バイオメディカル・ファジィ・システム学会誌　Vol.21，No.1，pp.41－48（2019）

[Original article]

（2019 年 3 月 26 日　Accepted）

推薦システムにおける全結合ニューラルネットワークを用いた強化学習

前田　康成[1]

1）北見工業大学・地域未来デザイン工学科

要約：　従来研究では，推薦システムを表現するための確率モデルとしてマルコフ決定過程が採用されている．他方，多くの分野において，ニューラルネットワークを用いた強化学習方法が提案されている．しかし，推薦システムにおけるニューラルネットワークを用いた強化学習方法は提案されていない．そこで，本研究では，マルコフ決定過程の真のパラメータが未知の仮定のもとで推薦システムにおける全結合ニューラルネットワークを用いた強化学習方法を提案する．提案方法では顧客の性質を表現するために顧客の履歴情報を利用する．シミュレーションによって提案方法の有効性を示す．シミュレーション結果では，提案方法の出力が最適解と一致した．

キーワード：推薦システム，マルコフ決定過程，強化学習，全結合ニューラルネットワーク

Reinforcement Learning using Fully Connected Neural Networks in Recommender System

Yasunari MAEDA[1]

1) School of Regional Innovation and Social Design Engineering, Kitami Institute of Technology

Abstract: *Markov decision processes are applied to recommender system in previous research. Reinforcement learning methods using neural networks have been proposed in many fields. But a reinforcement learning method using neural networks has not been proposed in recommender system. In this research we propose a reinforcement learning method using fully connected neural networks in recommender system under the condition that the true parameters of Markov decision processes are unknown. The proposed method uses historical data of customers to represent customers' properties. The effectiveness of the proposed method is shown by some simulations. The output of the proposed method is equal to the optimal solution in the simulation result.*

Keywords: *recommender system, Markov decision processes, reinforcement learning, fully connected neural networks*

Yasunari MAEDA
165 Koen-cho, Kitami-shi, Hokkaido, 090-8507, Japan
Phone:　+81-157-26-9328,　Fax: +81-157-26-9344,　E-mail: maedaya@mail.kitami-it.ac.jp

1. はじめに

本研究では，インターネット上の通信販売サイトなどで顧客に商品を推薦する際に利用されている推薦システム[1]を扱う．推薦システムに関しては，嗜好が類似した顧客は似たような購買活動をするという仮定のもとで，協調フィルタリングなどによって購入する確率の高い商品を推薦する方法が多く検討されてきた．しかし，推薦システムの本来の目的は売上高の最大化である．そこで，近年の従来研究[2][3][4][5][6][7]では他分野でも売上高などの最大化に利用されている確率モデルであるマルコフ決定過程（MDP）[8]を，推薦システムを表現するための確率モデルとして採用している．これらの従来研究では顧客の性質を表現するために顧客の購買履歴情報を利用している．本研究でもマルコフ決定過程を採用し，顧客の購買履歴情報を利用して売上高の最大化を検討する．

近年，多くの分野においてニューラルネットワークを用いた強化学習方法が提案されている．囲碁ソフトへの適用例[9]は特に有名である．強化学習にもいろいろな学習方法があるが，未知情報を伴うマルコフ決定過程を学習対象としたQ学習[10]がよく利用されており，推薦システムに関しても有用な学習方法だと考えられる．なお，ニューラルネットワークを用いたQ学習は，厳密にはニューラルネットワークを用いたQ学習の近似に相当する．しかし，推薦システムに関してはニューラルネットワークを用いた強化学習方法はまだ提案されていない．

そこで，本研究では推薦システムにおけるニューラルネットワークを用いた強化学習方法の基礎検討の一例として，推薦システムにおける全結合ニューラルネットワーク[11][12]を用いたQ学習アルゴリズムを提案する．本研究の特徴は，従来から検討されている売上高の最大化を目的としたマルコフ決定過程を用いた推薦システムに対して，新たにニューラルネットワークを用いた強化学習を適用する点である．ニューラルネットワークにはさまざまな種類があるが，本研究で採用する全結合ニューラルネットワークは前の層のすべてのユニットと次の層のすべてのユニットが結合している基本的なニューラルネットワークの１つである．なお，Q学習ではマルコフ決定過程の状態遷移確率を支配する真のパラメータと利得が未知の場合を学習対象とすることが多いが，本研究では真のパラメータが未知で，利得については利得の最大値（推薦システムにおける最高金額の商品の金額（売上高））は既知と仮定する．ただし，提案方法で用いる全結合ニューラルネットワークの出力層の活性化関数を変更することによって，利得の最大値も含めて利得について完全に未知の場合に対しても提案方法は容易に拡張可能である．

以下，2章で本研究で使用する各種記号などについて説明する．3章で推薦システムにおける全結合ニューラルネットワークを用いたQ学習アルゴリズムを提案し，4章で提案方法の有効性をシミュレーションによって確認する．5章で提案方法の計算量と改善案について考察し，最後に6章でまとめと今後の課題について述べる．

2. 準備

ここでは，本研究で用いる各種記号などを説明する．

2.1 推薦システムに関する各種記号など

m_i，$m_i \in M$は推薦対象の商品を示し，$M = \{m_1, m_2, \ldots, m_{|M|}\}$は商品集合である．本研究では，従来研究と同様に推薦システムを表現する確率モデルとしてマルコフ決定過程を採用し，商品m_iの推薦がマルコフ決定過程の行動選択に相当する．n_i，$n_i \in N$は推薦に対する顧客の反応を示し，$N = \{n_1, n_2, \ldots, n_{|N|}\}$は反応集合である．$1 \le i \le |M|$では$n_i = m_i$であり，顧客の反応が商品$m_i$の購入に相当する．$|N| = |M| + 1$であり，$n_{|M|+1}$は顧客が何も購入しなかったことを示す．$r(n_i)$，$1 \le i \le |M|$は商品$m_i$の売上高を示し，顧客が何も購入しなかった場合には$r(n_{|M|+1}) = 0$である．$r(n_i)$はマルコフ決定過程の利得に相当する．

Y_tはマルコフ決定過程におけるt期の行動を示し，推薦システムでは商品の推薦Y_t, $Y_t \in M$に相当する．Z_t, $Z_t \in N$は推薦Y_tに対する顧客の反応を示す．X_t，$X_t = Y_{t-2}Z_{t-2}Y_{t-1}Z_{t-1} = (Y_{t-2}, Z_{t-2}, Y_{t-1}, Z_{t-1}) \in M^2N^2$はマルコフ決定過程の$t$期の状態を示す．状態は過去2期間の推薦と反応の組に相当する．状態定義については，過去のより長い期間の履歴や顧客の反応のみで構成するなど，さまざまな定義が考えられる．本研究の提案方法は他の状態の定義に対しても容易に拡張可能である．

$\Pr(Z_t|X_t, Y_t, \theta^*)$は状態$X_t$において商品$Y_t$を推薦した場合に顧客の反応が$Z_t$になる確率でマルコフ決定過程の状態遷移確率に相当する．状態X_tは過去2期間の推薦商品と顧客の反応の組$(Y_{t-2}, Z_{t-2}, Y_{t-1}, Z_{t-1})$なので，$\Pr(Z_t|X_t, Y_t, \theta^*) = \Pr(Z_t|Y_{t-2}, Z_{t-2}, Y_{t-1}, Z_{t-1}, Y_t, \theta^*)$である．本研究における状態遷移確率は2期前の推薦商品がY_{t-2}，2期前の顧客の反応がZ_{t-2}，1期前の推薦商品がY_{t-1}，1期前の顧客の反応がZ_{t-1}という条件の顧客に

対して商品Y_tを推薦した場合に当該顧客の反応がZ_tになる確率である．遷移先の状態X_tを明記する表記にすると，$\Pr(X_{t+1}|X_t, Y_t, \theta^*) = \Pr(Z_t|X_t, Y_t, \theta^*)$となる．$\theta^*$は状態遷移確率を支配する真のパラメータで本研究では未知である．

本研究の目的は無限期間の割引総利得である$\sum_{t=1}^{\infty}\beta^{t-1}r(Z_t)$の期待値に相当する期待割引総利得を最大にする商品の推薦の仕方の学習である．ただし，β，$0<\beta<1$は割引率である．なお，本研究では全結合ニューラルネットワークの出力層の活性化関数として1以下の値を返す関数を利用するため，常に最大の利得を獲得し続けた場合の割引総利得が1に変換されるように利得を$\frac{1-\beta}{\max_{m_i} r(m_i)}$倍した変換後の値を利用する．変換後の利得の値に基づいて期待割引総利得の最大化を検討しても，元の利得を定数倍しているだけなので元の利得に対する最適な行動選択と変換後の利得に対する最適な行動選択は同じである．

2.2　全結合ニューラルネットワークに関する各種記号など

本研究では，入力層，中間層（1層），出力層の3層構造の全結合ニューラルネットワークを用いる（図. 1）．各層はいくつかのユニット（ニューロン）で構成され，各ユニットでは入力値を受け取って何らかの処理結果を出力値として出力する．入力層のユニット数と中間層のユニット数は推薦システムにおけるマルコフ決定過程の状態数と同じ$|M|^2|N|^2$とし，出力層のユニット数はマルコフ決定過程の行動数に相当する推薦システムの商品数と同じ$|M|$とする．

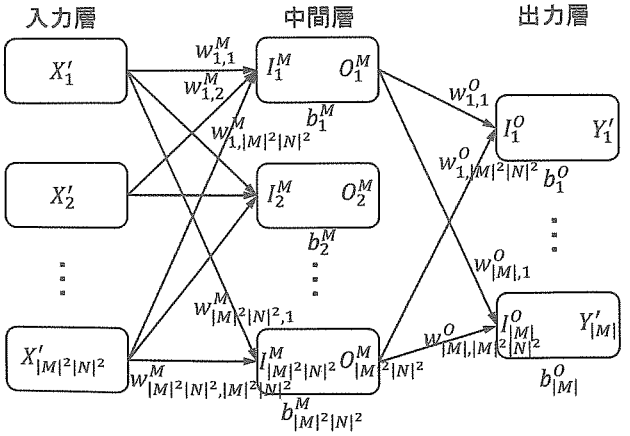


図. 1　本研究の全結合ニューラルネットワーク

X'_i，$X'_i \in \{0,1\}$，$1 \le i \le |M|^2|N|^2$は入力層におけるi番目のユニットの入力値で，マルコフ決定過程の状態を1から$|M|^2|N|^2$の通し番号で番号付けした場合のi番目の状態s_iに相当する．推薦システムにおける状態がX_t，$X_t = Y_{t-2}Z_{t-2}Y_{t-1}Z_{t-1} = m_i n_j m_k n_l$のとき，マルコフ決定過程の状態番号$N(X_t)$は次式で算出できる．

$$N(X_t) = (i-1)|N||M||N| + (j-1)|M||N| + (k-1)|N| + (l-1) + 1. \tag{1}$$

入力層のユニットは何も処理を実施せずに入力値をそのまま出力値として出力する．I_i^Mは中間層のi番目のユニットの入力値で次式による．

$$I_i^M = \sum_{j=1}^{|M|^2|N|^2} w_{i,j}^M X'_j + b_i^M, \tag{2}$$

ただし，I_i^M，$w_{i,j}^M$，b_i^Mの添え字Mは中間層を示す添え字で，$w_{i,j}^M$は中間層のi番目のユニットにおける入力層のj番目のユニットの出力値X'_jに対する重み，b_i^Mは中間層のi番目のユニットにおけるバイアスである．O_i^Mは中間層のi番目のユニットの出力値で次式による．

$$O_i^M = f_M(I_i^M), \tag{3}$$

ただし，O_i^M，f_Mの添え字Mは中間層を示す添え字で，$f_M(I_i^M)$は入力値I_i^Mを出力値O_i^Mに変換する中間層の活性化関数である．中間層の活性化関数にはいろいろな関数が利用されるが，本研究では次式によるシグモイド関数を利用する．

$$f_M(I_i^M) = \frac{1}{1+e^{-I_i^M}}. \tag{4}$$

I_i^Oは出力層のi番目のユニットの入力値で次式による．

$$I_i^O = \sum_{j=1}^{|M|^2|N|^2} w_{i,j}^O O_j^M + b_i^O, \tag{5}$$

ただし，I_i^O，$w_{i,j}^O$，b_i^Oの添え字Oは出力層を示す添え字で，$w_{i,j}^O$は出力層のi番目のユニットにおける中間層のj番目のユニットの出力値O_j^Mに対する重み，b_i^Oは出力層のi番目のユニットにおけるバイアスである．Y'_iは出力層のi番目のユニットの出力値で次式による．

$$Y'_i = f_O(I_i^O), \tag{6}$$

ただし，f_Oの添え字Oは出力層を示す添え字で，$f_O(I_i^O)$は入力値I_i^Oを出力値Y'_iに変換する出力層の活性化関数である．出力層の活性化関数にもいろいろな関数が利用されるが，本研究では中間層と同様に次式によるシグモイド関数を利用する．

$$f_O(I_i^O) = \frac{1}{1+e^{-I_i^O}}. \tag{7}$$

入力層の入力値に対して式(2)，式(3)，式(5)，式(6)を用いることによって，全結合ニューラルネットワークの

最終的な処理結果（出力値）が得られる．例として，入力値が$X'_j = 1$で，j以外のすべてのkについて$X'_k = 0$，$1 \le k \le |M|^2|N|^2$，$k \ne j$であれば入力値はj番目の状態s_jを示しており，このときの出力値Y'_iはQ学習における状態s_jにおいて行動（推薦商品）m_iを選択する場合のQ値$Q(s_j, m_i)$の推定値である．Q学習では学習に成功すると，$Q(s_j, m_i)$が状態s_jで行動m_iを実施して以降，最適な行動を選択し続ける場合の割引総利得の期待値に収束する．

3. 推薦システムにおける全結合ニューラルネットワークを用いたQ学習アルゴリズムの提案

全結合ニューラルネットワークの学習では，入力層の入力値に対して目的の出力値が得られるように，中間層の重み$w^M_{i,j}$，バイアスb^M_i，出力層の重み$w^O_{i,j}$，バイアスb^O_iを学習（調整）する．学習データは入力値と目的の出力値（教師データ）の組で構成され，出力層のi番目のユニットに対する教師データをt_iと表記すると，出力値に対する誤差（二乗誤差）eは次式で算出される．

$$e = \frac{1}{2}\sum_{i=1}^{M}(t_i - Y'_i)^2, \tag{8}$$

ただし，係数$\frac{1}{2}$は後述の式を見易くするためであり，特に意味はない．

以下でQ学習における誤差逆伝播法（バックプロパゲーション法）による全結合ニューラルネットワークの各種重みなどの更新方法を説明する．出力層のi番目のユニットにおける更新後のバイアス$\hat{b}^O_i$を次式に示す．

$$\hat{b}^O_i = b^O_i - \eta\frac{\partial e}{\partial b^O_i} = b^O_i - \eta\delta^O_i, \tag{9}$$

ただし，

$$\delta^O_i = \begin{cases} \frac{\partial e}{\partial Y_i} f'_O(I^O_i), & \text{実施された推薦が}m_i; \\ 0, & \text{実施された推薦が}m_i\text{以外}, \end{cases} \tag{10}$$

$$\frac{\partial e}{\partial Y_i} = Y'_i - t_i, \tag{11}$$

$$f'_O(I^O_i) = f_O(I^O_i)\left(1 - f_O(I^O_i)\right), \tag{12}$$

ηは学習係数，式(11)は式(8)の二乗誤差の微分，式(12)は式(7)のシグモイド関数の微分である．Q学習ではシミュレーション環境／実環境で実際に選択（実施）した行動（推薦システムでは推薦商品）に関する教師データのみ入手でき，その他の行動（商品）に関する教師データは入手できない．よって，δ^O_iは式(10)のようになる．例として，学習のt'回目の繰返し（本研究では，全結合ニューラルネットワークの1回の学習中にQ学習も1回の学習とする．）において，推薦システムの状態が$s_{N(X_{t'})}$，推薦商品が$Y_{t'}$，顧客の反応が$Z_{t'}$，次の状態が$s_{N(X_{t'+1})}$の場合の教師データ$t_{N(Y_{t'})}$（$N(Y_{t'})$は$Y_{t'}$の商品番号）を次式に示す．

$$t_{N(Y_{t'})} = (1-\alpha)Q\left(s_{N(X_{t'})}, Y_{t'}\right) + \alpha\left(r(Z_{t'}) + \beta\max_{m_j} Q\left(s_{N(X_{t'+1})}, m_j\right)\right), \tag{13}$$

ただし，αはQ学習における学習率，$Q\left(s_{N(X_{t'})}, Y_{t'}\right)$および$Q\left(s_{N(X_{t'+1})}, m_j\right)$は学習中の当該時点の全結合ニューラルネットワークに状態$s_{N(X_{t'})}$を入力した際の出力$Y'_{N(Y_{t'})}$と状態$s_{N(X_{t'+1})}$を入力した際の出力Y'_jである．式(11)中の教師データの算出に式(13)を利用する．なお，式(13)はもとのQ学習[10]におけるQ値$Q\left(s_{N(X_{t'})}, Y_{t'}\right)$の更新式に相当する．

出力層のi番目のユニットにおける中間層のj番目のユニットの出力値O^M_jに対する更新後の重み$\hat{w}^O_{i,j}$を次式に示す．

$$\hat{w}^O_{i,j} = w^O_{i,j} - \eta\frac{\partial e}{\partial w^O_{i,j}} = w^O_{i,j} - \eta\delta^O_i O^M_j. \tag{14}$$

中間層のi番目のユニットにおける更新後のバイアス$\hat{b}^M_i$を次式に示す．

$$\begin{aligned} \hat{b}^M_i &= b^M_i - \eta\frac{\partial e}{\partial b^M_i} = b^M_i - \eta\delta^M_i \\ &= b^M_i - \eta\left(\sum_{j=1}^{M}\delta^O_j w^O_{j,i}\right) f'_M(I^M_i) \\ &= b^M_i - \eta\left(\sum_{j=1}^{M}\delta^O_j w^O_{j,i}\right) f_M(I^M_i)\left(1 - f_M(I^M_i)\right), \end{aligned} \tag{15}$$

ただし，$f'_M(I^M_i)$はシグモイド関数の微分である．

中間層のi番目のユニットにおける入力層のj番目のユニットの出力値X'_jに対する更新後の重み$\hat{w}^M_{i,j}$を次式に示す．

$$\hat{w}^M_{i,j} = w^M_{i,j} - \eta\frac{\partial e}{\partial w^M_{i,j}} = w^M_{i,j} - \eta\delta^M_i X'_j. \tag{16}$$

次に，学習の流れを以下に示す．

【ステップ1】全結合ニューラルネットワークの中間層の重み$w_{i,j}^{M}$，バイアスb_i^M，出力層の重み$w_{i,j}^{O}$，バイアスb_i^Oの初期値と初期状態$state_{now}$を設定する．

【ステップ2】以下の【処理2-1】から【処理2-4】を繰返回数分（または，得られる利得が十分大きくなるまで）繰返す．

【処理2-1】全結合ニューラルネットワークに現在の状態$state_{now}$を入力して，商品m_i，$m_i \in M$を推薦する場合のQ値$Q(state_{now}, m_i)$に相当する出力値Y_i'を算出する．

【処理2-2】確率εでランダムに現在の行動（商品）$action_{now}$を選択し，確率$1-\varepsilon$で次式によって選択する．

$$action_{now} = \arg\max_{m_i} Q(state_{now}, m_i), \quad (17)$$

【処理2-3】状態$state_{now}$で行動$action_{now}$を実施（商品を推薦）し，顧客の反応に応じた利得rと次の状態$state_{next}$を観測する．観測結果に基づいて教師データ$t_{N(action_{now})}$を次式によって算出する．

$$t_{N(action_{now})} = (1-\alpha)Q(state_{now}, action_{now}) + \alpha\left(r + \beta \max_{m_i} Q(state_{next}, m_i)\right). \quad (18)$$

【処理2-4】中間層の重み$w_{i,j}^{M}$，バイアスb_i^M，出力層の重み$w_{i,j}^{O}$，バイアスb_i^Oを更新し，$state_{now} = state_{next}$として【処理2-1】に戻る．

4. シミュレーション結果

小規模ではあるが，3章における提案アルゴリズムの有効性を確認するためのシミュレーション結果を報告する．シミュレーションにおける各種設定は著者の主観による設定である．

推薦システムにおける商品数$|M| = 2$，顧客の反応数$|N| = |M| + 1 = 3$，割引率$\beta = 0.6$，利得$r(n_1 = m_1) = 100$，$r(n_2 = m_2) = 80$，$r(n_3 = \text{無購入}) = 0$とする．ただし，本研究の全結合ニューラルネットワークの出力層の出力値が1以下なので，期待割引総利得も1以下に変換する必要がある．そこで，上記の利得を$\frac{1-\beta}{\max_{m_i} r(m_i)} = \frac{1}{250}$倍した値を学習アルゴリズム中の利得として利用する．推薦商品に対する顧客の反応確率である状態遷移確率は次式のように設定した．

$$\Pr(Z_t | Y_{t-2}, Z_{t-2}, Y_{t-1}, Z_{t-1}, Y_t, \theta^*) = \begin{cases} 0.8, & Y_{t-1} = Z_{t-1} = Y_t = Z_t; \\ 0.1, & Y_{t-1} = Z_{t-1} = Y_t \text{かつ} Y_t \neq Z_t; \\ 0.4, & Y_{t-1} = Z_{t-1} = Y_t \text{以外かつ} Y_t = Z_t; \\ 0.2, & Y_{t-1} = Z_{t-1} = Y_t \text{以外かつ} Z_t \text{が} Y_t \text{以外の商品}; \\ 0.4, & Y_{t-1} = Z_{t-1} = Y_t \text{以外かつ} Z_t \text{が無購入} n_3. \end{cases} \quad (19)$$

Q学習でQ値によって行動（推薦商品）を選択するかランダムに選択するかの閾値の確率を$\varepsilon = 0.4$，Q学習の学習率を$\alpha = 0.9$，初期状態を$X_1 = s_1 = m_1 n_1 m_1 n_1$とした．全結合ニューラルネットワークの学習の繰返し回数は1000万回とした．本研究では全結合ニューラルネットワークの1回の学習に対してマルコフ決定過程の行動選択，状態遷移，利得獲得という一連のプロセスとQ値の更新を1回実施するので，Q学習の繰返し回数も1000万回である．全結合ニューラルネットワークの学習係数を$\eta = 0.001$とし，中間層の重み$w_{i,j}^{M}$，バイアスb_i^M，出力層の重み$w_{i,j}^{O}$，バイアスb_i^Oの初期値を正規乱数で設定した．選択した行動（推薦した商品）に対応した状態遷移（顧客の反応の生起）をシミュレートする際には，一様乱数を利用して状態遷移確率に従って遷移先の状態を生起させた．本研究ではシミュレーション用のプログラムをJava言語で作成し，OSがWindows 10 Pro 64ビットOS，CPUがIntel Core i7-7500U CPU 2.70GHz，メモリ（RAM）が12GBの計算機を使用した．シミュレーションの処理時間は約3分間だった．

1000万回の学習後の全結合ニューラルネットワークに全状態である状態s_1から状態s_{36}相当の入力値を入力した場合の出力結果Y_j'がQ値$Q(s_i, m_j)$（正確にはQ値の推定値）に相当する．各状態s_iに対するQ値最大の行動$\arg\max_{m_j} Q(s_i, m_j)$が学習結果による選択行動（推薦商品）である．状態遷移確率などすべての情報が既知の場合に期待割引総利得を最大にするという意味で最適な行動選択と最適な行動選択による期待割引総利得と1000万回の学習後の全結合ニューラルネットワークの出力を比較した．すべての情報が既知の場合の最適行動などは政策反復法（政策改良法），価値反復法（逐次近似法）などで算出できる[8]．全36状態における行動選択は最適な行動とすべて一致した．各状態s_iに対するQ値の最大値$\max_{m_j} Q(s_i, m_j)$は初期状態がs_iで無限期間にわたって最適な行動を選択し続ける場合の期待割引総利得に収束することが期待される．全36状態におけるQ値の最大値と状態遷移確率などすべての情報が既知の場合の

最適な行動選択に基づく期待割引総利得（商品の売上に相当する元の利得を$\frac{1-\beta}{\max_{m_i} r(m_i)} = \frac{1}{250}$倍した変換後の期待割引総利得）の比較を図. 2，図. 3，図. 4に示す．図のv価値（左側）が最適な行動選択による期待割引総利得，qNNW（右側）がQ値の最大値，s1からs36は状態s_1から状態s_{36}である．図のv価値とqNNWを状態ごとに比較することにより，ほぼ同程度の値であることが確認できる．v価値とqNNWの絶対誤差の平均は0.0091だった．状態遷移確率などすべての情報が既知の場合の最適な行動選択に基づく期待割引総利得（v価値）が最も大きな状態で0.829，最も小さな状態で0.661であることをふまえると，最適な行動選択の学習のみではなく，期待割引総利得の推定精度も十分だと考えられる．

以上より，小規模なシミュレーション例ではあるが提案アルゴリズムの有効性が確認できた．本研究はニューラルネットワークによる強化学習の推薦システムへの適用に関する基礎検討であり，提案方法に関しては計算量の視点から改善の必要性がある．提案方法の計算量に関する考察を5章で述べる．

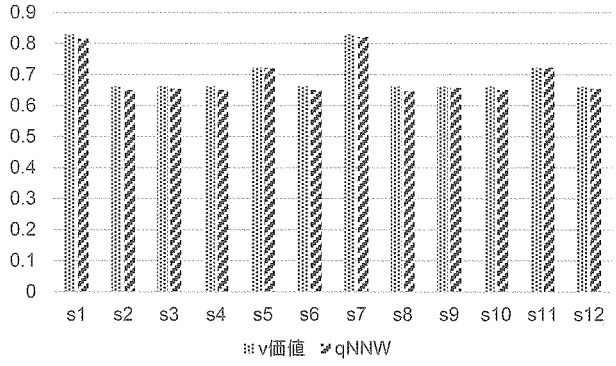


図. 2　Q値と期待割引総利得の比較（その1）

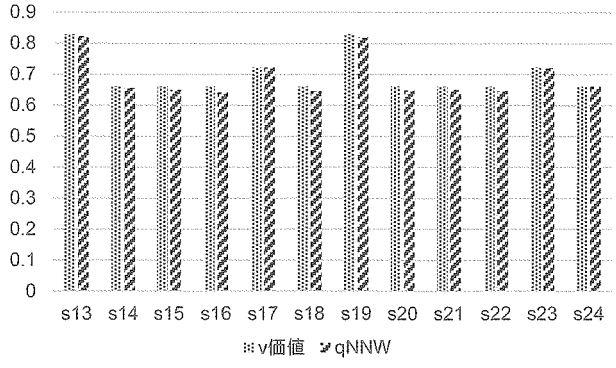


図. 3　Q値と期待割引総利得の比較（その2）

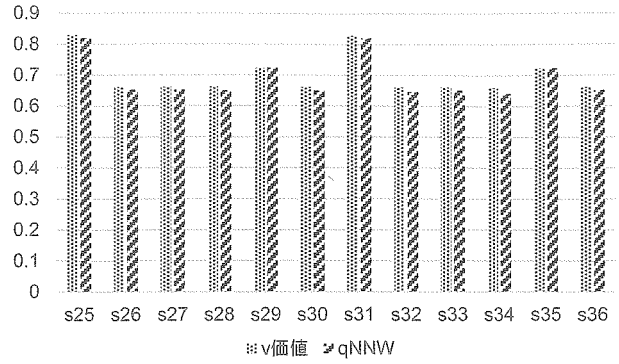


図. 4　Q値と期待割引総利得の比較（その3）

5. 考察

4章で紹介したシミュレーションでの商品数は$|M| = 2$であり，現実で想定される規模よりも明らかに小さい．提案方法に関するより詳細な評価のためには，より大きな商品数でのシミュレーションが必要である．しかし，より大きな商品数に対して提案方法の処理時間が膨大になることが推察されるため，今回は前述の商品数$|M| = 2$の場合の紹介のみとした．

前述のシミュレーションの処理時間は約3分間だった．使用した全結合ニューラルネットワークの構造から考えると，処理時間の大半は中間層の重みの更新だと推察される．中間層の重みの変数の数は図. 1の入力層から中間層への矢印の数と同じである．提案モデルでは入力層と中間層のユニット数がともに状態数と同じ$|M|^2|N|^2$なので，矢印の総数は$|M|^4|N|^4$である．顧客の反応の数$|N|$は$|N| = |M| + 1$なので，商品数$|M| = 2$の場合であれば矢印の総数は$2^4 3^4 = 1296$である．仮に商品数$|M| = 10$の場合であれば，矢印の総数は$10^4 11^4 = 146410000$で約11万倍となり処理時間も同様に膨大になることが推察される．

以上より，提案方法は条件次第では前述のシミュレーション例のように最適解と一致するような有効な性質があるものの，計算量の視点から改善が必要である．提案方法では，過去2期間の推薦と顧客の反応の組（顧客の履歴）を状態としているため，状態数$|M|^2|N|^2$が商品数$|M|$の増加に伴い大きくなる傾向にある．例えば，使用する顧客の履歴を過去1期間にしたり，顧客の反応のみに変更することによって計算量の軽減が可能である．

また，提案方法ではニューラルネットワークの中間層のユニット数を入力層のユニット数と同様に状態数としているが，状態数が増加した場合には中間層のユニット数が多過ぎて学習精度を低下させる可能性（一般的にニューラルネットワークではユニット数が多過ぎると

学習精度が低下することがある）もある．よって，計算量および学習精度の視点から中間層のユニット数の軽減も要検討である．

本研究では，商品のクラス分類は考慮しなかったが，商品の性質などによって分類したクラスによって状態（顧客の履歴情報）を表現することによって状態数を軽減することも可能である．さらに本研究では中間層が1層の全結合ニューラルネットワークを利用したが，畳み込みニューラルネットワークなど全結合以外のニューラルネットワークによる計算量の軽減や，中間層の多層化（深層学習の適用）による少ないユニット数のもとでの多層化による高精度化などの改善案も考えられる．

計算量の視点による提案方法の改善案の具体的な検討については今後の課題としたい．6章では，本研究のまとめと上記以外の今後の課題について説明する．

6. まとめと今後の課題

従来から確率モデルとしてマルコフ決定過程を採用した推薦システムについて数多く検討されてきた．他方，深層強化学習などニューラルネットワークを用いた強化学習（Q学習）がさまざまな分野に適用されつつある．推薦システムに関してもニューラルネットワークを用いた強化学習（Q学習）が適用可能と考えられるが，従来は未検討だった．そこで，本研究では推薦システムにおける全結合ニューラルネットワークを用いたQ学習アルゴリズムを提案した．また，小規模ではあるがシミュレーションによって提案方法の有効性を確認した．本研究はニューラルネットワークによる強化学習の推薦システムへの適用に関する基礎検討であり，計算量の視点より提案方法には改善が必要である．提案方法の計算量および改善案に関する考察は5章で述べたとおりである．

深層強化学習などのニューラルネットワークを用いた強化学習の従来研究では，シミュレーション例などで良好な行動選択結果が数多く報告されているが，状態遷移確率などが既知の場合の最適解との比較に関する報告はほとんどない．本研究のシミュレーション例では，提案アルゴリズムによる行動選択が最適な行動選択と一致し，期待割引総利得の推定値に相当するQ値の推定精度も十分高いことが確認できた．

本研究のシミュレーション例では商品の価格や顧客の反応確率（マルコフ決定過程の状態遷移確率）などの設定を著者の主観に基づいて設定したが，より詳細な評価を行うためには実データに基づく設定のもとで評価する必要がある．実データを用いたシミュレーションに基づく評価については今後の課題としたい．

次に本研究内容の健康・医療分野への適用可能性について述べる．本研究ではインターネット上の通信販売サイトなどにおける推薦システムを対象としたが，扱う商品は限定していない．よって，本研究の検討内容は医療／健康関連商品や医療保険，各種医療サービスを商品として扱う推薦システムにも適用可能である．また，本研究で確率モデルとして採用しているマルコフ決定過程は，従来研究[13]において医療検査項目の選択や医療アドバイスの選択などのヘルスケア支援を表現する確率モデルとしても採用されている．従来研究では患者の真の健康状態をマルコフ決定過程の状態，検査項目やアドバイスの選択をマルコフ決定過程における行動選択としてモデル化し，ヘルスケアに関する総コストの最小化を検討しているが，状態遷移確率などの各種確率は既知のもとで検討している．しかし，より現実に近い設定としては各種確率が未知の場合が想定される．よって，本研究の検討内容をヘルスケア支援に適用することにより，より現実に近い設定のもとでのヘルスケアに関する総コストの最小化が検討できる．具体的な検討については今後の課題としたい．

謝辞

本研究の一部はJSPS 科研費 JP16K00417 の助成による．

参考文献

[1] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich：情報推薦システム入門，共立出版，東京，2012.

[2] G. Shani, D. Heckerman, and R. I. Brafman, : An MDP-Based Recommender System, Journal of Machine Learning Research, Vol.6, pp.1265-1295, 2005.

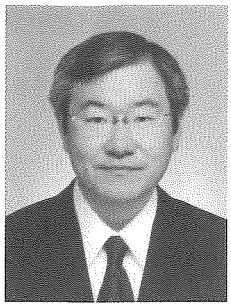
[3] 桑田修平，前田康成，松嶋敏泰，平澤茂一：推薦システムのための状態遷移確率の構造を未知としたマルコフ決定過程，情報処理学会論文誌 数理モデル化と応用, Vol.6, No.1, pp.20-30, 2013.

[4] 前田康成，鈴木正清，松嶋敏泰：顧客クラスが変化する推薦システムに関する一考察，電気学会論文誌C, Vol.137, No.6, pp.815-816, 2017.

[5] 前田康成，山内翔，鈴木正清，松嶋敏泰：推薦システムにおける新規顧客問題に関する一考察，バイオメディカル・ファジィ・システム学会誌, Vol.19, No.2, pp.13-19, 2017.

[6] 前田康成，山内翔，鈴木正清，松嶋敏泰：顧客クラスが変化する推薦システムにおける半教師付き学習，バイオメディカル・ファジィ・システム学会誌, Vol.20, No.1,

pp.15-22, 2018.
[7] 前田康成，山内翔，鈴木正清，松嶋敏泰：推薦システムの新規顧客問題における半教師付き学習，バイオメディカル・ファジィ・システム学会誌, Vol.20, No.1, pp.37-46, 2018.
[8] 金子哲夫：マルコフ決定理論入門，槙書店，東京，1973.
[9] D. Silver et.al., : Mastering the game of Go with deep neural networks and tree search, Nature, Vol.529, pp.484-503, 2016.
[10] C.J.C.H. Watkins, and P. Dayan : Q-Learning, Machine Learning, Vol.8, pp.279-292, 1992.
[11] 巣籠悠輔：詳解ディープラーニング，マイナビ出版，東京，2017.
[12] 涌井良幸，涌井貞美：ディープラーニングがわかる数学入門，技術評論社，東京，2017.
[13] 前田康成，山内翔，鈴木正清，高野賢裕，松嶋敏泰：マルコフ決定過程を用いたヘルスケア支援に関する一考察，バイオメディカル・ファジィ・システム学会誌, Vol.19, No.2, pp.21-27, 2017.



前田康成（まえだやすなり）
平成7年早大・理工卒．平成9年同大学院理工学研究科修士課程修了．日本電信電話（株），東日本電信電話（株），北見工大助手，助教，准教授を経て平成28年同大学教授，現在に至る．博士（工学）．統計的決定理論の学習問題への応用に関する研究に従事．電子情報通信学会等各会員．