# クラス分類問題の強化学習による解釈

大庭隆伸　南泰浩
日本電信電話株式会社 コミュニケーション科学基礎研究所

## 1 はじめに

強化学習は報酬期待値を最大化する方策関数の推定問題である．方策関数とは，現在の状態に応じた行動を返す関数である．一方，クラス分類問題は入力に対応する適切なクラスを有限のクラス集合の中から選択する問題である．今，行動を離散的かつ有限とする強化学習を考えた場合，その方策推定も，クラス分類関数の推定も，有限のシンボル集合の中から適切なシンボルを選択する問題という意味において共通であり，強化学習とクラス分類問題の類似性が窺える．我々の研究の目的は，数式上での両者の関係性を示すことにある．本稿では，クラス分類問題を強化学習上で定義することを試みる．

## 2 強化学習

状態および行動の集合を $S, A$ と表記する．ある方策 $\pi : S \to A$ が与えられた下で，時刻 $t = 0$ における状態 $s_0 = s \in S$ から，その方策に従って行動を生成し，将来に渡って得られる報酬の期待値 (状態価値関数) を次式で定義する．

$$V^\pi(s) = E[\sum_{t=0}^{\infty} g(t) r_{t+1} | s_0 = s, \pi] \quad (1)$$

ただし，$g(t)$ は収束を得るための関数．最も典型的な強化学習の定義は，これを最大化する方策を決定することである [1]．すなわち，

$$\pi^* = \arg\max_\pi V^\pi(s) \quad {}^\forall s \in S \quad (2)$$

これに対し次のような定式化も可能である．

$$\pi^* = \arg\max_\pi \sum_{s \in S} p^\pi(s) V^\pi(s) \quad (3)$$

この定式化は真の状態価値関数との最小二乗誤差に基づく強化学習問題の解法 [2] と密接に関連している．実際，最小二乗誤差法では状態価値関数の上限値 (真の値) との残差を最小化するのに対し，上記定式化では単純にその上限値を与える方策を探索する．

ところで，式 (2) を満たす方策が存在し，かつ状態出現確率が方策に依存しない，すなわち $p^\pi(s) = p(s)$ のとき，次式が成り立つ．

$$\begin{aligned} \pi^* &= \arg\max_\pi \sum_{s \in S} p(s) V^\pi(s) && (4) \\ &= \arg\max_\pi p(s) V^\pi(s) \quad {}^\forall s \in S && (5) \\ &= \arg\max_\pi V^\pi(s) \quad {}^\forall s \in S && (6) \\ &= \pi^* && (7) \end{aligned}$$

このことから，式 (2) と式 (3) の各定義に基づく強化学習は同一の方策を与えることがわかる．ただし，$p(s)^\pi \neq p(s)$ の場合でも，同一の方策が得られる場合もある．

## 3 クラス分類問題

入力集合を $X$，クラス集合を $Y$ とする．学習の目的はクラス分類関数 $f : X \to Y$ の推定である．しかし一般には $f$ を支配するパラメータ $\mathbf{A}$ を導入し，目的関数 $\sum_{(x,y) \in D} m_{\mathbf{A}}(x, y)$ を最大化するパラメータを求める問題に置換する．すなわち，

$$\mathbf{A}^\star = \arg\max_{\mathbf{A}} \sum_{(x,y) \in D} m_{\mathbf{A}}(x, y) \quad (8)$$

なお，$D$ は標本集合である．$m_{\mathbf{A}}(x, y)$ の定義により，分類手法の定義され，例えば

**[Maximum Entropy Approach [3]]**

$$m_{\mathbf{A}}(x, y) = \frac{\exp(\mathbf{A}^\top \mathbf{F}(x, y))}{\sum_{y'} \exp(\mathbf{A}^\top \mathbf{F}(x, y'))} \quad (9)$$

**[サポートベクタマシン** ($Y = \{-1, 1\}$) [4]]

$$m_{\mathbf{A}}(x, y) = -\max(1 - y\mathbf{A}^\top \mathbf{F}(x), 0) - c||\mathbf{A}|| \quad (10)$$

**[ブースティング** [5]]

$$m_{\mathbf{A}}(x, y) = -\sum_{y'} \exp(\mathbf{A}^\top \mathbf{F}(x, y') - \mathbf{A}^\top \mathbf{F}(x, y)) \quad (11)$$

**Clustering problem on Reinforcement Learning**
Takanobu OBA and Yasuhiro MINAMI
NTT Communication Science Laboratories, NTT Corporation.
619-0237, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{oba, minami}@cslab.kecl.ntt.co.jp

などが用いられる．ただし，$\mathbf{F}$ は素性ベクトル，$c$ は正則化定数，$\|\cdot\|$ はノルム．

クラス分類は推定されたパラメータ $\mathbf{A}^\star$ を使用し，$\arg\max_{y\in Y} m_{\mathbf{A}^\star}(x,y)$ により行われる．

## 4 強化学習としてのクラス分類問題

今，関数 $e_f(x)$ を $\sum_{(x,y)\in D} m_{\mathbf{A}}(x,y) = \sum_{x\in X} p(x)e_f(x)$ により定義する．このときクラス分類関数の学習に関わる式 (8) は次式と等価である．

$$f^\star = \arg\max_f \sum_{x\in X} p(x)e_f(x) \tag{12}$$

また以下の関係が成り立つ．

$$\sum_{x\in X} p(x)e_f(x) = E[e|f] \tag{13}$$

$$= E[\sum_{t=0}^{\infty} \delta(t)e_{t+1}|f] \tag{14}$$

$$= \sum_{x\in X} p^f(x)V^f_{\delta(t)}(x) \tag{15}$$

ただし，$\delta(t)$ は離散デルタ関数．また，

$$V^f_{\delta(t)}(x) = E[\sum_{t=0}^{\infty} \delta(t)e_{t+1}|x_0 = x, f] \tag{16}$$

式 (12) および (15) から，クラス分類関数の推定が式 (3) で定義された強化学習に一致することがわかる．具体的には，$S = X$, $\pi = f$, $g(t) = \delta(t)$, $r = e$ の下での強化学習と位置づけられる．$S = X$, $\pi = f$ から $A = Y$ である．

$g(t) = \delta(t)$ により，クラス分類問題では時間発展を考慮していないことがわかる．このとき $p^f(x) = p(x)$ である．よって式 (15) から $e_f(x) = V^f_{\delta(t)}(x)$ という解釈が可能となる．また，もし，

$$\arg\max_f V^f_{\delta(t)}(x) \quad {}^\forall x \in X \tag{17}$$

を満たす解 $f^*$ が存在するならば，強化学習の定義式 (式 (3)) を式 (4) に変形する条件に一致するので，クラス分類関数の推定問題の定義式 (式 (12)) に関しても式 (4) から式 (7) と同一の変形を得る．つまりクラス分類関数の推定問題を式 (2) で定義される強化学習上で表現できることを示す．

$r = e$ であるが，$e$ は説明用の記号であり，クラス分類の学習自体には使用しない．そこで次に $m_{\mathbf{A}}(x,y)$ について考える．強化学習分野でQ関数と呼ばれる関数は時間発展のない環境において，

$$Q^\pi(s,a) = E[\sum_{t=0}^{\infty} \delta(t)r_{t+1}|s_0 = s, a_0 = a, \pi] \tag{18}$$

$$= E[r|s,a,\pi] = E[r|s,a] \tag{19}$$

$$= Q(s,a) \tag{20}$$

すなわち，即時報酬の期待値に一致し，将来に渡り使用する方策も無用となる．Q関数の形，それ自身が報酬期待に一致すること，また，方策を $\arg\max_{a\in A} Q(s,a)$ により設計し得ることを考えると，$m_{\mathbf{A}}(x,y)$ をQ関数と考えるのが妥当であろう．

ただ，$m_{\mathbf{A}}(x,y)$ には一般に他のクラスラベルも考慮した関数が用いられる．また，通常 $Q^{\pi'}(s,\pi(a)) \neq V^\pi(s)$ であるが，時間発展がない環境では $Q(s,\pi(s)) = V^\pi(s)$ と，ある固定方策の下で両者が同一となる．それらの影響により，クラス分類が $Q(s,\pi(s))$, $V^\pi(s)$ すなわち $m(x,y)$, $e(s)$ のどちらを使用した学習であるのか曖昧となり，その解釈には議論の余地がある．

## 5 まとめ

本稿では強化学習とクラス分類問題の類似性に着目し，クラス分類関数の推定問題を強化学習上で定義した．基本的な両者の相違は時間発展の有無である．また，強化報酬が即時報酬を獲得しながら学習を行うのに対し，クラス分類では報酬期待値にあたる関数を設計した上で学習を行う．ただ，この関数には複数のラベルを考慮した関数の使用が許されるので，一般的にひとつの行動を起してひとつの即時報酬を受け取る強化学習の枠組み上での解釈を難しくしている．

## 参考文献


[1] 三上貞芳, 皆川雅章 : 強化学習, 森北出版株式会社 (2000).

[2] Schoknecht, R.: Optimality of reinforcement learning algorithms with linear function approximation, in *Advances of Neural Information Processing Systems*, Vol. 15, pp. 1555–1562 (2003).

[3] Berger, A., Pietra, V. D. and Pietra, S. A. D.: A Maximum Entropy Approach to Natural Language Processing, *Computer Linguistics*, Vol. 22, No. 1, pp. 39–71 (1996).

[4] Joachims, T.: A Support Vector Method for Multivariate Performance Measures, in *Proceedings of Machine learning*, pp. 377–384 (2005).

[5] Freund, Y. and Schapire, R. E.: A Decision-Theoretic Generalization of On-line Learning and An Application to Boosting, *Journal of Computer and System Sciences*, Vol. 55, (1997).