数理解析研究所講究録
第899巻 1995年 1-6

An algorithm for multiobjective Markov decision processes:Discounted reward case

長岡工業高等専門学校　涌田和芳（Ｋａｚｕｙｏｓｈｉ　Ｗａｋｕｔａ）

1.はじめに

多目的マルコフ決定過程　（ＭＭＤＰ）　のためのアルゴリズムは，すでにいくつか提案されている．Ｖｉｓｗａｎａｔｈａｎ　ｅｔ　ａｌ．［5］と Ｎｏｖａｋ［3］は，ＭＭＤＰを多目的ＬＰ問題として定式化し，これを解いた．（Ｎｏｖａｋ[3]は，完全エルゴード過程の平均利得型も扱っている）．この方法により得られる非支配解の集合は，本論で述べる最適解の集合より一般に小さい．

Ｗｈｉｔｅ ＆ Ｋｉｍ［7］は，ＭＭＤＰが特殊な構造をもつ部分的観測ＭＤＰであることを述べ，部分的観測ＭＤＰのためのアルゴリズムにもとづいた解法を示した．

Ｆｕｒｕｋａｗａ[1]は，割引利得型ＭＭＤＰのための政策反復法を考案した．Ｔｈｏｍａｓ[4]は，Ｆｕｒｕｋａｗａ[1]の方法を平均利得型に応用した．Ｔｈｏｍａｓ[4]は完全エルゴード過程の場合を考えたが，Ｉｋｉ ＆ Ｆｕｒｕｋａｗａ[2]は多重連鎖過程の場合の政策反復法について議論した．しかし，これらの政策反復法は，確定的定常政策の中だけでしか議論されていない．Ｎｏｖａｋ［3］によれば，確率的政策を考慮したときに比べ数倍の解が得られる例がある．

本論では，割引利得型ＭＭＤＰのための政策反復法について議論する．最近著者は，確率的で，システムの履歴を記憶するすべての政策の中で最適な確定的定常政策を求める政策反復法を与えた([6])．そこでは，まず政策反復によって最適である可能性のある候補を絞り，次に最適政策を特徴付ける線形不等式系をフーリエ消去法で解いて最適性の判定をした．しかし，フーリエ消去法は変数を消去する際に多くの不等式が生ずることがあるため，制限された形では適用可能であるが，大きなモデルでは実行が困難である．そこで本論では，線形不等式系を解くときフーリエ消去法の代わりにＬＰ法を適用した解法を述べる．

2．多目的マルコフ決定過程

$a=(a_1,\ldots,a_m), b=(b_1,\ldots,b_m)\in R^m$ に対して

$a \underline{\geq} b \Leftrightarrow a_k \underline{\geq} b_k, k=1,\ldots,m$

$a \geq b \Leftrightarrow a \underline{\geq} b, a \neq b$

$a > b \Leftrightarrow a_k > b_k, k=1,\ldots,m$

$U \subset R^m$に対して

$e(u)=\{x\in U \mid x \underline{\leq} y \text{ for some } y\in U \text{ implies } x=y\}$

多目的マルコフ決定過程

$S=\{1,\ldots,N\}$ ：状態空間

A =有限集合：行動空間，$A(i): i\in S$ で実行可能な行動集合

$GrA=\{(i,a)\,|\,i\in S, a\in A(i)\}$

$p(j|i,a), i,j\in S, a\in A(i)$ ：推移確率

$r(i,a)=(r^1(i,a),\ldots,r^m(i,a))\in R^m$ ：利得関数

β ($0\leqq\beta<1$) ： 割引因子

Π： すべての政策の集合

Π_D： すべての確定的定常政策の集合

$$I_\pi(i_1)=E_\pi\left[\sum_{n=1}^{\infty}\beta^{n-1}r(i_n,a_n)\ \middle|\ i_1\right]$$

$$V_\pi(i_1)=\bigcup_{\pi\in\Pi} I_\pi(i_1)$$

すべての $i_1\in S$ に対して $I_{\pi^*}(i_1)\in e(V(i_1))$ であるとき，π^* は最適であるという

3．最適性の判定

$f\in\Pi_D$ に対して

$$v(i)=r(i,f(i))+\beta\sum_{j\in S}p(j|i,f(i))v(j), i\in S$$

を解いて，$v(i)=I_f(i), i\in S$ を求める．そして，$1\times m$ 行ベクトル

$$q_f(i,a)=r(i,a)+\beta\sum_{j\in S}p(j|i,a)I_f(j)-I_f(i), i\in S, a\in A(i)$$

を用いて，次の3つの行列を定める．

Q_f :$q_f(i,a),(i,a)\in GrA$ を行ベクトルにもつ行列

$\overline{Q}_f(i_1)$:$q_f(i_n,a_n),(i_n,a_n)\in GrA, p_f(i_1)\{i_n\}>0$

を行ベクトルにもつ行列

$\overline{\overline{Q}}_f(i_1)$:$q_f(i_n,a_n),(i_n,a_n)\in GrA, p_\pi(i_1)\{i_n\}>0, \pi\in\Pi$

を行ベクトルにもつ行列

Theorem 3. 1.

(i) $f \in \Pi_D$ が最適ならば，次の各線形不等式系

$$(S_1):\begin{cases} x>0 \\ \overline{Q}_f(1)x \leqq 0 \end{cases}, \cdots\cdots, (S_N):\begin{cases} x>0 \\ \overline{Q}_f(N)x \leqq 0 \end{cases}$$

は解をもつ.

(ii) 各線形不等式系

$$(T_1):\begin{cases} x>0 \\ \overline{\overline{Q}}_f(1)x \leqq 0 \end{cases}, \cdots\cdots, (T_N):\begin{cases} x>0 \\ \overline{\overline{Q}}_f(N)x \leqq 0 \end{cases}$$

が解をもてば，$f \in \Pi_D$ は最適である.

Assumption 3. 1. 任意の確定的定常政策により生ずるマルコフ連鎖において，すべての状態はひとつのエルゴード集合に属する.

Corollary 3. 1. Assumption 3. 1を仮定する. このとき，$f \in \Pi_D$ が最適であることは線形不等式系

$$(S_0):\begin{cases} x>0 \\ Q_f x \leqq 0 \end{cases}$$

が解をもつことと同値である.

$1 \times m$ 行ベクトル

$$d_f^g(i_1) = I_g(i_1) - I_f(i_1)$$

を用いて，$d_f^g(i_1), g \in \Pi'_D$ を行ベクトルにもつ行列 $D_f(i_1)$ を定める. ただし，Π'_D は，最適である可能性のある政策の集合である.

Theorem 3. 2. $f \in \Pi_D$ が最適であることは，$f \in \Pi'_D$ で，かつ各線形不等式系

$$(D_1):\begin{cases} x>0 \\ D_f(1)x \leqq 0 \end{cases}, \cdots\cdots, (D_N):\begin{cases} x>0 \\ D_f(N)x \leqq 0 \end{cases}$$

が解をもつことと同値である．

Theorems 3. 1, 3. 2, Corollary 3. 1の各線形不等式系

$$(S):\begin{cases} x>0 \\ Bx \leqq 0 \end{cases}$$

が解をもつかどうかという問題は，次のようにLP 問題として定式化できる．

$$p(S):\ Maximize\ \ v=y$$
$$subject\ to$$
$$\begin{cases} x_1 \geqq y, \cdots\cdots, x_m \geqq y \\ Bx \leqq 0 \\ x_1 \geqq 0, \cdots\cdots, x_m \geqq 0, y \geqq 0 \end{cases}$$

Theorem 3. 3. (S)は，Theorems 3. 1, 3. 2, Corollary 3. 1 の任意の線形不等式系とする．

(i) $P(S)$の最大値が正であるとき，またその時に限り(S)は解をもつ．このとき，$P(S)$は非有界である．

(ii) $P(S)$の最大値がゼロのとき，またその時に限り(S)は解をもたない．

$P(S)$ はスラック変数を導入して実行可能基底形式をつくるので，容易に解くことができる．

4．政策反復法

S上のm−値関数 u,v について

$$u \geqq v \Leftrightarrow u(i) \geqq v(i), i \in S$$
$$u \geq v \Leftrightarrow u \geqq v, u \neq v$$

任意の$f,g\in\Pi_D$ に対して，S上のm−値関数を

$$I_f^g(i)=r(i,g(i))+\beta\sum_{j\in S}p(j|i,g(i))I_f(j),i\in S$$

で定義する．ここで，特に $I_f^f(i)=I_f(i),i\in S$である．

Lemma 4. 1.

(*i*) $I_f^g-I_f^f\geq 0\Rightarrow I_g\geq I_f$
(*ii*) $I_f^g-I_f^f\leq 0\Rightarrow I_g\leq I_f$
(*iii*) $I_f^g-I_f^f=0\Rightarrow I_g=I_f$

アルゴリズムの概要

Ｐｈａｓｅ Ｉ． Ｌｅｍｍａ ４．１に基づき政策反復を行って，最適である可能性のある政策の集合と最適でない政策の集合に分類する．

Ｐｈａｓｅ ＩＩ．Ｔｈｅｏｒｅｍ ３．２に基づきＬＰ問題を解いて，最適である政策の集合に属する政策が最適かどうか判定する．

Ｒｅｍａｒｋ ４．１．$g(i)=a$ のとき，$q_f(i,a)=I_f^g(i)-I_f^f(i)$ なので，Ｐｈａｓｅ Ｉのデータを Ｐｈａｓｅ ＩＩで利用できる．

References

[1] N. Furukawa, Vector-valued Markovian decision processes with countable state space, in : R. Hartley et al. Ed.,Recent Developments in Markov Decision Processes (Academic Press, New York, 1980) pp. 205-223.

[2] T. Iki & N. Furukawa, Vector-valued Markov decision processes with average criterion, Memoirs of the Faculty of Education, Miyazaki University, 54・55 (1984) 1-10

[3] J. Novak, Linear programming in vector criterion Markov and semi-Markov decision processes, Optimization 20 (1989) 651-670.

[4] L.C. Thomas, Constrained Markov decision processes as multi-objective problems, in: S. Fench et al. Ed., Multi-Objective Decision Making (Academic Press, New York, 1983) pp. 77-94.

[5] B. Viswanathan et al., Multiple criteria Markov decision processes, TIMS Studies in the Management Sciences 6 (1977) 263-272.

[6] K. Wakuta, Vector-valued Markov decision processes and the systems of linear inequalities, To appear in Stochastic Processes and its Application (1995).

[7] C.C. White & K.W. Kim, Solution procedures for vector criterion Markov decision processes, Large Scale Systems 1 (1980) 129-140.