ランダムアルゴリズムによる帰納学習の特性解析†

4B－9

徳永大輔　上原邦昭††
神戸大学工学部情報知能工学科

1. はじめに

機械学習の分野において，様々な帰納学習アルゴリズムが研究されているが，その特性を解析することは，研究，開発を進める上で重要な指針となっている．従来は実験による特性解析手法，もしくはPAC学習モデル[1]や平均的事例解析[2]に代表される理論的アプローチによる特性解析手法が主に用いられてきた．しかし，これらは複雑なアルゴリズムの解析を取り扱うことが困難である等の問題点がある．本研究では，帰納学習アルゴリズムの特性解析手法の問題点を解決する手段として，ランダムアルゴリズムを用いた特性解析手法であるRandom Case Analysisを提案する．

2. PAC学習と平均的事例解析

PAC (Probably Approximately Correct) 学習は，学習アルゴリズムの挙動を数学的に解析し，概念の学習可能性を確かめるものである．PAC学習の問題点として，解析はワーストケースを定式化して行なうため，得られた解析値が現実の帰納学習アルゴリズムの挙動と大きく異なる場合があること，定式化に高度な確率的知識を必要とするために，複雑な帰納学習アルゴリズムの特性解析は困難であること等がある．

平均的事例解析では，帰納学習アルゴリズムの平均的な挙動を表す数式モデルを作成して特性を解析している．平均的事例解析の利点として，学習アルゴリズムの分類精度（得られる概念が正しい分類を行なう確率）の正確な期待値が算出できること等がある．しかし，学習アルゴリズムの定式化には高度な数学的知識が必要であり，複雑な帰納学習アルゴリズムへの適用は困難である．また，解析条件によっては数式モデルが複雑になり，計算不可能となる場合がある．

†Random Case Analysis of Inductive Learning Algorithms
††Daisuke Tokunaga and Kuniaki Uehara
Department of Computer and Systems Engineering,
Faculty of Engineering, Kobe University

3. Random Case Analysis

ランダムアルゴリズムとは，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等の多くの分野で用いられている．本研究では，ランダムアルゴリズムの手法のうちサンプリング法を用いている．サンプリング法とは，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するものである．本稿では，帰納学習における事例の記述空間を母集団として，ランダムサンプリングにより抽出した事例を用いてアルゴリズムの特性解析を行なっている．

Random Case Analysisでは，N回の試行が行なわれ，各試行では帰納学習アルゴリズムによる1回の学習と1回の分類テストが行なわれる．訓練事例数がl個の時の解析アルゴリズムの流れを以下に示す．

1．l個の事例をランダムサンプリング

2．学習アルゴリズムに適用して学習概念を得る

3．学習概念が正しく事例を分類できるかをテスト

N回の試行のうち正しく分類できた回数をX回とすると，分類精度Kは$K = X/N$で求まる．ここで，試行回数Nが数回や数十回では十分な信頼度を持った解析値を得ることは難しいが，Nが大きくなるにつれて信頼度が高くなる．しかし，試行回数を多くすると計算量が比例して大きくなるという問題がある．そこで十分な信頼度を持った分類精度を得る最小限の試行回数を得るためにチェルノフの定理[3]を導入する．

チェルノフの定理. $X_1, \cdots, X_N$ を $\{0,1\}$ の値をとる互いに独立な N 回のベルヌーイ試行とし，$\Pr(X_i = 1) = p_i$，$\Pr(X_i = 0) = 1 - p_i$ とする．$X_1, \cdots, X_N$ の和を $X = \sum_{i=1}^{N} X_i$，X の期待値を $\mu = \sum_{i=1}^{N} p_i$ とする．ある実数 $\delta \in (0,1]$ に対して，X の値がその期待値 μ の $1-\delta$ 倍以下である確率は，

$\Pr(X < (1-\delta)\mu) < \exp(-\mu\delta^2/2) \stackrel{\text{def}}{=} F^-(\mu,\delta)$
である．同様に，ある実数 $\delta \in (0,1)$ に対して，X の値が μ の $1+\delta$ 倍以上である確率は，
$\Pr(X < (1+\delta)\mu) < \left[\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}\right]^{\mu} \stackrel{\text{def}}{=} F^+(\mu,\delta)$
である．

目標概念を α，信頼区間を δ，危険率（分類精度の測定値を中心とした信頼区間内から理論値が外れる確率）を F^+, F^-，分類精度を μ_K とする時，チェルノフの定理の N を試行回数，X を分類に成功した回数として適用すると，分類精度が理論値に対して十分な信頼度を持つためには，上式より，
$N \geq \frac{f(F^+,F^-,\delta)}{\mu_K}$ 又は $N \geq \frac{f(F^+,F^-,\delta)}{1-\mu_K}$
のいずれかを満たすまで試行を繰り返せば良いことがわかる．概念 α から事例を発生する関数を **EXAMPLE**(α)，訓練事例 L から学習を行なう関数を **LEARN**(L)，概念記述 β から属性 T_a のクラスを決定する関数 を **CLASSIFY**(β, T_a) と定義すると，上式の試行回数 N を用いた解析アルゴリズムは図1で示されるアルゴリズムとなる．

Algorithm RandomCaseAnalysis$(\alpha, l, F^+, F^-, \delta)$
begin
  $X \leftarrow 0$;
  $N \leftarrow 0$;
  while $(N < \frac{f(F^+,F^-,\delta)}{\frac{X}{N}}$ **and** $N < \frac{f(F^+,F^-,\delta)}{1-\frac{X}{N}})$ **do**
  begin
    $L \leftarrow \phi$;
    $N \leftarrow N+1$;
    repeat l **do**
      $L \leftarrow L \cap \{$**EXAMPLE**$(\alpha)\}$;
    $\beta \leftarrow$ **LEARN**(L) ;
    $(T_c, T_a) \leftarrow$ **EXAMPLE**(α) ;
    if $T_c =$ **CLASSIFY**(β, T_a) **then**
      $X \leftarrow X+1$;
  end ;
  output X/N ;
end.

図 1: Random Case Analysis アルゴリズム

4. 解析実験

Random Case Analysis によって得られる値の信頼性を確認するために，図1のアルゴリズムを用いて評価実験を行なう．評価対象として 2-level 決定木アルゴリズムを用いている．解析条件は $F^+ = F^- = 0.05, \delta = 0.01$ である．平均的事例解析による分類精度の理論値を折れ線で表し，Random Case Analysis による解析値とその信頼区間の上限と下限を + で表している．図2より，Random Case Analysis によって得られる値の精度が十分高くなっていることがわかる．

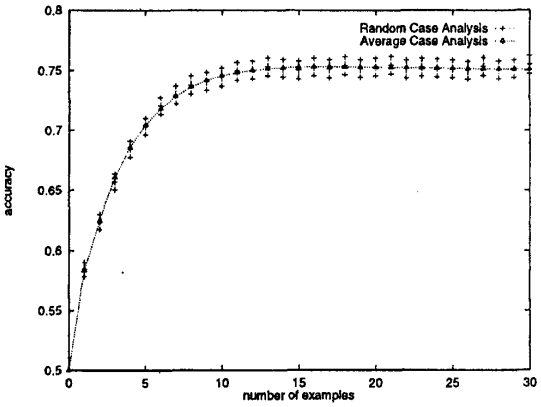

図 2: 平均的事例解析との比較

5. おわりに

Random Case Analysis の利点として以下のことが挙げられる．第一に，複雑なアルゴリズムに対して計算量が非常に大きくなりがちな PAC 学習や平均的事例解析に比べて，Random Case Analysis は学習アルゴリズムの種類やその学習方法，概念の記述方法などに依存しないため，複雑なアルゴリズムに対しても容易に，かつ少ない計算量で解析を行なえる点である．第二に，得られた分類精度が実際の挙動に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．第三に，目標概念の変更，設定が容易にでき，かつアルゴリズムの定式化の必要が無く高度な確率的知識をほとんど必要としないため，様々な環境下での特性解析を容易に可能にしている点である．

参考文献

[1] Valiant, L. G., "A Theory of the Learnable," *C. ACM*, pp. 1134–1142 (1984).

[2] Pazzani, M. J. and Sarett, W., "Average Case Analysis of Conjunctive Learning Algorithms," *Proc. of the Seventh International Conference on Machine Learning*, pp. 339–347 (1990).

[3] Raghavan, P., "Lecture Notes on Randomized Algorithms," Research Report, RC 15340, IBM (1990).