解　説　能動学習

4. 能動学習に関する計算論的学習理論の研究

Research on Active Learning in Computational Learning Theory by Naoki ABE and Atsuyoshi NAKAMURA (C & C Research Laboratories, NEC).

安倍直樹[1]　中村篤祥[1]

1 NEC C & C研究所

1. はじめに

「能動学習」とは，学習者が自発的に行動することにより学習を進めていくような学習形態の総称であり，さまざまな形態を含んでいる．この中で代表的なものに「質問による学習」と「強化学習」がある．前者は，質問を行うことによりデータを得て，概念や言語を学習する問題であり，関数を実験から推定する実験計画法のモデル化と考えることもできる．各時点でそれまでに得た情報から考えて最も効果的な質問点を逐次的に決めてやることにより，少ない質問数で学習対象の概念（言語）のよい近似を求めることが目的となる．これに対して後者は，環境内を動き回り，行動規則を獲得していくロボットなどのモデル化であり，統合的な学習モデルである．ここでは，学習者の目的は学習対象の概念のよい近似を求めること自体にはなく，自らとる行動に対し環境から与えられる報酬（これを強化信号と呼ぶ）をもとに行動規則を獲得し，それを用いて累積の利潤を最大化することである．

計算論的学習理論の分野においては，前者の定式化である「質問学習モデル」[2),3)]は，この分野の主要な学習モデルの1つとして活発に研究されてきた．また，後者の強化学習は応用システムに盛んに適用されており，最近ではその性能評価を目的とした「最適選択の逐次型学習モデル」[1)]も提案されている．本稿では，これら2つの能動学習のモデルに関する理論的な研究成果について概観する．（2章で「質問による学習」について，3章で「強化学習」について解説する．）

2. 質問による学習

2.1 単調DNFの学習

まず最初に具体例として，単調DNFというブール関数のクラスを能動的に学習する方法を説明しよう．

DNF（Disjunctive Normal Form）とは，積和形式と呼ばれているブール関数の表現形であり，次のように定義される．$x_1, \cdots, x_n$ をブール変数（0と1の2値の値をとる変数）とする．ブール変数 x_i とその否定 $\overline{x_i}$ をリテラルという．リテラルの論理積（ANDで繋いだもの）を項といい，項の論理和（ORで繋いだもの）がDNFである．否定を含まないDNFを単調DNFという．

さて，次のような問題を考えてみよう．

問題1　ブール変数 $x_1, \cdots, x_n$ に対応する n 個の値を入力すると，0か1の値を出力するブラックボックスがあるとする．このブラックボックスは単調DNFで表現される関数であることが分かっているとする．このブラックボックスに対する m 個の入出力ペア $(\vec{x}_1, y_1), \cdots, (\vec{x}_m, y_m)$ が与えられたとき，それらすべてに矛盾しない単調DNFを求めよ．

この問題は，次の意味で実は手に負えないと考えられている問題である：ブラックボックスの関数が高々 k 項からなる単調DNFであるとき，与えられた入出力ペアの集合に無矛盾な高々 k 項からなる単調DNFを求める問題は，NP完全（手に負えないと考えられているクラスの問題）である[14)]．

しかし，学習アルゴリズムが選んだ入力に対するブラックボックスの出力が得られる場合には，手に負える問題となるのである．以下にValiant

```
H←∅
for i=1 to m {
  if H(x_i)≠y_i then {
    z←x_i
    for j=1 to n {
      if z の第jビットが1 then
        z の第jビットを反転
        if F(z)=0 then {
          z の第jビットを反転(元に戻す)
        }
      }
    }
    z で値1が割りあてられている変数すべてからなる項をH
    に加える
  }
}
```

図-1 単調DNFの能動学習アルゴリズム

によるアルゴリズム[3),20)]を示す（図-1）．ブラックボックスの関数が，単調DNF F で冗長性なく表現されるとする．今，すでにみつかっている項の集合を H とする．Valiantのアルゴリズムは $H=\emptyset$ から始めて，与えられた入出力ペアの集合に H の表す関数が矛盾するかぎり F の項を1つずつみつけて，H に加えるというものである．つまり，H は常に F（の項の集合）の部分集合であり，$F(\vec{x})=0$ であれば $H(\vec{x})=0$ になっている*．今，与えられた入出力ペアの中から $y_i=1$ であるのに $H(\vec{x}_i)=0$ であるようなペア $(\vec{x}_i, y_i)$ をみつけたとする．これはまだみつかっていない F の項が $\vec{x}_i$ により満たされていることを意味する．この $\vec{x}_i$ を種に，入力値を選んでブラックボックスの出力値を得ることにより，新しい項をみつけることができる．$\vec{x}_i$ を0と1のビット列とみなす．値が1のビット1カ所を反転させてその入力に対するブラックボックスの出力値をもらう．もし出力値が1ならばそのまま，0ならばビットを反転して元に戻す．これを値が1のすべてのビットに対して行う．このようにして作られたビット列により，値1が割りあてられている変数すべてからなる項が新しくみつかった項である．

F に含まれる項の数を s とすると，上記アルゴリズムは明らかに，自ら選んだ高々 sn 個の入力に対するブラックボックスの出力を貰うことにより，n，m，s に関して多項式時間で問題1を解くアルゴリズムなのである．

* ここで，F，H は項の集合としてもそれが表す関数の意味でも使うことにする．

2.2 PAC学習と所属性質問

さて，前節でみたように，能動的にサンプルを選ぶ場合と，受動的に与えられたサンプルしか使わない場合とでは，明らかにアルゴリズムの能力に差が出る場合がある．計算論的学習理論の分野では，高い確率でよい近似を行うことを基準とするPAC（Probably Approximately Correct）学習において，能動的なサンプリングである所属性質問（membership query）を許すことにより，学習能力にどのくらい差が出るかという研究が盛んに行われてきた．次節でそれらの研究成果について紹介するための準備として，ここで考える学習の枠組みについて簡単に説明する．詳しくは文献7）を参照されたい．

ここでは概念学習というものを考える．概念学習では，ある概念クラスに属する概念を事例から推定することが目標となる．概念とは，学習領域と呼ばれる集合の部分集合である．前節の例では，$\{0,1\}^n$ が学習領域であり，その上で定義される任意のブール関数 f に対し，集合 $\{\vec{x}: f(\vec{x})=1\}$ が概念である．単調DNFのクラスは，単調DNFで表現される関数が表す概念の集合とみることができるので，1つの概念クラスである．また逆に概念を，含む点の上のみで1の値をとるブール関数とみなすことができる．学習領域上の点は，ある分布 D に従って発生するものと仮定する．2つの概念 c と h の間の距離は，2つの集合の対象差の点が発生する確率とする．概念 c と h の間の距離が ε 以下の場合，h は c の ε 近似であるという．

概念クラス C がPAC学習可能であるとは，C に属する任意の概念 c と学習領域 X 上の任意の分布 D，任意の $0<\varepsilon, \delta<1$ に関して，$1-\delta$ の確率で ε 近似仮説 h を出力する多項式時間アルゴリズムが存在すること，と定義される．ただし，ここでの多項式時間とは，$1/\varepsilon$，$1/\delta$ と学習領域サイズ，概念サイズ** に関して多項式で表現される時間を意味する．また受動的な情報として一般に，分布 D に従って発生した点 x と $c(x)$ の値が1ステップで得られると仮定する．$(x, c(x))$ を例と呼び，例の集合をサンプルと呼ぶ．出力される仮説 h は，学習領域上の任意の点 x

** 前出の単調DNFの例では，学習領域サイズとしては n，概念サイズとしては s などを用いる．

に対し多項式時間で $h(x)$ を計算できれば何でも構わないことにする．

所属性質問とは，学習領域内の点 x が学習対象の概念 c に含まれているかを尋ねる質問で，入力 x に対し1ステップで $c(x)$ の値を返すブラックボックスの存在を仮定する．所属性質問は，欲しい情報を積極的に尋ねる質問であり，能動学習の典型的な形態といえる．

2.3 所属性質問の威力

この節では，所属性質問の威力に関する計算論的学習理論分野での研究成果について，盛んに研究されているDNFおよびその双対形であるCNF（Conjunctive Normal Form：和積形式）を中心に説明する．DNF（CNF）全体のPAC学習可能性については，Valiant[20]がPACモデルを提案して以来の重要な未解決問題であり，否定的にみられている問題の1つである．

2.3.1 DNF

DNF（CNF）の部分クラスで，所属性質問を使えばPAC学習可能であるが使わなければ手に負えそうもないというクラスで，まずあがるのは単調DNF（単調CNF）である．実際，**2.1**節で説明した図-1のアルゴリズムは，所属性質問を用いたPAC学習アルゴリズムであることが示せる．所属性質問を使わなければ手に負えそうもないということは，問題1がNP完全であることからわかりそうであるが，NP完全になるのは出力する仮説を高々 k 項からなる単調DNFにかぎった場合であることに注意してほしい．ここでのPAC学習可能性の定義は，仮説クラスに制限を加えない場合であるので，別の証拠が必要である．実は証拠として，「もし，単調DNFが所属性質問を使わないでPAC学習可能であれば（手に負えないと思われている）DNF全体が学習可能である．」ということが示されている．つまり，単調DNFの学習アルゴリズム A' を用いて，DNF全体の学習アルゴリズム A を次のように構築することができるのである．

A を適用する学習領域を $\{(x_1, \cdots, x_n)\}$ とすれば，A' は2倍の次元の領域 $\{(x_1, \cdots, x_n, x_1', \cdots, x_n')\}$ に適用する．A を適用する学習対象のDNFを $f(x_1, \cdots, x_n)$ とすれば，A' は $f'(x_1, \cdots, x_n, x_1', \cdots, x_n')$ を学習する．ただし，f' は f に出現するすべての負のリテラル $\overline{x_i}$ を x_i' により置き換えて得られる単調DNFである．A は A' を起動し，A' が例を要求した時，得られた例 $((a_1, \cdots, a_n), b)$ を $((a_1, \cdots, a_n, \overline{a_1}, \cdots, \overline{a_n}), b)$ に変換して A' に渡す．ただし，$\overline{a_i}$ は a_i を反転（0ならば1，1ならば0）したものである．A' が仮説として単調DNF $h(x_1, \cdots, x_n, x_1', \cdots, x_n')$ を出力したら，A は，DNF $h(x_1, \cdots, x_n, \overline{x_1}, \cdots, \overline{x_n})$ を出力する．

これは，DNF全体の所属性質問を使わないPAC学習問題が，単調DNFの学習問題に還元できることを意味する．還元の厳密な定義に関しては，文献18）を参照されたい．

所属性質問なしでは単調DNFの双対形である単調CNFの学習問題でさえ手に負えそうもないが，所属性質問を使えばもっと大きなクラスが学習可能である．リテラルの論理和で $\overline{x_1} \vee \overline{x_2} \vee \overline{x_3} \vee x_4$ のように正のリテラル（否定なしの変数）を高々1つしか含まないものをホーン節という．$\overline{x_1} \vee \overline{x_2} \vee \overline{x_3} \vee x_4$ は $x_1 \wedge x_2 \wedge x_3 \Rightarrow x_4$ と等価であり，ホーン節は論理でいうところの“含意”を表しているともいえる．ホーン節の論理積，つまり各節（リテラルの論理和）に高々1つの正のリテラルを含むCNFをホーン文（Horn sentence）という．Angluinら[4]は，ホーン文が所属性質問を使えばPAC学習可能であることを示した．正のリテラルを1つも含まないCNFは，正負のリテラルおよび0と1の値を逆にみれば単調CNFであるから，ホーン文は単調CNFより大きなクラスであり，ほとんど単調なCNFのクラスといえる．

しかし，非単調の領域にもう一歩踏み込むと，そこは所属性質問を使っても手に負えそうもない領域に突入してしまう．つまり，各節が高々2つの正のリテラルを含むCNF（2-quasi-Horn式）にCNF全体の学習が（所属性質問を使うPAC学習の意味で）還元できることが知られている．2-quasi-Horn式の所属性質問を使う学習アルゴリズム A' を用いて，CNF全体を学習するアルゴリズム A を構築できるのである．DNFに関して上で説明したように，所属性質問なしの場合にはCNF全体の学習問題を単調CNFの学習問題まで還元できるのに対し，所属性質問ありの場合には，2-quasi-Horn式の学習問題までしか還元できない．これはアルゴリズム A' を用いてCNF全体を学習するアルゴリズム A を構築する

際，A'が尋ねるすべての所属性質問にAが答えなければならないことが問題を難しくしているからである．つまり，Aを適用する学習領域XをA'を適用する学習領域X'へ対応させる写像が上への写像でない場合，X'上で対応するXの点が存在しない場合がある．そこで，そのような点に対してAがA'から質問された時に，常に0（または常に1）と答えても矛盾しないような概念の学習にA'の適用範囲をかぎる必要がある．実際，変数x_iとx_i'が異なる値をとるという条件は，$(x_i \vee x_i') \wedge (\overline{x_i} \vee \overline{x_i'})$と2-quasi-Horn式で書けるので，$A$の学習対象のCNF $f(x_1, \cdots, x_n)$に対し，A'に$f'(x_1, \cdots, x_n, x_1', \cdots, x_n')$とすべての$i$に対する$(x_i \vee x_i') \wedge (\overline{x_i} \vee \overline{x_i'})$を繋げた2-quasi-Horn式を学習させることにすれば，AはA'の所属性質問に対し，対応するX上の点がない場合には常に0を返してやればよいことになる．ただし，f'はfに出現するすべての正のリテラルx_iを$\overline{x_i'}$により置き換えることにより得られる正のリテラルを含まないCNFである．

DNFの部分クラスの学習可能性を示した結果で，任意のDNFの学習可能性に最も近づいていると思われる結果はすべて所属性質問を使った学習アルゴリズムを用いている：2μDNF（各変数が高々2回しか現れないDNF）[8),10),17)]，CDNF（CNF表現によるサイズがDNF表現の多項式サイズのDNF）[9)]，一様分布の下でのDNF[13)]．DNF全体の所属性質問を使ったPAC学習問題は3μDNF（各変数が高々3回しか現れないDNF）の学習問題に還元できるので，2μDNFはホーン文と同様に学習できそうなクラスの端に位置していることになる．

さて，このように所属性質問を使うことにより，より一般的なクラスの学習可能性が導かれることを説明したが，未解決のDNF（CNF）全体の学習に関して所属性質問は役に立たないという次のような結果が示されている：ある暗号学上の仮定のもとでは，DNF（CNF）は所属性質問を使わないでPAC学習可能であるか，または所属性質問を使ってもPAC学習不可能かのいずれかである[6)]．

2.3.2 その他のクラス

所属性質問なしでは手に負えそうもないが，所属性質問を使えばPAC学習できるというそのほかの結果としては，DFA（決定性有限オートマトン）[2)]，μブール式（各変数が高々1回しか現れないブール式）[5)]が有名である．これらの学習困難性の結果は，公開鍵暗号の安全性を根拠としている[15)]．所属性質問を使ってもPAC学習が手に負えそうもないクラスとしては，3μブール式，NFA（非決定性有限オートマトン），文脈自由文法などがある[6)]．

3. 強化学習と最適選択の逐次型学習

3.1 マルコフ決定プロセス

強化学習の代表的な枠組みがマルコフ決定プロセス（Markov Decision Process）である．このモデルでは，学習者のおかれた環境は学習者の行動により制御されたマルコフ・プロセスであると仮定される．まず，環境中には有限個の状態が存在する．各時間ステップtにおいて，学習者が有限個の行動のうち1つ$(a \in A)$を選択すると，確率的に状態遷移が起こる．すなわち，次の時間ステップ$t+1$における状態は，現在の状態$s \in S$ととられた行動$a \in A$により定まる確率$P(s, a, s')$で$s' \in S$となる．（Pは$\sum_{s' \in S} P(s, a, s') = 1$なる遷移確率関数．）また，この時学習者は利潤として$r_t = R(s, a)$を得る．学習者は，過去の履歴と状態から行動への写像であり，その目的は累積利潤の最大化である．累積利潤の定義としては以下の2つのものが知られている：（1）有限時間の場合（Finite horizon）；$\sum_{i=1}^{t} r_i$，（2）無限時間の場合（Infinite horizon）；$\sum_{i=1}^{\infty} \gamma^{i-1} r_i$．Infinite horizonの定義では，$i$ステップ目に得られる利潤は今得られる利潤よりも一定倍率価値が低く見積もられているのである$(0 < \gamma^{i-1} < 1)$．**図-2**にマルコフ決定プロセスの簡単な例を示した．この例では，利潤が得られるのは，状態4で行動aをとった時と状態3で行動bをとった時にかぎる．また，状態0から始めると，行動aを続けて3度選択した場合か，行動bを続けて3度選択した場合にもっとも高い確率で利潤が得られる．

Infinite horizonの場合のマルコフ決定プロセスに対しては，つねに最適な戦略が存在し，しかも漸近的に最適な戦略に収束するような単純な学習方式が存在することが知られている[21)]．この学習方式（Q-learning）を以下に説明する．まず，

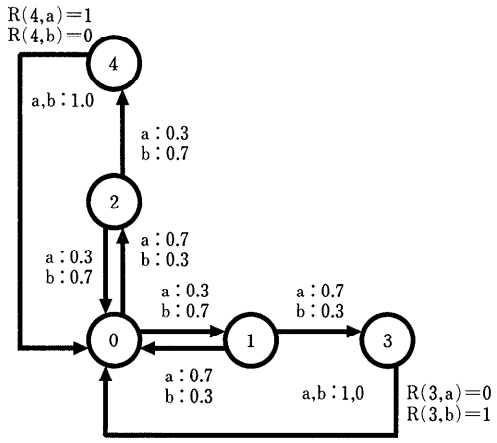


図-2 マルコフ決定プロセス

状態を行動へ写像する戦略で最適なものを π^* とする．(そのような最適戦略の存在が保証されている．) 次に，状態 s において行動 a をとった時に期待される累積利潤 (Q-value) を $Q(s, a)$ とすると，これは以下のように再帰的に与えられる．

$$Q(s, a)=R(s, a)+\gamma \sum_{s' \in S} P(s, \pi^*(s), s') \cdot \max_{a' \in A} Q(s', a')$$

また，逆に Q-value が定められると，最適な戦略は以下のように与えられる．

$$\pi^*(s)=\arg \max_{a \in A} Q(s, a)$$

したがって，環境を定めている P，R をあらかじめ知らずに，環境を探索することによって Q-value を学習することができれば，最適戦略を獲得できるのである．実際，環境内を動き回り，状態 s から行動 a をとって状態 s' に遷移するたびに以下の手続きに従って推定値 Q_t を更新することにより，Q-value を学習することができる．

$$Q_t(s, a)=(1-\alpha)Q_{t-1}(s, a)+\alpha(r_t+\gamma \max_{a' \in A} Q_{t-1}(s', a'))$$

ここで α は学習レートである．この方法を用いると，十分大きい t に対して，Q_t が Q に漸近的に確率収束していくことが知られている[21]．

このような結果により Q-learning は，単純でしかも理論的に性能を保証された学習法として人気がある．しかし，強化学習の枠組みにおける学習者（または行動者）の真の目的は最適戦略の獲得ではなく累積利潤の最大化であることを思い起こすと，上のような理論的な保証の現実問題における意味に疑問が起こる．つまり，現実問題においては，実行可能な試行回数での利潤の最大化が目的なのであり，「t が十分大きい」頃になって最適戦略が学習されてもあまり意味がないのではないか．また，無限試行数を仮定すると，最適な戦略の学習のための探索と獲得された知識をもとにした利潤の最大化の間のトレードオフの問題が顕在化しないのではないか．このような問題意識から提案されたのが，文献 1) の「最適選択の逐次型学習モデル」である．

3.2 最適選択の逐次型学習モデル

ここでは上記の強化学習の枠組みに沿ってこのモデルを説明しよう．マルコフ決定プロセスにおいては，環境を決定するパラメータである P と R を固定すると，最適戦略 $\pi^*_{P,R}$ が定まった．したがって学習者/行動者の善し悪しは，最適戦略 $\pi^*_{P,R}$ と比較して累積利潤がどの程度少なくなるかによって測ることができる．すなわち，環境パラメータ (P, R) に対する戦略 π の t 回の試行における期待累積利潤を $R(\pi, (P, R), t)$ とする時，戦略 π の期待 regret $G(\pi, (P, R), t)$ を以下のように定義する．

$$G(\pi, (P, R), t)=R(\pi^*_{P,R}, (P, R), t)-R(\pi, (P, R), t)$$

固定された環境パラメータ (P, R) に対してはもちろん戦略 $\pi^*_{P,R}$ がゼロ regret を達成するが，この戦略はそれ以外の環境パラメータに対してはよい性能をあげることは期待できない．可能な環境パラメータ (P, R) のいずれに対しても低い regret を達成するためには，学習型の戦略を採用することが必要なのである．すなわち，環境パラメータ (P, R) のとり得る範囲を C とする時，C における最悪の場合の regret を最小化することが望まれる．そこで，戦略 π の C に対する期待 regret $G(\pi, C, t)$ を以下のように定義する．

$$G(\pi, C, t)=\max_{(P,R) \in C} G(\pi, (P, R), t)$$

「最適選択の逐次型学習モデル」では，どれだけ緩やかな関数 $F(t)$ によって期待 regret $G(\pi, C, t)$ を上限できるかによって C の潜在的な学習の困難性を特徴づける．(たとえば，$\sqrt{t}$ よりは $t^{1/3}$ が，$t^{1/3}$ よりは $\log t$ の方が緩やかである．)

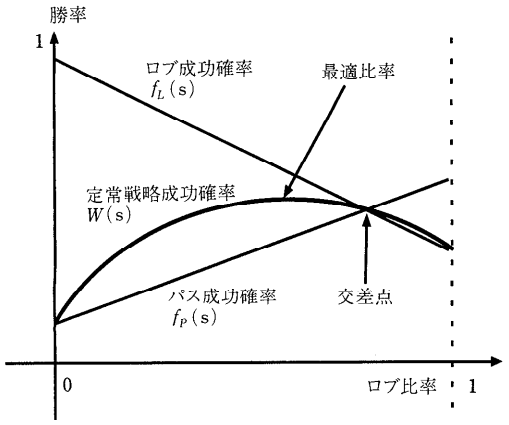


図-3 ロブ・パス問題

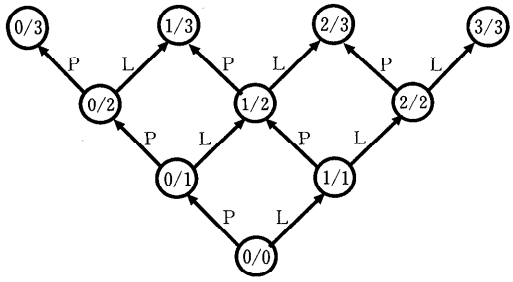


図-4 ロブ・パス問題のマルコフ決定プロセス

3.3 ロブ・パス問題

一般のマルコフ決定プロセスに関して上記の最適選択の逐次型学習モデル内での評価をすることはかなり困難である．そこでここでは，「ロブ・パス問題」というある単純な問題に関する評価[1]を紹介する．「ロブ・パス問題」は，テニスゲームのモデル化であるが，もともと動物行動学の文脈で発案されたモデル[11]を逐次型学習問題として定式化したものである．まず，学習者のとれる行動はロブかパスの2種類のショットのいずれかしかないと仮定する．また，学習者の状態 s はそれまでに打ったショット中でのロブの比率と同一とする．また行動に対する報酬については，単純のため各ショットの勝負で測るとして，ロブの勝率 $f_L(s)$ は s に関する単調減少関数，パスの勝率 $f_P(s)$ は s に関する単調増加関数で定められるとする．文献1）においては，とくに $f_L(s)$，$f_P(s)$ ともに線形関数で表される場合を考察している（**図-3** 参照）．ロブ・パス問題は，状態数が無限のマルコフ決定プロセスの特殊なケースとも考えることができる（**図-4** 参照）．

まず，ロブ，パスの勝率関数が既知であり，さらに許される戦略が定常ランダム戦略にかぎる，すなわち常に一定の確率 r でロブを打つ戦略であるとする．すると，漸近的にロブ比率 s は r に収束するので，勝率は漸近的に，$w(r)=r\cdot f_L(r)+(1-r)\cdot f_P(r)$ で与えられる．$w(r)$ は r に関する2次式であるから最大値が存在し，それを与える $r=r^*$ が最適なロブ比率であることが導かれる．また，$f_L(0)=f_P(1)$（「Matching Shoulders 条件」[1]）が成り立つ場合には，ロブ比率 r の定常ランダム戦略がすべての戦略の中で最適な戦略となる[*3]．（加えて最適比率が f_L と f_P の交点で与えられる．）これにより，任意の戦略の regret の評価も最適な定常ランダム戦略との比較をすることで行うことができる．ここでは「Matching Shoulders 条件」が成り立っていることを仮定して話をすすめることにする．

3.4 ロブ・パス戦略

まず考えられる戦略としては，試行を繰り返すうちに r^* の推定値を計算し，ロブとパスをうまく打ち分けることによりなるべく早く推定された最適値に到達する，という戦略である．しかし，ここで問題になるのが，最適ロブ比率の推定にはロブ比率を変化させた方がよいが，勝率を高めるには最適なロブ比率からあまり離れない方がよいというトレードオフである．文献1）で提案されている Arthur[*4] は，このトレードオフの具体的な解消法を提示している．Arthur の行動は，推定と移動からなるステージからなる．各ステージにおいて，試行数が過去の試行数と同じオーダになるように，それまでの全ステージでの試行数と同じ数の試行を行いながら最適ロブ比率を推定/接近する．あるステージのはじめにロブ比率 r にいるとすると，ロブ，パスを適当に打ち分けて，ロブ比率 $r+\Delta$ に移動し，$r, r+\Delta$ 2点でのロブとパス各々の勝率を推定する．ロブ，パス勝率関数はともに線形であるので2点における推定値をもとに全関数を推定することができ，またその交点として最適比率も推定することができる．そして最適比率の推定値が計算できたら，ショッ

[*3] 「Matching Shoulders 条件」が成り立たない場合には，驚くべきことに必ずしもそうならない．
[*4] 往年の名テニス選手の Arthur Ashe から名づけられた．

トを適当に打つことによりロブ比率をその推定値に合わせて，次のステージに進む．Arthurの各ステージでのテスト幅Δを$O(t^{1/4})$に設定することによりトレードオフを解決して，期待regretを$O(\sqrt{l})$に抑えることができることが（ある緩やかな条件のもとで）示されている．

上記のArthurについては，環境パラメータに関して成り立つ仮定によってテスト幅を変えることにより，さまざまな異なる期待regretの上限（たとえば，$O(\log t)$，$O(t^{\frac{1}{3}})$，$O(t^{\frac{1}{2}})$，$O(t^{\frac{3}{5}})$，$O(t^{\frac{2}{3}})$，$O(t^{\frac{5}{7}})$など）を導くことができることが示されている[1]．しかし，$O(\log t)$の上限を導くには，「勝率関数の傾き（の絶対値）の和が既知である」などといった強い仮定が必要であった．その後，より弱い仮定のもとで$O(\log t)$の上限を達成する戦略が複数提案されている[16),19)]．ここでは，文献19）で提案されているChris[*5]という方法を簡単に紹介しよう．Chrisは「確率的近似法」に基づく以下のような戦略である．Chrisは推定ロブ比率rをパラメータとして保持し，それを更新しながら学習/行動する．推定ロブ比率の現在値をrとすると，Δをある小さい正の定数として以下の2つの試行を行う．（i）確率$r-\Delta$でロブを打つ．（ii）確率$r+\Delta$でロブを打つ．そしてもし（i）で勝ち（ii）で負けたならば，rの値を減る方向に微量修正し，逆に（i）で負け（ii）で勝ったならば，rの値を微量だけ増やす．それ以外の場合はrは変更しない．以上のことを際限なく繰り返すのである．r^*の値と勝率関数の傾斜の差aがある仮定を満たす時，Chrisの期待regretが$O(\log t)$で抑えられることが示せる[19)]．

上記のロブ・パス問題の定式化においてとくに非現実的であると思われる点に，勝率関数の線形関数による表現があげられる．この問題に対処するために，平岡ら[12)]は成功確率関数がパラメータにより指定される関数ではなくて，単に単調増加（減少）な凸関数であるという仮定のもとに解析した．彼らは，やはり「確率近似法」に基づく戦略により，期待regretの上限$O((\log t)^2)$を達成できることを示している．

[*5] Chris Evertから名づけられた．

最適選択の逐次型学習モデル内における以上のような結果を通して，ロブ・パス問題は知識の獲得と利用の最適な解消に関する解析を許す単純だが興味深い構造をもったマルコフ決定プロセスの部分クラスであることがわかる．今後は，マルコフ決定プロセス一般，もしくはより広い部分クラスに関して同様な解析がなされ，またその過程で効率的かつ期待regretの意味で高性能を有した戦略が発見されることが期待される．

4. おわりに

能動学習の理論的解析の枠組みとして，「質問学習モデル」と「最適選択の逐次型学習モデル」に関する研究成果について解説した．今後，学習理論の研究成果が機械学習の実際の応用の場面で活かされていくこと，また機械学習の応用の観点からこれらのモデルがさらに発展していくことが期待される．

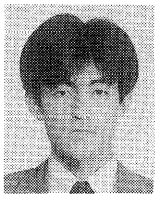
参考文献

1) Abe, N. and Takeuchi, J.: The 'Lob-pass' Problem and an On-line Learning Model of Rational Choice, Proc. of the 6 th Annual ACM Conference on Computational Learning Theory, pp. 422-428 (1993).
2) Angluin, D.: Learning Regular Sets from Queries and Counterexamples, Information and Computation 75, pp. 87-106 (1987).
3) Angluin, D.: Queries and Concept Learning, Machine Learning 2, pp. 319-342 (1988).
4) Angluin, D., Frazier, M. and Pitt, L.: Learning Conjunctions of Horn Clauses, Machine Learning 9, pp. 147-164 (1992).
5) Angluin, D., Hellerstein, L. and Karpinski, M.: Learning Read Once Formulas with Queries, Journal of the ACM 40 (1), pp. 185-210 (1993).
6) Angluin, D. and Kharitonov, M.: When Won't Membership Queries Help?, Journal of Computer and System Science 50, pp. 336-355 (1995).
7) 有川節夫, 西野哲朗他：(特集)計算論的学習理論とその応用, 情報処理, Vol. 32, No. 3, pp. 217-281 (Mar. 1991).
8) Aizenstein, H. and Pitt, L.: Exact Learning of Read-Twice DNF Formulas, In Proceedings of the 32 th Annual IEEE Symposium on Foundations of Computer Science, pp. 170-179 (1991).
9) Bshouty, N. H.: Exact Learning Boolean Functions via Monotone Theory, Information and Computation 123, pp. 146-153 (1995).

10) Hancock, T. R. : Learning 2μDNF Formulas and $k\mu$ Decision Trees, In Proceedings of the 4 th Annual Workshop on Computational Learning Theory, pp. 199-209 (1991).
11) Herrnstein, R. J.: Rational Choice Theory, American Psychologist, 45 (3), pp. 356-367 (1990).
12) Hiraoka, K. and Amari, S.: Stochastic Game under Unknown Stochastic Environment—Nonparametric Lob-pass Problem, Submitted (1996).
13) Jackson, J.: An Efficient Membership-Query Algorithm for Learning DNF with Respect to the Uniform Distribution, In Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science, pp. 42-53 (1994).
14) Kearns, M., Li, M., Pitt, L. and Valiant, L.: On the Learnability of Boolean Formulae, Proc. of the 19th ACM Symposium on the Theory of Computing, pp. 285 296 (1987).
15) Kearns, M. and Valiant, L.: Cryptographic Limitations on Learning Boolean Formulae and Finite Automata, Journal of the ACM 41 (1), pp. 67-95 (1994).
16) Kilian, J., Lang, K. J. and Perlmutter, B.: Playing the Matching-shoulders Lob-pass Game with Logarithmic Regret, Proc. of the 7th Annual ACM Conference on Computational Learning Theory, pp. 159-164 (1994).
17) Pillaipakkamnatt, K. and Raghavan, V.: Read-Twice DNF Formulas Are Properly Learnable, Information and Computation 122, pp. 236-267 (1995).
18) Pitt, L. and Warmuth, M. K.: Prediction-Preserving Reducibility, Journal of Computer and System Science 41, pp. 430-467 (1990).
19) Takeuchi, J., Abe, N. and Amari, S.: The Lob-pass Problem, Submitted (1997).
20) Valiant, L. G.: A Theory of the Learnable, Communications of the ACM 27, pp. 1134-1142 (1984).
21) Watkins and Dayan, P.: Q-learning, Machine Learning 8, pp. 279-292 (1992).

(平成9年5月8日受付)

安倍 直樹 (正会員)
1960年生. 1984年マサチューセッツ工科大学情報科学科学士/修士課程修了. 同年IBMワトソン研究所にて研究員. 1989年ペンシルバニア大学情報科学科博士課程修了. 同年カリフォルニア大学サンタクルーズ校にて研究員. 1990年NEC入社, 現在C&C研究所研究専門課長. 計算論的学習理論およびその応用研究に従事. Ph. D.
e-mail: abe@sbl. cl. nec. co. jp



中村 篤祥 (正会員)
1963年生. 1986年東京工業大学理学部情報科学科卒業. 1988年同大学院理工学研究科修士課程修了. 同年NEC入社. 現在, 同社C&C研究所に勤務. 計算論的学習理論の研究に従事. e-mail: atsu@sbl. cl. nec. co. jp