ファジィ理論を用いた自動装置の制御技術に関する研究

シルベスター・コバチ　ピーター・バラーニ　浅井　博次

Study on Control Techniques by Fuzzy Reasoning in Automata

Szilveszter Kovács　　Péter Baranyi　Hirotsugu Asai*

あらまし　多機能な制御システムの構築に必要な制御規則獲得に有効な強化学習法に着目し，効率的な制御規則獲得のために問題となる計算コストを削減する手法として，SVD(Singular Value Decomposition；特異値分解)に基づいて代数積・加算重心(PSG ; potential based guiding)法に基づいたファジィ規則を近似する技法を適応することを提案した．

キーワード　強化学習，特異値分解，potential based guiding model

１．　まえがき

近年，さまざまな分野で使用されている自動装置は高性能化や多機能化が図られている．その反面，機能向上に伴い制御システムが複雑化し，計算機に求められる能力は急激に増加している．また，自動装置の適用環境についても拡大への要求が増大している．つまり，限られた環境で開発・実行できる多機能システムが要求されている．そのため，少ない計算コストで多機能な（複雑な）システム構築が可能な手法について検討することは非常に意味がある．

複雑な制御を実行するための制御規則を獲得する方法として強化学習法がある．強化学習法はタスクを達成したら“報酬”を与えるように設定するだけで複雑な制御規則を自動的に取得できるため，未知環境での制御など自律ロボット工学の世界で近年ますます注目を集めている．しかし，複雑な状況で強化学習法を応用する場合には，１つ問題がある．状態評価関数あるいは行動評価関数の表現が巨大になってしまうことである[1]．連続的な環境(連続的に評価される)場合，状態評価関数あるいは行動評価関数が連続関数になるため，強化学習はさらに複雑になってしまう．

そこで，本研究では，多機能な制御システムの構築に必要な制御規則獲得に有効な強化学習法に着目し，効率的な制御規則獲得のために問題となる計算コストを削減する手法として，SVD(Singular Value Decomposition；特異値分解)を応用した方法を提案する．

２．　導入

強化学習法の一般的な目標は，最適な政策を発見することであり，ほとんどの場合状態評価関数あるいは行動評価関数を構築することによって行われる[1]．

状態評価関数 $V^p(s)$ は出発点として与えられた状態 $s \in S$ と関連付けられ，与えられた政策p に従う場合の予測される結果の関数である．行動評価関数$Q^p(s,a)$ は与えられた状態 s で行動$a \in A_s$ をとり，与えられた政策p に従う場合の予測される結果の関数である． 行動評価関数により，最適な政策は次式で与えられる[1]．

$$\boldsymbol{p}(s) = \arg\max_{a \in A_s} Q^p(s,a) \qquad (1)$$

すなわち最適政策を評価するために行動評価関数 $Q^p(s,a)$ を近似する必要がある．離散的な（状態・行動が離散的に定義される）環境において，少なくとも $\sum_{s \in S} \|A_s\|$ の要素を扱われなくてはならないことを意味している(ここで $\|A_s\|$ は状態 s での可能な行動の集合数)．複雑なタスクを適応する場合，可能な状態数と可能な行動数の両方が非常に大きい値になる．連続的な（状態・行動が連続的に定義される）環境で強化学習を実行するために関数近似法が広く使用されている．これらの方法の多くは，離散的な場合と同様に連続的な状態・行動空間を扱うために戦略分割を適用している．適当な分割構造を構築する場合の問題点の１つは，行動評価関数がわからないことである．細かく分割すると状態数が大きくなる一方，粗く分割すると不正確で適応能力のないシステムになってしまう．扱う状態数の増加は計算コストの増加に繋がる．それは，多くの実時間アプリケーシ

ョンでは，許容できないことである．こういった問題の簡単な解決法として，連続的環境における強化学習アプリケーションの高いストレージ・コストと計算コストに対処するため，文献[2,3]で Yam と Baranyi によって提案された SVD に基づいて代数積・加算重心(PSG; potential based guiding)法に基づいたファジィ規則を近似する技法を適応することを提案する．

3．SVD に基づいた PSG 近似

複雑さの削減において SVD を使用する鍵となる考え方は，“特異値が与えられたシステムを分解するために適用でき，分解されたパーツの重要度を示す”というものである．割り当てられた特異値に応じて出力に，あまり，または全く寄与しない部分を切り取ることで削減を実施できる．SVD に基づく削減の初期研究で計算の複雑さと近似誤差の関係が述べられている[2]．シングルトンに基づく代数積・加算重心(PSG)近似は，入力空間 X_n 上で定義された前件ファジィ集合 $m_{j_n,n}(x_n)$ によって与えられる N 変数ファジィ・ルールベースを仮定している．前件部のすべての組み合わせは出力空間 Y 上で定義されたひとつの後件部ファジィ集合に対応する．これらの関係は規則によって以下のように表現される．

If $m_{j_1,1}(x_1)$ **And** $m_{j_2,n}(x_2)$ **And** … **And** $m_{j_N,N}(x_N)$ **Then** $b_{j_1 j_2 \cdots j_N}$

シングルトン後件ファジィ集合 $b_{j_1 j_2 \cdots j_N}$ はその位置毎に $b_{j_1 j_2 \cdots j_N}$ で定義される．シングルトンに基づく PSG 近似の通常の手法は以下の式で表される[2,3].

$$f = \sum_{j_1,j_2,\cdots,j_N}^{J_1,J_2,\ldots,J_N} \prod_{n=1}^{N} m_{j_n,n}(x_n) b_{j_1 j_2 \cdots j_N} \qquad (2)$$

文献[2,3]で紹介された SVD に基づくファジィ・ルールベース削減によって，(2)式は次式に変換できる．

$$\tilde{f} = \sum_{j_1,j_2,\ldots,j_N}^{J_1^r,J_2^r,\ldots,J_3^r} \prod_{n=1}^{N} m^r_{j_n,n}(x_n) b^r_{j_1 j_2 \cdots j_N} \qquad (3)$$

ここで，削減の結果として必然的に $\forall n : J_n^r \le J_n$ となる．本手法を今後 SVD 削減と呼ぶこととする．

強化学習に SVD 削減を適応する我々の場合，ゴールは基本的に未知の関数 $y = f(x_1, x_2, \ldots, x_n)$ を近似することであると仮定する．適当な強化学習法を使用し，可能な限り多くの近似点を収集する．すなわち，近似 $\tilde{y} = f^a(x_1, x_2, \ldots, x_n)$ において P 個の点が使用されると仮定する．近似点数が増加するとその計算によって有効な計算能力 C は急激に食いつぶされる．そこで，文献[2,3]で提案された SVD 削減を適用することにより与えられた誤差の閾値に従い，計算の複雑さを $D \le C$ に減少させることが，本論文の目的である．これにより，$C - D$ の計算能力を開放できる．開放された計算能力を使用して，更に近似点数を増やすことによって，誤差を改善することができる．

3． 応用例

強化学習に SVD に基づく PSG ファジィ・ルールベース近似手法を適用する方法を紹介するために，2つの学習手法を選択した．1つ目が代表的な強化学習法 Q-Learning [4]，2つ目が Fuzzy Prioritised Sweeping 法[5]に基づいたモデルである．離散的な環境での Q-Learning では，行動評価関数は次式を繰り返すことで近似される．

$$Q_{i,u} \approx \tilde{Q}^{k+1}_{i,u} = \tilde{Q}^{k}_{i,u} + g^{k}_{i,u} \cdot \left(g_{i,u,j} + \alpha \cdot \max_{v \in U} \tilde{Q}^{k}_{j,u} - \tilde{Q}^{k}_{i,u} \right)$$

$$\forall i \in I, \forall u \in U \qquad (4)$$

ここで $\tilde{Q}^{k+1}_{i,u}$ は状態 S_i において実行した行動 A_u の行動評価値を $k+1$ 回繰り返し更新した値である．α は割引率，$g^{k}_{i,u} \in [0,1]$ は学習間隔を決定する学習率である．連続的な環境における Q-Learning に移行するためには，次式に示すように PSG 近似で $\tilde{Q}(s,a)$ を近似することによって連続行動評価関数 $Q(s,a)$ を近似することができる．

If s **is** S_i **And** a **is** A_u **Then** $\tilde{Q}(s,a) = Q_{i,u}$ (5)

$$i \in I, u \in U$$

そして，関数(1)によって最適政策を構築し，文献[5]で提案されている発見的教授法によって連続的政策へと変更できる．

$$\tilde{p}(i) = \arg\max_{u \in U} Q(i,u) \qquad (6)$$

If s **is** S_i **Then** $a = \tilde{a}_{\tilde{p}(i)}, \; i \in I$ (7)

PSG 近似によって$\tilde{Q}(s,a)$を近似する場合，前節で提案したように SVD 削減法を適用できる行動評価関数のすべての評価値をゼロに設定し，行動を適当な探索政策によって選択する．状態空間と行動空間の分割は三角形状のファジィ集合形式で状態評価値と行動評価値から得 Ruspini ファジィ分割として構築される(fig.1).

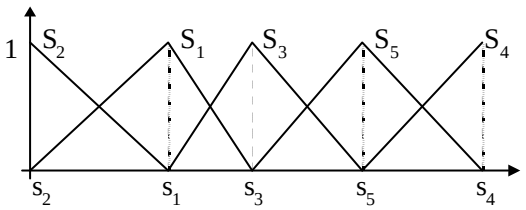

Fig.1. 状態評価値から得られた状態分割例

2つ目の例は，文献[5]で提案された Fuzzy Prioritised Sweeping 法に基づくモデルである．ここで連続行動評価関数$Q(s,a)$は(5)式の PSG 近似で$\tilde{Q}(s,a)$を近似することによって近似される．
文献[5]では，行動評価関数は次式の the Bellman equation [6]で近似されると言われている．

$$Q_{i,u} \approx \tilde{Q}_{i,u} = \sum_{j\in I} \tilde{p}_{i,j}(u)\cdot\left(\tilde{g}_{i,u,j} + \alpha \cdot \max_{v\in U} Q_{j,v}\right) \qquad (8)$$

$$\forall i \in I, \forall u \in U$$

ここで，$p_{i,j}$は行動A_uの場合における状態遷移$S_i \to S_j$の可能性見積もり値，$g_{i,u,j}$は対応する平均報酬の見積もりである．状態遷移の可能性見積もり$\tilde{p}_{i,j}(u)$と対応する平均報酬の見積もり$\tilde{g}_{i,u,j}$は，文献[5]で提案されている最もありそうなモデルの評価戦略によって評価することができる．行動A_uの場合における状態遷移$S_i \to S_j$の可能性見積もり値は次式で表される．

$$\tilde{p}_{i,j}^{k+1}(u) = \frac{M_{i,u,j}^{k+1}}{L_{i,u}^{k+1}},\ i,j \in I, u \in U \qquad (9)$$

ここで，

$$L_{i,u}^{k+1} = L_{i,u}^{k} + m_i^S(s_k)\cdot m_u^A(a_k),\ i \in I, u \in U \qquad (10)$$

は，ファジィ状態S_iにおいて$k+1$回繰り返されたファジィ行動A_uの試行回数，m_i^Sはファジィ集合S_iのメンバーシップ関数である．

$$M_{i,u}^{k+1} = M_{i,u}^{k} + m_i^S(s_k)\cdot m_u^A(a_k)\cdot m_j^S(s_{k+1}) \qquad (11)$$

$$i,j \in I, u \in U$$

は，ファジィ状態S_iにおいて$k+1$回繰り返されたファジィ行動A_uの試行の中で状態がS_jへ遷移した回数である．

(ファジィ状態S_iにおけるファジィ行動A_uの実行に対する) 平均報酬の見積もり$\tilde{g}_{i,u,j}$を対応させると[5],

$$\tilde{g}_{i,u,j}^{k+1} = \frac{\sum_{l=0}^{k} m_i^S(s_l)\cdot m_u^A(a_l)\cdot m_j^S(s_{l+1})\cdot g_l}{\sum_{l=0}^{k} m_i^S(s_l)\cdot m_u^A(a_l)\cdot m_j^S(s_{l+1})} \qquad (12)$$

$$i,j \in I, u \in U$$

ここでg_lはステップlにおいて得られる報酬である．

行動評価関数を近似すると最適政策は文献[5]で提案されているように関数(6), (7)で構築される．

1つ目の例と比較すると，この場合における SVD 削減法の唯一の違いは状態空間と行動空間に前もって定義されたファジィ分割が必要であることである(状態遷移の可能性と対応する平均報酬の評価計算に必要)．これらの分割を許容される計算コスト内で構築する．

4． 結論

複雑な状況で強化学習法を応用する場合の問題の1つは，状態評価関数あるいは行動評価関数の表現が巨大になってしまうことである[1]．連続的な環境での強化学習は更に複雑である．連続的な空間を記述するため，即ち基本的に未知の状態評価関数または行動評価関数を正確に近似するために密な分割を適用するからである．細かく分割すると状態数が増加し，多数の状態を扱うと計算コストが増大する．これは，実時間アプリケーションだけでなく実際に使用する（限られた）計算機能力において許容できない問題である．こういった問題の簡単な解決法として，連続的環境における強化学習アプリケーションの高いストレージ・コストと計算コストに対処する

ため， SVD **に基づいて代数積・加算重心**(PSG)**法**[2,3]**に基づいたファジィ規則を近似する技法を適応することを提案した．本手法の適用により，従来計算機の処理能力に制限されていた強化学習法の近似能力を大幅に改善し，効率的な制御規則の獲得が可能となった．**

文　献

[1] R. S. Sutton, A. G. Barto: “Reinforcement Learning: An Introduction”, MIT Press, Cambridge, 1998.

[2] Y. Yam, P. Baranyi, C. T. Yang: “Reduction of Fuzzy Rule Base Via Singular Value Decomposition”, IEEE Transaction on Fuzzy Systems. Vol.: 7, No. 2, 1999, pp. 120-131.

[3] P. Baranyi, Y. Yam: “Fuzzy rule base reduction”, Chapter 7 of Fuzzy IF-THEN Rules in Computational Intelligence: Theory and Applications Eds., D. Ruan and E.E. Kerre, Kluwer, 2000, pp 135-160.

[4] C. J. C. H. Watkins: “Learning from Delayed Rewards”, Ph.D. thesis, Cambridge University, Cambridge, England, 1989.

[5] M. Appl: “Model-based Reinforcement Learning in Continuous Environments”, Ph.D. thesis, Technical University of München, München, Germany, dissertation.de, Verlag im Internet, 2000.

[6] R. Bellman: “Dynamic Programming”, Princeton University Press, 1957.