# ランダムアルゴリズムによる帰納学習の特性解析

徳永大輔[†] 上原邦昭[†,††]

機械学習アルゴリズムを評価する際には，計算時間の削減は大きな課題となっている．学習アルゴリズムは高い分類特性を持つものほどアルゴリズムの構造は複雑になり，特性を解析するために必要な計算時間は膨大なものとなっている．現在，学習アルゴリズムの近似的な数学的モデルを用いて特性を調べる手法が提案されているが，数学的モデルはモデル化を行い過ぎると，本来のアルゴリズムが持つ特性を失い，分類精度が不正確になる問題がある．逆に，モデル化が不十分だと解析に要する計算時間はほとんど削減されないという問題があった．本稿では，ランダムアルゴリズムを用いて，訓練事例集合のすべての組合せの中からサンプリングを行い，解析の精度を保ちつつ，計算時間を大幅に削減する Random Case Analysis を提案する．さらに，既存の手法では解析が困難な ID3 アルゴリズムや C4.5 アルゴリズムといった，複雑な帰納学習アルゴリズムに対して実験を行い，Random Case Analysis の有効性を確認する．

# Random Case Analysis of Inductive Learning Algorithms

DAISUKE TOKUNAGA[†] and KUNIAKI UEHARA[†,††]

In machine learning, it is important to reduce computational time to analyze learning algorithms. Learning algorithms have become complicated, and it requires much computational time to analyze them. Many researchers have presented analytic methods of learning algorithm by using approximately mathematical model. If we simplify the model too much, we may lose the essential behavior of the original algorithm, but if we don't simplify it enough, we need much computational time to analyze it. In our framework, called Random Case Analysis, we adapt the idea of Randomized Algorithms. By using Random Case Analysis, we can predict various aspects of learning algorithm's behavior, requiring much less computational time than analyses presented so far. Furthermore, we can easily apply our framework to practical learning algorithms, such as ID3 or C4.5.

## 1. はじめに

機械学習の分野で提案されている学習アルゴリズムを評価する際には，いくつかの検討すべき観点がある．まず第一に，ベンチマークとなる共通のテストデータの存在である．テストデータとしては UCI のデータベース[3]が多くの研究で利用されているが，データ中にはクラスや属性値にノイズが含まれている可能性があったり，クラスの決定に不要な属性が含まれていることがある．このため，属性がアルゴリズムに与える影響を特定することは困難であり，学習アルゴリズムの一般的な特性を推察するにとどまるという問題がある．もちろん，人工的に事例を発生させて統計学の背景から解析する手法も多く用いられているが，信頼区間の比率は分類精度の値によって異なるため，実験者は十分に考慮して評価を行わなければならないという問題がある．

第二に，学習に要する計算コストの問題がある．近年，計算コストの削減のために，学習アルゴリズムの分類精度の近似的な数学的モデルを用いて特性を調べる手法が提案されている．しかし，数学的モデルは，モデル化を行い過ぎると本来のアルゴリズムが持つ特性を失い，分類精度が不正確になる危険性があるため，実験者は十分注意して数学的モデルを作成しなければならない．たとえば，PAC（Probably Approximately Correct）学習モデル[9]は，概念クラスの学習可能性を議論するものであるが，PAC 学習モデルはワーストケースという特殊な状況を想定してモデル化を行っているために，実際の特性と大きく異なることが多いため，アルゴリズムの特性解析手法として用いられることは少ない．

† 神戸大学工学部情報知能工学科
Faculty of Engineering, Kobe University
†† 神戸大学都市安全研究センター
Research Center for Urban Safety and Security, Kobe University

逆に，モデル化が十分に行われないと，数学的モデルは学習アルゴリズムそのものに近くなり，事例の組合せを総当たりして分類精度を求めなければならず，計算時間が大きくなるという問題がある．たとえば，平均的事例解析（Average Case Analysis）[4]は学習アルゴリズムの分類精度の正確な期待値が算出できること，目標概念などの解析条件を実験者が自由に設定できることなどの利点を持っている．問題点としては，目標概念などの解析条件によっては数学的モデルが複雑になり，解析に必要な計算量が増大し，事実上，計算不可能となる場合があることがあげられる．このため，多くの平均的事例解析手法では，訓練事例数や属性数が小さい場合の解析しか行えていないことが述べられている[2),10)]．すなわち，平均的事例解析は計算コストの削減を行えるほど十分な数学的モデル化がされておらず，膨大な数の訓練事例集合の組合せを考慮しなければ分類精度を算出できないためである．

以上の観点から，平均的事例解析の利点であるモデル化による解析の柔軟さと，実験的手法の利点である適用の容易さを統合し，さらにランダムアルゴリズム[7]の考え方を用いて訓練事例集合からのサンプリングによって計算時間を大幅に削減する手法としてRandom Case Analysisを提案する．また，従来の手法では解析が困難であった複雑な学習アルゴリズムの特性を解析して，Random Case Analysisの有効性を示す．

## 2. Random Case Analysis

Random Case Analysisは，訓練事例集合の組合せすべてについて期待値を計算するのではなく，その中の任意の個数の組合せについて期待値を算出する手法をとっている．このため，属性数や訓練事例数が増え，訓練事例集合の組合せが非常に大きくなる場合でも，解析に必要な計算時間を非常に小さくおさえることができるようになっている．

### 2.1 基本的な考え方

Random Case Analysisは，ランダムアルゴリズム[7]の考え方を応用したものである．ランダムアルゴリズムは，乱数を利用して設計したアルゴリズム全般を指し，ソーティングやグラフ理論，暗号理論，線形計画法などの多くの応用分野で用いられている手法である．ランダムアルゴリズムの主な手法として，サンプリング法やランダムウォーク，ランダマイズドラウンディングなどがある．これらの手法のうち，Random Case Analysisはサンプリング法を用いて解析を行っている．

サンプリング法は，母集団からランダムサンプリングにより抽出したサンプルの集合が，母集団の性質をかなり良く受け継いでいるという性質を利用したものである．Random Case Analysisは，訓練事例集合の組合せを母集団として，この母集団が非常に大きなものである場合でも，サンプル（訓練事例）の集合を用いて計算される分類精度が非常に精度の良いものであることを利用している．具体的には，まず訓練事例集合を1つ生成し，学習アルゴリズムに適用して正しく学習されたかを確認する．これを1回の試行として，$N$ 回の試行を行い，正しく学習された回数 $X$ と全試行回数 $N$ との比を，学習アルゴリズムの分類精度の測定値 $K$ としている．解析アルゴリズムの基本的な流れを**図1**に示す．

(1) $n \leftarrow 0$, $X \leftarrow 0$.
(2) $l$ 個の事例をサンプリング（各属性ごとに定義された属性値の出現分布に基づいて得られる属性値集合を目標概念に適用し，クラスを決定する）し，訓練事例集合 $L$ とする．
(3) $L$ を評価を行いたい学習アルゴリズムに適用させ，正しく学習が行われていれば $X \leftarrow X+1$.
(4) $n \leftarrow n+1$.
(5) $n \geq N$ ならば(6)へ．$n < N$ ならば(2)へ．
(6) $K \leftarrow X/N$.

**図1** 解析の流れ
Fig. 1 A framework of analysis.

一般的に，適切な実験データが入手困難な場合，ランダムに訓練事例集合を生成して，学習アルゴリズムが正しく分類できる割合を算出する手法がよく用いられている．この手法は，中心極限定理を利用して，二項分布を正規分布に近似して区間推定を行うものである．一般には，すべての区間にわたって一定数の試行回数のもとで解析を行う手法がとられているが，分類精度の値によってその信頼区間の比率が異なるような解析結果が得られたり，論文によっては試行回数を十分とっていないために，精度の悪い解析結果となっているなどの問題が残されている．試行回数を固定して信頼区間の比率が変化する従来の手法より，信頼区間の比率を実験者が設定して，それにあわせて試行回数が変化する手法がより本質的であるといえるため，本稿ではランダムサンプリングの手法を採用している．

### 2.2 Chernoffの定理

Random Case Analysisは各試行がBernoulli試行であるため，二項分布に従っている．二項分布の挙動を数学的背景から説明するものとしてChernoffの定理[7]がある．Chernoffの定理は，二項分布の期待値（平均）と，測定値と期待値との誤差の比率（測定値が期待値から5%上に外れるなら比率は1.05となる）か

ら，その誤差の範囲を外れる確率を不等式の形で表した定理である．Chernoffの定理に基づいて，Random Case Analysisでは分類精度の測定値が誤差の比率の範囲内にあれば十分な精度を持つものと見なし，そのために必要な試行回数を決定している．Chernoffの定理を以下に示す．

**定理1（Chernoffの定理）：** $X_1,\cdots,X_N$ を $\{0,1\}$ の値をとる，互いに独立な $N$ 回のBernoulli試行とし，$\Pr(X_i=1)=p_i$，$\Pr(X_i=0)=1-p_i$ とする．さらに $X=\sum_{i=1}^{N}X_i$，$\mu=\sum_{i=1}^{N}p_i$ とする．ある実数 $\delta\in(0,1]$ に対して，$X$ の値がその期待値 $\mu$ の $1-\delta$ 倍以下である確率は，

$$\Pr(X<(1-\delta)\mu)<\exp(-\mu\delta^2/2)\stackrel{\text{def}}{=}F^-(\mu,\delta) \tag{1}$$

である．同様に，ある実数 $\delta(>0)$ に対して，$X$ の値が $\mu$ の $1+\delta$ 倍以上である確率は，

$$\Pr(X>(1+\delta)\mu)<\left[\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}\right]^{\mu}\stackrel{\text{def}}{=}F^+(\mu,\delta) \tag{2}$$

である．

Chernoffの定理を用いて $i$ 回目の試行の成否を $X_i=1$，$X_i=0$ の場合と考えると，$X$ は $N$ 回の試行において学習アルゴリズムが正しく分類を行った回数を指している．つまり，分類精度の測定値は $K=X/N$ となり，その期待値を $\mu_K=\mu/N$ とすると，次の補助定理1′を得ることができる．

**定理1′：** $X_1,\cdots,X_N$ を $\{0,1\}$ の値をとる，互いに独立な $N$ 回のBernoulli試行とし，$\Pr(X_i=1)=p_i$，$\Pr(X_i=0)=1-p_i$ とする．さらに $X=\sum_{i=1}^{N}X_i$，$\mu=\sum_{i=1}^{N}p_i$ とする．ある実数 $\delta\in(0,1]$ に対して，$X$ の平均 $K=\sum_{i=1}^{N}X_i/N$ の値がその期待値 $\mu_K=\sum_{i=1}^{N}p_i/N$ の $1-\delta$ 倍以下である確率は，

$$\Pr(K<(1-\delta)\mu_K)<\exp(-\mu_K\delta^2N/2)\stackrel{\text{def}}{=}F^-(\mu_K,\delta) \tag{3}$$

である．同様に，ある実数 $\delta(>0)$ に対して，$K$ の値が $\mu_K$ の $1+\delta$ 倍以上である確率は，

$$\Pr(K>(1+\delta)\mu_K)<\left[\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}\right]^{\mu_K N}\stackrel{\text{def}}{=}F^+(\mu_K,\delta) \tag{4}$$

である．

この補助定理より，分類精度 $K$ に対して，その期待値 $\mu_K$ の信頼区間 $[(1-\delta)\mu_K,(1+\delta)\mu_K]$ を考えると，測定された分類精度が信頼区間を外れる確率を式(3)，式(4)で表すことができることになる．

### 2.3 必要最小限の試行回数

式(3)，式(4)より，危険率 $F^-(\mu_K,\delta)$，$F^+(\mu_K,\delta)$ を定数として，$N$ を大きくすれば $\delta$ が小さくなることが分かる．つまり式(3)，式(4)は，試行回数 $N$ を増やせば分類精度の測定値 $K$ が分類精度の理論値に近づくことを表している．

$N$ を無限大にしたときの $K$ の値について考えると，式(3)と式(4)より以下の2つの式が導かれる．

$$\Pr(K<(1-\delta)\mu_K)<\exp(-\mu_K\delta^2N/2)\stackrel{N\to\infty}{\longrightarrow}+0 \tag{5}$$

$$\Pr(K>(1+\delta)\mu_K)<\left[\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}\right]^{\mu_K N}\stackrel{N\to\infty}{\longrightarrow}+0 \tag{6}$$

式(5)は，$N$ を無限大にすれば左辺の確率が0，$K$ が $(1-\delta)\mu_K$ より小さくなることはないことを意味している．言い換えると，$K\geq(1-\delta)\mu_K$ である．このことはすべての $\delta$ について成立する．ここで，$\delta\to0$ とすると，$K\geq\mu_K$ が導出できる．同様に，式(6)から $N\to\infty$ のとき $K\leq\mu_K$ となり，これらの式を同時に満たすのは $K=\mu_K$ のときのみとなる．すなわち，試行回数を無限大にしたときの値は理論値そのものとなる．

以上のことは2.1節の中心極限定理の考え方とも一致している．しかし，実際には有限回しか試行を行えないため，分類精度の測定値は理論値の近似値をとることになる．したがって，試行回数 $N$ を理論的に検討し，解析を行うために必要な最小限の試行回数を導出する必要がある．

信頼区間の幅を決定するしきい値 $\delta$ と信頼区間を下回る確率 $F^-$ を変数とする関数

$$f^-(F^-,\delta)\stackrel{\text{def}}{=}\frac{-2\ln F^-(\mu_K,\delta)}{\delta^2} \tag{7}$$

を定義する．この関数は，測定値が信頼区間を下回らないために必要最小限の試行回数を決定する関数である．このとき，式(3)を満たす $N$ の最小値 $N^-_{\min}$ は

$$N=\frac{-2\ln F^-(\mu_K,\delta)}{\mu_K\delta^2}=\frac{f^-(F^-,\delta)}{\mu_K}\stackrel{\text{def}}{=}N^-_{\min} \tag{8}$$

となる．同様に，$\delta$ と信頼区間を上回る確率 $F^+$ を変数とする関数

$$f^+(F^+,\delta) \stackrel{\text{def}}{=} \log_{\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}} F^+(\mu_K,\delta) \quad (9)$$

を定義する．この関数は，測定値が信頼区間を上回らないために必要最小限の試行回数を決定する関数である．このとき，式(4)を満たす $N$ の最小値 $N^+_{\min}$ は

$$N = \frac{\log_{\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}} F^+}{\mu_K} = \frac{f^+(F^+,\delta)}{\mu_K} \stackrel{\text{def}}{=} N^+_{\min} \quad (10)$$

となる．ここで，式(8)，式(10)より，測定値が信頼区間を外れないために必要最小限の試行回数を決定する関数

$$f(F^+,F^-,\delta) \stackrel{\text{def}}{=} \max\{f^+(F^+,\delta), f^-(F^-,\delta)\} \quad (11)$$

を定義すると，式(3)，式(4)を同時に満たす $N$ の最小値 $N_{\min}$ は，

$$\begin{aligned} N_{\min} &= \max\{N^+_{\min}, N^-_{\min}\} \\ &= \frac{\max\{f^+(F^+,\delta), f^-(F^-,\delta)\}}{\mu_K} \\ &= \frac{f(F^+,F^-,\delta)}{\mu_K} \end{aligned} \quad (12)$$

と表すことができる．

以上で，分類精度の測定値が分類精度の理論値に対して十分な信頼度を持つための必要最小限の試行回数 $N$ が導出できた．しかしながら，式(12)の定義では，分類精度 $\mu_K$ が小さいときには試行回数が反比例的に増えるという問題がある．この問題を避けるために，定理 $1'$ をさらに応用し，改めて検討を行う．

定理 $1'$ の $\mu_K$ は分類精度の期待値，つまり分類に成功する回数の平均の期待値であるが，ここでは分類に失敗する回数の平均の期待値（これを分類失敗精度の理論値と呼び $\mu_{K\text{miss}}$ と表すことにする）を考える．すなわち，分類失敗精度の測定値（$K_{\text{miss}}$ と表すことにする）が理論値に対し十分な信頼度を持つ場合，分類精度の測定値が $1-K_{\text{miss}}$ となることを利用すると，式(3)，式(4)より分類に失敗する確率は以下のように表すことができる．

$$\Pr(K_{\text{miss}} < (1-\delta)\mu_{K\text{miss}}) < \exp(-(1-\mu_K)\delta^2 N/2) \quad (13)$$

$$\Pr(K_{\text{miss}} > (1+\delta)\mu_{K\text{miss}}) < \left[\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}\right]^{(1-\mu_K)N} \quad (14)$$

式(7)と式(9)で定義した関数を利用すると，式(13)，式(14)はそれぞれ，

$$N = \frac{-2\ln F^-(\mu_K,\delta)}{(1-\mu_K)\delta^2} = \frac{f^-(F^-,\delta)}{1-\mu_K} \stackrel{\text{def}}{=} N^-_{\text{missmin}} \quad (15)$$

$$N = \frac{\log_{\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}} F^+}{1-\mu_K} = \frac{f^+(F^+,\delta)}{1-\mu_K} \stackrel{\text{def}}{=} N^+_{\text{missmin}} \quad (16)$$

と変形できる．さらに，両式を同時に満たす $N$ の最小値は式(11)を用いて，

$$\begin{aligned} N_{\text{missmin}} &= \max\{N^+_{\text{missmin}}, N^-_{\text{missmin}}\} \\ &= \frac{\max\{f^+(F^+,\delta), f^-(F^-,\delta)\}}{1-\mu_K} \\ &= \frac{f(F^+,F^-,\delta)}{1-\mu_K} \end{aligned} \quad (17)$$

と表すことができる．以上のことから，分類失敗精度の測定値が理論値に対して十分な信頼度を持つために必要最小限の試行回数 $N_{\text{missmin}}$ が求められる．

式(12)と式(17)より，分類精度または分類失敗精度が理論値に対して十分な信頼度を持つためには，

$$N \geq N_{\min} = \frac{f(F^+,F^-,\delta)}{\mu_K} \quad (18)$$

または

$$N \geq N_{\text{missmin}} = \frac{f(F^+,F^-,\delta)}{1-\mu_K} \quad (19)$$

のいずれかを満たすまで試行を繰り返せばよいことが分かる．ここで $\mu_K = X/N$ より，式(18)，式(19)は，

$$X \geq f(F^+,F^-,\delta) \quad (20)$$

$$N - X \geq f(F^+,F^-,\delta) \quad (21)$$

と変形される．この2つの不等式が Random Case Analysis の解析の終了条件となる．すなわち，分類精度の値にかかわらず試行回数は $f(F^+,F^-,\delta)$ 回以上，$2\cdot f(F^+,F^-,\delta)$ 回以下になるため，2.1節で述べたような問題もなく，計算時間が大きくなり過ぎる問題もないために，Random Case Analysis は有効な解析手法となっている．

### 2.4 解析アルゴリズムの改良

図1のアルゴリズムに式(20)，式(21)で与えられる試行回数を適用した Random Case Analysis の解析アルゴリズムを図**2**に示す．

この解析アルゴリズムでは，事例を発生する関数 **EXAMPLE**$(\alpha)$，訓練事例集合 $L$ をもとに学習を行う関数 **LEARN**$(L)$ と，概念記述 $\beta$ に従って属性 $T_a$ のクラスを決定する関数 **CLASSIFY**$(\beta, T_a)$ が必要となる．関数 **LEARN** は帰納学習アルゴリズム本体であり，関数 **CLASSIFY** は概念記述，属性値集合を入力として，関数 **LEARN** によって得られた

```
RandomCaseAnalysis(α, l, F+, F−, δ)
   α  : 目標概念
   l  : 訓練事例数
   F+: 上側危険率
   F−: 下側危険率
   δ  : 信頼区間幅
begin
   X ← 0 ;
   N ← 0 ;
   while (X < f(F+, F−, δ)
         and N − X < f(F+, F−, δ)) do
   begin
      L ← φ ;
      N ← N + 1 ;
      repeat l do
         L ← L ∩ {EXAMPLE(α)} ;
      β ← LEARN(L) ;
      (Tc, Ta) ← EXAMPLE(α) ;
      if Tc = CLASSIFY(β, Ta) then
         X ← X + 1;
   end ;
   output X/N ;
end.
```

図2 Random Case Analysis の解析アルゴリズム
Fig. 2 Algorithm of Random Case Analysis.

概念記述に基づいて属性値集合の属するクラスを決定する関数である．関数 **EXAMPLE** は，目標概念，属性ごとの属性値の発生確率分布，ノイズの発生確率を入力とする．具体的には，乱数を発生させ，その値と発生確率分布から各属性の値を決定し，次に得られた属性値集合を目標概念に適用してクラスを決定する．最後に乱数を発生させ，その値とノイズの発生確率によってノイズを含むかを決定し，それに応じてクラス，属性値にノイズを含める．これらの操作で得られたものを関数 **EXAMPLE** の出力とする．これらの関数と解析アルゴリズムによって，Random Case Analysis は目標概念や学習アルゴリズムに依存しない柔軟な解析が可能となっている．

## 3. 解析実験

### 3.1 従来の解析手法との比較

Random Case Analysis によって求められる分類精度が十分な信頼度を持つことを確かめるために，単調単項式（たとえばブール関数 $mono_{[001]}(x_1, x_2, x_3) = x_3$）の学習アルゴリズム（wholist）を PAC 学習，平均的事例解析，Random Case Analysis によって解析する．

単調単項式の PAC 学習に最小限必要な訓練事例数 $m$ は，誤分類率の上限 $\epsilon$，単調単項式の変数の数 $n$，確率 $\epsilon$ 以上で誤分類するような仮説を得る確率の上限 $\delta$ を用いて，

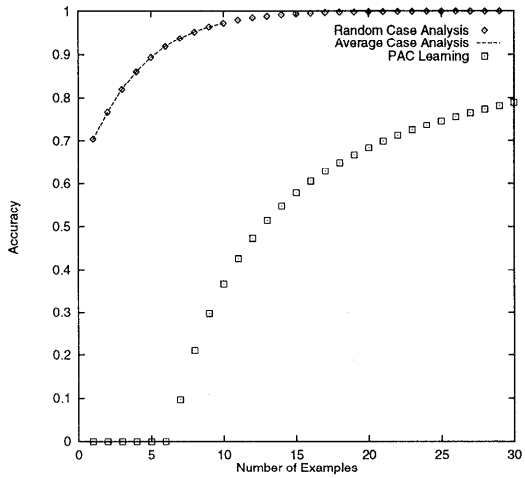



図3 単調単項式の解析結果
Fig. 3 Analysis of wholist.

$$m > \frac{1}{\epsilon}\left(n + \log_2 \frac{1}{\delta}\right) \tag{22}$$

と表すことができる[11]．これより，

$$1 - \epsilon < 1 - \frac{1}{m}\left(n + \log_2 \frac{1}{\delta}\right) \tag{23}$$

が導出される．式 (23) の左辺はワーストケースの分類精度の値であり，その値の上限が式 (23) の右辺で示される．

PAC 学習（解析条件は $\delta = 0.1$, $n = 3$）と平均的事例解析（解析条件は $mono_{[001]}$），Random Case Analysis（解析条件は $mono_{[001]}$, $F^+ = 0.05$, $F^- = 0.05$, $\delta = 0.005$）による解析の結果を**図3** に示す．

図3 では，Random Case Analysis が（分類精度の理論値を算出する）平均的事例解析とほぼ等しい精度の値を示しており，十分に満足できる性能を持っていることが確認できる．反面，PAC 学習で得られる値はワーストケースを想定した分類精度であるために，現実の挙動を示す平均的事例解析の値と大きく異なる結果となっている．なお，図3 において，平均的事例解析は最後まで解析が行えていないが，平均的事例解析は解析に要する計算時間が膨大になるという問題点を抱えているためである．**図4** に図3 の解析に要した計算時間を示す☆．

平均的事例解析では，訓練事例数が 10 のときの解析を計算するために約 1 カ月を要している．以後，訓練事例数が 1 増えると計算時間は約 8 倍ずつ増加するため，訓練事例数が 10 を超える場合の解析は事実

☆ SGI の Indy ワークステーション．CPU は R4600，133 MHz．メモリ 64 M．OS は IRIX 5.3．GNU Common Lisp 2.2 を使用．

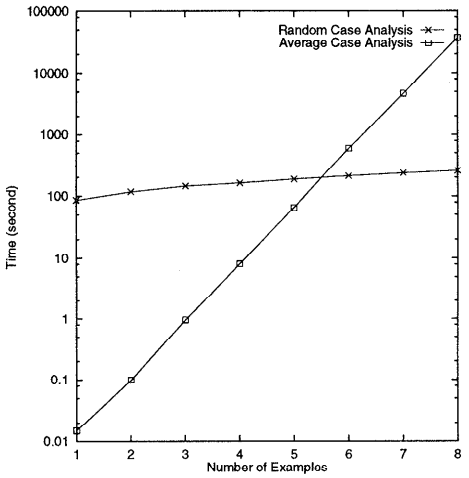



図4 単調単項式の解析に要する時間
Fig. 4 Computational time required to analyze wholist.

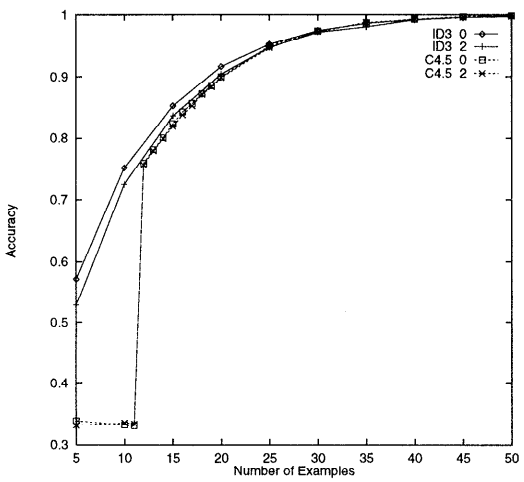



図5 不要属性を含む解析結果（ID3, C4.5）
Fig. 5 Analysis including irrelevant attributes (ID3, C4.5).

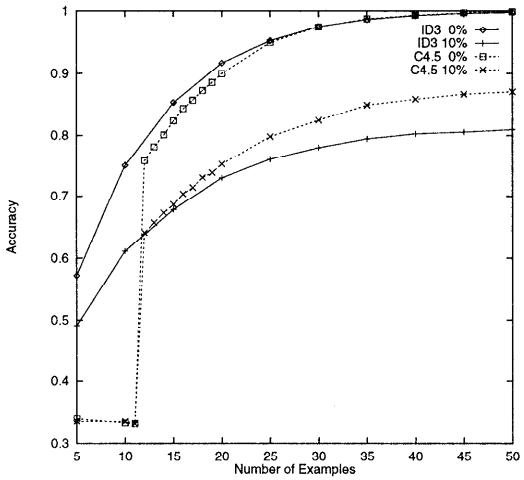



図6 クラスノイズを含む解析結果（ID3, C4.5）
Fig. 6 Analysis including class noise (ID3, C4.5).

上不可能になる．しかし，Random Case Analysis は訓練事例数が増えて，訓練事例集合の組合せが増えても試行回数に影響しないため，計算時間がほとんど増加せず，その結果，計算時間が著しく増加することがないという利点を持っていることが分かる．

### 3.2 ID3，C4.5 アルゴリズムの評価

従来の数学的手法では，ID3 アルゴリズム[5]や C4.5 アルゴリズム[6]などは数学的モデル化が困難なために解析が困難であった．Random Case Analysis は，これらのアルゴリズムの解析も容易に行うことができる．なお，C4.5 は ID3 を拡張したものであり，拡張によって分類精度がどのように影響するかを比較して調べている．帰納学習アルゴリズムに学習させる目標概念は次のとおりである．ただし，解析条件は $F^+ = 0.05$, $F^- = 0.05$, $\delta = 0.01$ としている．

- クラス数 3，属性数 3，各属性は 3 種類の属性値を 1/3 ずつの確率で発生する．
- 上の条件のもとで 24 種類の事例を一般化した概念を作り，目標概念とする．

まず，**図5** に不要属性を含まないときと 2 つ含むときの解析結果を示す．不要属性に関しては ID3，C4.5 ともに分類精度の著しい減少は見られない．両者ともに，不要属性に対して強いアルゴリズムであるといえる．

次に，**図6** にクラスノイズを含まないときと 10% 含むときの，各帰納学習アルゴリズムの解析結果を示す．C4.5 の分類精度の減少率に比べて，ID3 の分類精度の減少率は高い．クラスノイズに対しては C4.5 アルゴリズムの方が優秀である．C4.5 が ID3 から改良された点の 1 つとして，クラスノイズに強くなっていることが分かる．

最後に，**図7** に属性ノイズを含まないときと 10% 含むときの，各帰納学習アルゴリズムの解析結果を示す．属性ノイズについても同様に，C4.5 の分類精度の減少率より ID3 の分類精度の減少率の方が高く，属性ノイズに対しても C4.5 が ID3 よりも優秀であることが分かる．

また，図5，図6，図7 において，訓練事例数が少ないとき（訓練事例数が 11 以下のとき）に C4.5 の分類精度が極端に小さくなっているが，これは C4.5 が一定数以上の訓練事例がなければ決定木を生成しない（根だけの決定木ができる）仕様となっているためである．したがって，3 つのクラスのうちの 1 つが

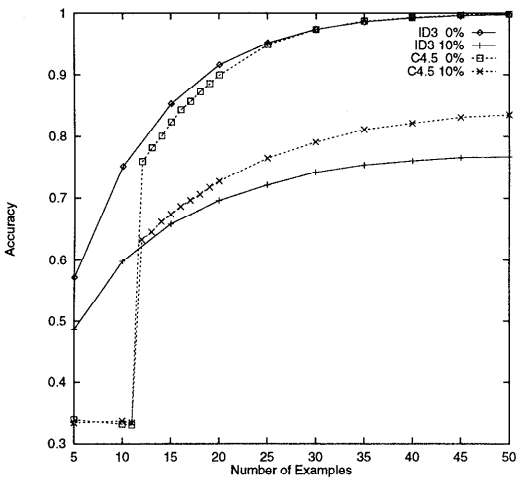



図7 属性ノイズを含む解析結果（ID3, C4.5）
Fig. 7 Analysis including attribute noise (ID3, C4.5).

1/3 の確率で根に与えられるので，分類精度の期待値は $1/3 = 0.333\cdots$ となっている．実際に，Random Case Analysis によって得られた解析結果はこの値を近似したものとなっている．このような特性は学習アルゴリズムを研究，開発する段階では発見しにくいものである．

従来の手法では，上記の特性が発見できないまま数学的モデルを作成し，その結果，誤った分類精度を算出してしまう危険性があった．しかし，Random Case Analysis では，解析アルゴリズムが学習アルゴリズムから完全に独立しているために，学習アルゴリズムの特性が変化しても，解析アルゴリズムを変更する必要がないという利点を持っている．

### 3.3 CBL，AQ アルゴリズムの評価

紙面の都合上割愛したが，上記の他に AQ アルゴリズム[8]，CBL アルゴリズム[1]の解析も行っている．これらのアルゴリズムも，C4.5 と同様に数学的モデルを作ることが困難なために，解析が行われていなかったものである．これらのアルゴリズムの特性としては，AQ は不要属性を含むときに分類精度の収束が遅くなっている．また，クラスノイズ，属性ノイズを含むときには分類精度が大幅に低いものとなっている．これは，AQ が事例を一般化するアルゴリズムであるため，不要属性やノイズといった特殊な事例を含むときには一般化の結果（学習によって得られる仮説）が目標概念と大きく違ったものになるためであると考えられる．また，CBL は CBL1（事例をすべて蓄えて分類，学習するアルゴリズム）と，CBL4（事例の膨大な記憶量を減らすために，事例数を削減し，さらにノイズ，不要属性に対して考慮したアルゴリズム）の解析を行った．不要属性については，CBL1 が分類精度が低下するのに対して，CBL4 ではほとんど低下していない．また，ノイズについても CBL1 と比べて CBL4 の低下量が小さいことが，解析結果から判明している．

## 4. 結　　論

本稿では，平均的事例解析と実験的手法とランダムアルゴリズムの考え方を統合した手法として Random Case Analysis を提案した．Random Case Analysis は平均的事例解析の利点である解析の柔軟さと，実験的手法の利点である適用の容易さをあわせ持っている．Random Case Analysis の利点として，以下のことがあげられる．

第 1 に，学習アルゴリズムを解析する際の最小試行回数は Chernoff の定理によって理論的に裏付けられたものとなっている．このため，複雑なアルゴリズムの解析に計算時間が非常に大きくなる平均的事例解析と比べて，少ない計算時間で解析を行うことができる．Random Case Analysis の解析アルゴリズムの計算時間は，訓練事例集合のサンプリング，学習，テストにかかる時間（一般には学習時間の比率が高い）と，危険率や信頼区間の幅を表すしきい値から導出される試行回数とに比例している．すなわち，学習アルゴリズムの種類や学習方法，概念の記述方法などに依存していない．このことが，複雑なアルゴリズムに対しても容易にかつ少ない計算時間で解析を行える理由となっている．

第 2 に，得られた分類精度が実際の挙動にきわめて近い挙動を示すため，学習アルゴリズムの研究，開発の面で有効な点である．第 3 に，目標概念の変更，設定が容易にでき，かつアルゴリズムの数学的モデル化の必要がなく，高度な確率的知識をほとんど必要としない．また，解析アルゴリズムから学習アルゴリズムがカプセル化されているため，実験者が意図していなかった特性が出る場合にも正しい解析が可能であり，学習アルゴリズムの研究，開発を行う際に非常に有効な手段であるといえる．

しかしながら，厳密に正しい結果を算出する平均的事例解析と比べ，Random Case Analysis では近似的な数値しか算出できない欠点を持っている．また，Random Case Analysis では試行回数を決定するためのパラメータ $F^+$，$F^-$，$\delta$ を経験的に決定しているため，解析の自由度が高い反面，実験者は注意してパラメータを決定する必要がある．特に，試行回数を求めるための式 (12)，式 (17) は $1/\delta^2$ にほぼ比例する

ため，解析の精度と計算時間とのトレードオフを十分に考慮する必要がある．

本研究は，平均的事例解析の計算コストを削減する手法を考案するために始めたものである．当初は，数学的モデルに訓練事例を適用する際に，サンプリングを用いて計算コストを削減する試みを考えていた．最終的に，学習アルゴリズムに直接訓練事例を適用する手法を用いたが，数学的モデルが明らかな場合には，平均的事例解析に訓練事例集合からのサンプリングを利用して，解析に要する計算時間を短縮した解析手法の提案も可能である．今後は，数学的モデルに対するサンプリングの適用についても検討したい．

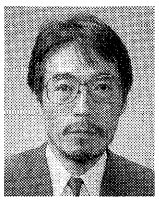
## 参考文献


1) Aha, D.W.: Case-Based Learning Algorithms, *Proc. Case-Based Reasoning Workshop*, pp.147–158 (1991).
2) Langley, P., Iba, W. and Thompson, K.: An Analysis of Bayesian Classifiers, *Proc. AAAI-1992*, pp.223–228 (1992).
3) Murphy, P.M. and Aha, D.W.: UCI Repository of Machine Learning Databases, Technical Report, University of California, Department of Information and Computer Science, Irvine, CA (1992).
4) Pazzani, M.J. and Sarrett, W.: Average Case Analysis of Conjunctive Learning Algorithms, *Proc. 7th International Conference on Machine Learning*, pp.339–347 (1990).
5) Quinlan, J.R.: Induction of Decision Trees, *Machine Learning*, Vol.1, pp.81–106 (1986).
6) Quinlan, J.R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).
7) Raghavan, P.: Lecture Notes on Randomized Algorithms, Research Report, RC 15340, IBM (1990).
8) Shi, Z.: *Principles of Machine Learning*, International Academic Publishers (1992).
9) Valiant, L.G.: A Theory of the Learnable, *C. ACM*, Vol.27, No.11, pp.1134–1142 (1984).
10) 坂本拓也，上原邦昭：N-Level 決定木アルゴリズムにおける平均的事例解析手法，情報処理学会論文誌，Vol.38, No.3, pp.419–428 (1997).
11) 篠原 歩，宮野 悟：PAC 学習，情報処理，Vol.32, No.3, pp.257–263 (1991).


(平成 9 年 5 月 19 日受付)
(平成 9 年 9 月 10 日採録)

**徳永 大輔**（学生会員）

昭和 47 年生．平成 8 年神戸大学工学部システム工学科卒業．現在，同大学院自然科学研究科情報知能工学専攻博士前期課程在学中．主に機械学習の研究に従事．



**上原 邦昭**（正会員）

昭和 29 年生．昭和 53 年大阪大学基礎工学部情報工学科卒業．昭和 58 年同大学院博士後期課程単位取得退学．大阪大学産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授を経て，同大学都市安全研究センター教授．情報知能工学科を兼担．平成元年より 2 年まで Oregon State University, Visiting Assistant Professor．平成 6 年より 8 年まで神戸大学総合情報処理センター副センター長．工学博士．人工知能，特に機械学習，マルチメディアデータベース，自然言語によるヒューマンインタフェースの研究に従事．1990 年度人工知能学会研究奨励賞受賞．人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，システム制御情報学会各会員．