epsilon)$ で表し，Nearest Neighbor 法によって訓練事例の分類を行った際の事例ベースの候補 Cd の誤分類率を $error(Cd)$ で表すものとする．最後に，境界値と事例の Dis の差の最大値を d_{max} とする．

SABI の目的は，$\sum_{i=1}^{n} |BI(\epsilon_{c_i})|/|TI| \leq UL$ であり，かつ $error(AI \cup \prod_{i=1}^{n} BI(\epsilon_{c_i}))$ が最小となるように，カテゴリごとに $0 < \epsilon_{c_i} < d_{max_{c_i}}$ を満たす選択値 ϵ_{c_i} を設定することである．本稿では，カテゴリごとに ϵ を単純に比較的小さい一定の間隔[2]で変動させて候補を獲得するようにしている．これは，ϵ ををなるべく小刻みに変動させ，厳密に事例を選択して候補を獲得しようというヒューリスティックな考え方である．このように獲得される候補に対して，順次 $error$ を求めており，次章以降の実験で精度の良い結果が得られている．しかし，本来ならば厳密に事例を選択することだけでなく，計算コストも考慮し効率的に候補を獲得していくことが望ましいため，ϵ を変動させる方法については，さらに検討が必要だと思われる．

[2]カテゴリごとに属するカテゴリとの相対距離の順に事例を並べた際の，事例間の Dis の差の平均値としている

また，Dis 軸上である事例が境界値を隔てて別のカテゴリ側に存在する場合がある．例えば，図 2 で c_j (c_i) に属する事例の中に Dis が境界値よりも大きい (小さい) 事例が含まれているような場合である．この時，別のカテゴリ側に存在する事例は，ノイズとなるような例外的な事例とみなすことができる．例外的な事例は誤った境界を形成して，将来，誤分類を増加させる原因となる．また，例外的な事例を分類すると誤分類してしまうことが多い．このため，SABI では，別のカテゴリ側に存在する事例は境界事例や候補の評価に用いる訓練事例から除くようにしている．最後に，本節までに述べた SABI アルゴリズムを図 3 に示す．

procedure SABI
input: TI: 訓練事例集合, UL: 選択率の上限;
output: IB: 事例ベース;
begin
  IB の誤分類率 $error(IB) \leftarrow \infty$
  平均事例集合 $AI \leftarrow \emptyset$
  for カテゴリ c とカテゴリ $\bar{c}$ の組
    c に属している $\bar{c}$ に近い事例の集合 $S_{c\bar{c}} \leftarrow \emptyset$
    $\bar{c}$ に属している c に近い事例の集合 $S_{\bar{c}c} \leftarrow \emptyset$
  TI をカテゴリごとの集合に分割する
  for 各カテゴリ c
    平均事例 I_{ave} を生成する
    $AI \leftarrow AI \cup I_{ave}$
    for 事例 $x \in$ カテゴリ c に属する事例
      距離 $distance$ が最小である別のカテゴリ $\bar{c}$ を求める
      $S_{c\bar{c}} \leftarrow S_{c\bar{c}} \cup x$
  for カテゴリ c とカテゴリ $\bar{c}$ の組
    for 事例 $x \in S_{c\bar{c}} \cup S_{\bar{c}c}$
      カテゴリ c との相対距離 Dis を求める
    境界値 $BV_{c\bar{c}}$ を推定する
  for カテゴリごとの選択値 ϵ_c の組合せ (n : カテゴリ数)
    if $\frac{\sum_{i=1}^{n} |BI(\epsilon_{c_i})|}{|TI|} \leq UL$
      事例ベースの候補 $Cd \leftarrow AI \cup \prod_{i=1}^{n} BI(\epsilon_{c_i})$
      $error(Cd)$ を計算する
      if($error(IB) > error(Cd)$)
        $IB \leftarrow Cd$
end

図 3: SABI アルゴリズム

3 実験による SABI の評価

本章では，さまざまな領域のデータベースに対して SABI の有効性を検証する．実験では，適切に事例の選択ができるかどうかを確かめるために，Nearest Neighbor 法との分類精度の比較を行っている．適用したデータベースは，UCI Machine Learning Repository [7] から得たものである．すべての結果は 30 回の試行の平均をとっている．1 回の試行では，データベース中のすべての事例からランダムに 5 分の 4 を訓練事例とし，残りの 5 分の 1 をテスト事例としている．また，SABI の選択率の上限は 30 % に設定している．

実験結果を表 1 に示す．表中の NN 法は，Nearest Neighbor 法を表している．また，カッコ内の数値は，得られた事例ベースの選択率を表している．太文字は，水準 1 % で平均の差の検定を行い，SABI の分類精度の方が Nearest Neighbor 法よりも明らかに良いと検定されたものを示している．斜体は，Nearest Neighbor 法の方が SABI の分類精度よりも良いと検定されたものを示している．

表 1: 分類精度と事例の選択率 (%)

データベース名	NN 法	SABI (UL=30 %)
breast-cancer	70.85(100.00)	*67.53*(28.70)
soybean	100.00(100.00)	*99.39*(27.78)
voting	87.90(100.00)	**91.55**(29.79)
hayes-roth	69.18(100.00)	**76.06**(1.85)
promoters	80.33(100.00)	77.95(7.35)
tic-tac-toe	97.76(100.00)	*88.94*(29.66)
cleveland	77.45(100.00)	**81.44**(28.76)
credit	80.19(100.00)	79.30(14.61)
iris	94.89(100.00)	94.34(30.00)
pima-diabetes	69.98(100.00)	71.26(29.06)

表 1 の結果から，SABI は 30 % 以下に事例を減らしても，ほとんどのデータベースにおいて分類精度の低下を 3 % 程度までに抑えていることが分かる．voting, hayes-roth, cleveland では分類精度の向上さえも見られる．これは，データベース中にノイズが含まれているため，Nearest Neighbor 法では残されてしまうノイズを，SABI では除去できていることを示している．また，hayes-roth と promoters では，選択率の上限を 30 % に指定したにもかかわらず，あまり事例を選択していないことが分かる．これは，カテゴリ間の境界が比較的単純であり，平均事例のみで十分に近似できるためと考えられる．しかし，tic-tac-toe では分類精度が 10 % 程度低下している．これは，tic-tac-toe の問題領域が特徴の選言的標準形 (Disjunctive Normal Form: DNF) [3] を用いて表現できる概念であることが原因と考えられる [8]．つまり，特徴の出現頻度ではなく，特徴間の関係によってカテゴリが決定されているのである．このため，このような問題領域に特徴の頻度情報

[3] $(x_1 \wedge x_2) \vee (x_3 \wedge x_4) \vee (x_5 \wedge x_6)$ のように，複数の and 記述を or 結合させた記述形式．

を用いて事例の生成，選択を行っている SABI を適用すると，特徴間の関係が反映されずに分類精度が悪くなってしまっていると考えられる．

次に，SABI と同様に，分類に有効な事例を境界事例として選択している Aha [9] の IB3 と性能を比較する．IB3 では，誤分類する事例を境界事例とみなして選択している．さらに，選択した事例に対して，その事例を使った分類の正誤を記録しておき，頻繁に誤分類に使われる事例はノイズを含む事例として削減するようにしている．

SABI と IB3 の選択率が同じである場合，選択対象の事例が同じであるため分類精度に差がでないと考えられる．このため，SABI と IB3 の分類精度がほぼ同じ状態で SABI の選択率を限界まで下げる実験を行い，双方の選択率の比較を行った．適用したデータベースは 3 節で適用したものと同じである．表 2 に結果を示す．表中のカッコ内の数値は，得られた事例ベースの選択率を表している．また，太文字は検定によって SABI の方が IB3 より分類精度または選択率が良いと検定されたものを示している．

表 2: 分類精度と事例の選択率 (%)

データベース名	SABI	IB3
breast-cancer	64.02(20.01)	65.01(20.66)
soybean	**99.39(0.00)**	88.18(27.88)
voting	91.67(**1.23**)	91.02(9.00)
hayes-roth	**76.79(0.00)**	55.32(9.65)
promoters	73.57(**2.00**)	74.29(20.17)
tic-tac-toe	81.42(16.21)	80.20(17.98)
cleveland	78.04(**11.85**)	77.87(13.19)
credit	78.91(**0.47**)	80.63(10.83)
iris	93.56(**4.94**)	93.90(11.51)
pima-diabetes	69.05(**15.69**)	68.18(16.32)

表 2 から，すべてのデータベースにおいて SABI の選択率が IB3 の選択率より下まわっていることが分かる．また，soybean と hayes-roth では，選択率が小さいにもかかわらず，IB3 よりも良い分類精度が得られている．このため，選択率の小さい SABI が IB3 に比べて適切に事例を選択していると言える．

しかし，SABI はノンインクリメンタルなアルゴリズムであるため，インクリメンタルな IB3 に比べると，事例ベース獲得までにかかる計算コストが非常に大きくなるという問題がある．実際には，いくつかのデータベースに対する 30 回の試行が終了するまでに，数日かかっているが，IB3 では，すべて数十秒で終了している [4]．

オーダによる計算コストの比較を行うと，まず IB3 では事例を一つ入力して事例ベースを更新するのに $O(|IB_i| \times |A|)$ の計算コストがかかるため，全体では，$\sum_{i=1}^{|TI|} O(|IB_i| \times |A|)$ となる．なお，$|TI|$ は訓練事例数，$|IB_i|$ は i 番目の事例が入力された際にすでに事例ベースに記憶している事例数，そして $|A|$ は事例の属性数を表している．これに対して，SABI は一つの候補の誤分類率を計算するのに $O(|TI| \times |Cd_j| \times |A|)$ の計算コストがかかる．このため，全体では $\sum_{j=1}^{|Candidate|} O(|TI| \times |Cd_j| \times |A|)$ となる．なお，$|Candidate|$ は 候補の数であり，$|Cd_j|$ は j 番目の候補の事例数を表している．さらに，$|IB_i|$ と $|Cd_j|$ は $|TI|$ に比べると十分小さく取り得る値の範囲が比較的狭いため，定数とみなすと，全体の計算コストはそれぞれ $O(|TI| \times |A|)$，$O(|Candidate| \times |TI| \times |A|)$ と表せる．ここで，$|Candidate|$ はカテゴリごとに考慮する選択値の数の積で表されるため，考慮する選択値の数やカテゴリ数が大きくなると非常に大きな値となる．特に，カテゴリ数に対して指数関数的に増加する値である．このため，SABI の方が IB3 よりも計算コストが非常に大きくなっていることが分かる．

4 計算コストの削減

前章の実験では，SABI による事例の選択の有効性を示すことができた反面，事例ベース獲得までにかかる計算コストが大きくなり過ぎることが明らかとなった．この原因の一つに，事例ベースの候補を得る際に選択値を単純に比較的小さい一定の間隔で変動させていることが考えられるが，本章では事例の選択の厳密性を重視したこの条件の下で計算コストを削減する手法を提案する．

SABI の事例ベースの獲得は，一種の探索問題であるため，効率的な探索手法を取り入れれば，結果として，事例ベース獲得までにかかる計算コストを削減できることになる．ここで，SABI の探索空間を例によって示す．カテゴリが 2 種類 $(c, \bar{c})$ あり，共に選択値を 3 個考慮するようにする．この問題領域における探索空間は，図 4 の実線部分で表される．点線部分は，制約条件を満たさないために探索する必要のない部分とする．各節点は，事例ベースの候補を表しているもの

[4] 利用した計算機は，SPARC Statoin 5 (CPU: microSPARCII(110MHz), Memory: 32MB) である．

とする．例えば，$Cd(1,2)$ はカテゴリ c の選択値を ϵ_{c1} に，カテゴリ $\bar{c}$ の選択値を $\epsilon_{\bar{c}2}$ に設定して選択する境界事例と平均事例を記憶している候補を表している．ここで，番号の数字が小さいほど選択する境界事例の数が少ないものとし，$Cd(0,0)$ は境界事例を一つも選択せず，平均事例だけを記憶している候補を表しているものとする．

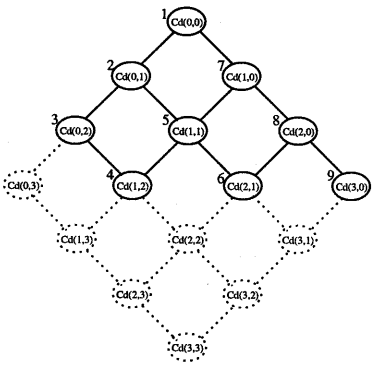

図 4: SABI の探索空間

3 章で述べたように，SABI の計算コストは $O(|Candidate| \times |TI| \times |A|)$ と表すことができる．$|Candidate|$ は候補の数であり，探索する節点の数と一致している．このため，探索する節点の数の削減が直接計算コストの削減につながることになる．そこで，バックトラック法を用いて探索範囲を狭める手法を 2 種類提案する．

(1) Topdown バックトラック法: Topdown バックトラック法は，選択する事例を増やしながら，言い換えると，選択値を大きくしながら探索を行うアルゴリズムである．また，探索は節点 1 を始点とした深さ優先探索を行うものとする．このため，節点は番号順に探索されることになる．探索中，ある節点の誤分類率がその前に探索した節点の誤分類率よりも増加すれば，前の節点にバックトラックしてそれ以上深く探索しないようにする．これは，それ以上深く探索を行っても，あるカテゴリの事例数が増加する一方で，カテゴリごとに記憶する事例数の偏りが大きくなってしまうためである．例えば，図 4 の場合では，節点 7 の次に節点 8 を探索して誤分類率が増加した場合，次の節点 9 は探索せず，節点 7 にバックトラックするようにしている．

(2) Bottomup バックトラック法: Bottomup バックトラック法は，Topdown バックトラック法とは逆に，すべての事例集合から選択する事例を小さくしながら探索を行うアルゴリズムである．探索は，どのカテゴリの選択値を広げても制約条件を満たさなくなるような節点を始点として，深さ優先探索を行うものとする．開始点は一つとは限らないため，すべての開始点から探索をする必要がある．例えば，図 4 の場合では，節点 4, 6, 9 を出発点として探索が行われる．そして，Bottomup バックトラック法と同様に，探索の途中である節点の誤分類率が増加すれば，前の節点にバックトラックしてそれ以上深く探索しないようにしている．

上記の二つの効率的な探索手法を用いた SABI を，それぞれ T-SABI, B-SABI と呼び，SABI との分類精度と事例ベース獲得までにかかる計算時間の比較を行う．用いたデータベース，実験方法，条件等は，3 節と同じにしてある．結果を表 3 に示す．なお，それぞれの手法の計算時間は SABI の計算時間を 100 として換算された値である．また，分類精度において太文字は水準 1 %で平均の差の検定を行い，SABI の分類精度の方よりも悪いと検定されたものを示している．

表 3 から，T-SABI, B-SABI のどちらも soybean を除くほとんどのデータベースで計算時間を大幅に削減できている．また，分類精度も変化することなく，探索範囲内で最適な事例ベースが見つかっている．soybean では，ほとんどの節点において誤分類率が 0 % で等しく，T-SABI, B-SABI を用いても探索する節点数を減らすことができていなかったために，計算時間を削減することはできなかった．これは，soybean の事例がカテゴリごとにはっきりと分かれており，どの事例を選択しても誤分類率が増加しないためだと考えられる．また，3 つのデータベースで T-SABI, B-SABI に適用した場合の方が SABI の分類精度よりも悪くなっている．これは，評価値として用いた誤分類率が探索空間において多峰性を持っており，局所最適解が得られたものと思われる．

結果的に，数日かかっていた SABI の計算時間を T-SABI では数秒に，B-SABI では数分に短縮することができた．また，T-SABI, B-SABI 両方の手法を用いても分類精度が悪くなるデータベースはなく，どちらかを用いれば最適解が得られているため，両方の手法を組み合わすことができれば局所探索を避けられる可能性がある．これらのことから，バックトラック法の導入による計算コストの軽減は有効であると思われる．

表 3: 分類精度と計算時間 (%)

データベース名	SABI	計算時間	T-SABI	計算時間	B-SABI	計算時間
breast-cancer	67.53	100.00	**56.73**	0.38	67.13	8.86
soybean	99.39	100.00	99.09	91.91	99.29	105.93
voting	91.55	100.00	91.67	0.46	91.02	30.97
hayes-roth	76.06	100.00	76.79	0.64	**61.37**	48.75
promoters	77.95	100.00	78.73	2.07	78.41	39.22
tic-tac-toe	88.94	100.00	**66.52**	0.02	89.53	3.98
cleveland	81.44	100.00	78.20	0.33	81.44	33.42
credit	79.30	100.00	78.91	0.04	78.09	11.87
iris	94.34	100.00	94.44	65.93	94.33	99.32
pima-diabetes	71.26	100.00	68.90	0.02	71.21	18.39

5 おわりに

本稿では，生成した平均事例に加えて境界事例を選択的に記憶して極小の事例ベースを獲得する SABI アルゴリズムを提案した．実験の結果，SABI はさまざまな問題領域で高い分類精度が得られた．また，最も類似したアルゴリズムと思われる IB3 との分類精度と選択率の比較を行い，SABI が非常に少ない事例の選択で高い分類精度を保つことができることを示した．さらに，T-SABI と B-SABI では，誤分類率が最小の事例ベースを探索する手法にバックトラック法を導入して，事例ベース獲得までにかかる計算コストを大幅に軽減することができた．

今後の課題としては，constructive induction [8] を導入して，tic-tac-toe のような特徴間の関係が分類に影響する問題領域に適用させることが考えられる．また，カテゴリごとに平均事例を一つだけ生成しているため，カテゴリにサブカテゴリがあるような領域では分類精度が悪くなると考えられる．このため，複数の平均事例を生成する必要がある [10].

参考文献

[1] 奥田健三，山崎勝弘: “事例ベース形推論とその応用”, 情報処理学会誌, Vol. 31, No. 2, pp. 244–254, 1990.

[2] Riesbeck, C. K. and Schank, R. C.: *Inside Case-Based Reasoning*, Lawrence Erlbaum, Hillsdale, 1989.

[3] Rosch, E. and Mervis, C. B.: “Family Resemblances: Studies in the Internal Structure of Categories”, *Cognitive Psychology*, Vol. 7, pp. 573–605, 1975.

[4] Cover, T. M. and Hart, P. E.: “Nearest Neighbor Pattern Classification”, *IEEE Transactions on Information Theory*, Vol. IT-13, pp. 21–27, 1967.

[5] 上原邦昭，谷澤正幸，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.

[6] Zhang, J.: “Selecting Typical Instances in Instance-Based Learning”, *Proc. of the Ninth International Conference on Machine Learning*, pp. 470–479, 1992.

[7] Murphy, P. M. and Aha, D. W.: *UCI Repository of Machine Learning Databases*, University of California, Department of Information and Computer Science, Irvine, CA, 1994.

[8] Matheus, C. J. and Rendell, L. A.: Constructive Induction on Decision Trees, *Proc. of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 645–650, 1989.

[9] Aha, D. W. and Kibler, D.: “Noise-Tolerant Instance-Based Learning Algorithms”, *Proc. of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 794–799, 1989.

[10] Datta, P. and Kibler, D: “Symbolic Nearest Mean Classifiers”, *AAAI-97 / IAAI-97 Proceedings*, pp. 82–87, 1997.