The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

隠れ状態を持つ多腕バンディット問題における方策の検討

A study on measures in multi-armed bandit problem with hidden state

工藤亘平[*1]
Kouhei Kudo

竹川高志[*1]
Takasi Takekawa

[*1] 工学院大学情報学部
Faculty of Informatics Kogakuin University

Abstract: The Bandit problem is a matter of maximizing the current reward by selecting one out of the options and acquiring the reward, while limiting it to one state. Reinforcement learning is a problem of maximizing rewards earned in the future by performing various actions from options, in the presence of multiple states. The difference between the two is that state information is known, and multiple states are taken into account. In this simulation, we consider a model in which the current state and state transition information is unknown, maintaining one state for a certain period of time and then transitioning to another state. Regarding this model, we compare the general Bandit problem policy and reinforcement learning policy by cumulative reward. As a result, the cumulative reward was higher for the reinforcement learning policy than for the Bandit problem policy.

1. はじめに

近年はインターネットの普及によってインターネットショッピングやニュースの利用や, インターネット広告を目にする機会が増えている. 総務省の調べによるインターネットショッピングの個人利用率は全年代平均で約 7 割を超える結果[1]となっていることから, 老若男女に関係なく多くの人が利用していることがわかる. このため企業側も利益を最大化する目的で個人ごとの購買意欲などを高めるためにパーソナライズを行っている. パーソナライズは, ユーザーが過去に選択したコンテンツの利用履歴や年齢や性別などの個人情報に合わせてシステムが推薦するコンテンツを最適化する手法である.

パーソナライズに用いる手法としてバンディット問題があり, バンディット問題によるニュースのパーソナライズ化 [2]による個人の趣味嗜好にあったニュースの推薦などバンディット問題を活用したパーソナライズ化の方法が多く考えられている. バンディット問題は探索と知識利用のトレードオフを解決する問題であり, 1 つの状態に限定した中で, 選択肢の集合から 1 つを選びその選択肢から報酬を獲得してその報酬から次の選択を決定することで, 現在獲得できる報酬を最大化する問題である. 一般的にバンディット問題では選択肢をアーム, アームを選択する戦略を方策と呼ぶ[3].

またバンディット問題に似た問題設定を持つ手法として強化学習がある. 強化学習は複数の状態が存在し全ての状態を知っている中で, 選択肢の集合から様々な試行を繰り返していくことで, 未来で獲得する報酬を最大化する問題である. バンディット問題と強化学習の違いは, 状態の情報が既知で複数の状態を考慮しているかである.

バンディット問題でパーソナライズを行う場合はユーザー情報や過去の評価などによってコンテンツを推薦する. しかし人は気分や状況など状態の変化によって異なる行動を起こすため, 同様のコンテンツを推薦していても良い評価を得られない場合がある. 本論文ではこのような人の行動を現在の状態や状態遷移の情報が未知の設定で, 一定時間は 1 つの状態を維持して, その後他の状態に遷移する対人行動モデルとして考える. このモデルに対してバンディット問題の方策と強化学習の方策を用いてシミュレーションを行い, 一般的なバンディット問題の設定と比較をした.

連絡先:竹川高志, 工学院大学情報学部, 163-8677 東京都新宿区西新宿 1-24-2, 033340-0103, takekawa@cc.kogakuin.ac.jp

2. シミュレーション設定

2.1 モデル作成

人の感情や状況で行動が変化する性質を各状態で高い確率のアームが異なる設定で, 毎回各状態の推移確率によって状態遷移を繰り返すが一定の時間は 1 つの状態を維持しその後他の状態に遷移することでアームの確率が変化するモデルとして作成する. 本モデルは現在いる状態や状態数などの情報については方策側に提示せずに獲得した報酬によって方策に状態を推定させる. そして遷移確率は「初期状態」「確率分配」「状態遷移」の 3 つの状況がある.

「初期状態」は複数の状態が存在する中でいずれか 1 つの状態の推移確率が 100%となる状況である.「確率分配」が起こるまでの一定時間はこの状況が続く.

「確率分配」は一定時間経過後に毎回獲得する報酬の結果によって各状態の推移確率を増減させる状況である. この状況は「状態遷移」が起こるまで続く.

「状態遷移」は増減した推移確率で他の状態へ遷移した場合の行動である. 遷移した状態の推移確率を 100%, 他の状態の推移確率を 0%にして「初期状態」を行う.

2.2 方策

今回のシミュレーションで用いる方策のバンディット問題の ε -グリーディ法, UCB 方策, Thompson Sampling と強化学習の Q 学習と Q-Network について説明する.

ε -グリーディ法は, パラメータ ε によって定められた回数で最適なアームを探索し, 残りの回数で探索された最適なアームを選択する.

UCB 方策は, 毎回各アームで報酬の平均とアームの選択回数を用いた補正項を合わせた UCB スコアを計算して, そのスコアが最も高いアームを選択する[4].

Thompson Sampling は, 各アームの報酬の当たりとはずれの回数によって事後分布を作成し, その分布より生成される乱数が最大となるアームを選択する[5].

Q 学習は各アームに行動価値関数 Q 値を設定して獲得した報酬で Q 値を最適化していく方法である. 各 Q 値とソフトマックス関数を用いて確率的にアームを選択する.

Q-Network は, Q 学習とニューラルネットワークを合わせた学習法である. ニューラルネットワークでは行動の履歴と報酬を入力とし, 各アームの報酬の期待値を出力として Q 値を教師データとして学習を行う方策である. またアームの選択はソフトマックス関数を用いて確率的に選択する.

2.3 設定

本シミュレーション用いるバンディット問題の設定としてアームに用いる分布をベルヌーイ分布, 報酬は当たりで 1, はずれで 0 とする. また 1 回の試行回数を 10000 回, 1 つの方策で試行を繰り返す回数 500 回とする. アーム数を 3 本とし, 各アームの当たりの確率は, (0.2, 0.2, 0.75)と設定する. このときの試行回数 10000 回での最大累積報酬は 7500 となる. モデルに用いる設定として状態数は 3 として各状態で高い確率のアームが異なる設定を持ち, 1 つの状態を維持する時間を 100～1000 回の範囲で乱数を用いてランダムに設定する. 方策のパラメータは ε -グリーディ法の ε を 0.2, Q 学習の割引率 γ を 0.1,温度 β を 10.0, Q-Network の履歴数を 1,学習率 α を 0.8,温度 β を 100.0, 割引率 γ を 0.8 とする.

3. シミュレーション結果

3.1 各方策の累積報酬

図 1 は UCB 方策の各アームの UCB スコアと Q 学習の各アームの Q 値とモデルの状態推移である. これよりバンディット問題と強化学習の方策でスコアの減少に差があることがわかる. 図 2 は今回の方策に一般的なバンディット問題の設定とモデルを用いた場合での累積報酬の結果である. 結果, 左の一般的な場合には Thompson Sampling と UCB 方策が約 7500 と最も報酬を獲得し, Q 学習と Q-Network が約 7400 と報酬を獲得した. 右のモデルを適用した場合には Q 学習と Q-Network が約 7200 と最も報酬を獲得し, UCB 方策が約 7000 と報酬を獲得した.

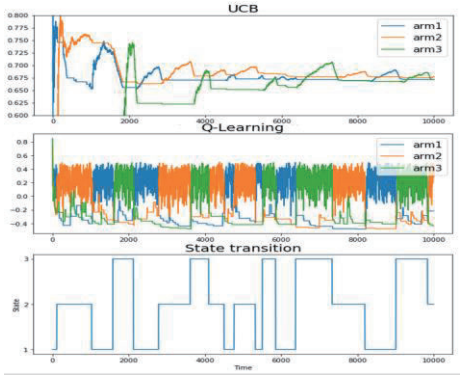

図 1　モデルの状態遷移

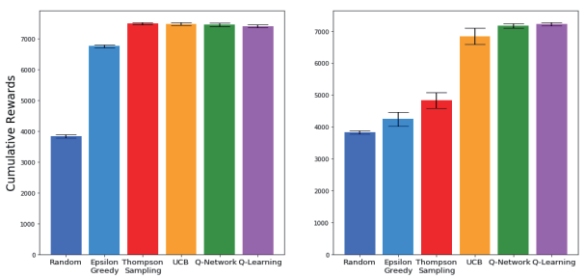

図 2　一般的な場合(左)とモデル(右)の方策の累積報酬

3.2 状態遷移回数の増加による累積報酬

今回のモデルは 1 つの状態を維持する時間が 100～1000 回と長く設定されている. この範囲を限定して状態遷移の回数を増加させた場合の各方策の累積報酬の変化をシミュレーションした. 図 3 は待機する時間を変化させた場合の各方策の累積報酬である. 横軸は維持する時間の最大値であり, 左側ほど状態遷移回数が多い. 結果, 1 つの状態を維持する時間の最大が 300 回以下でどの方策も累積報酬が大きく減少した.

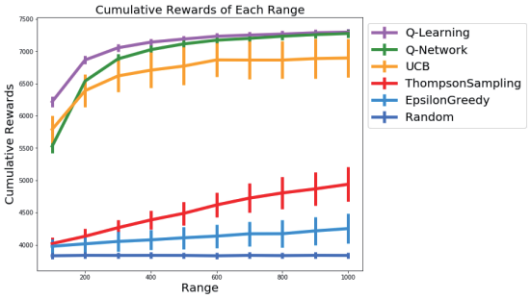

図 3　状態遷移回数の増加による各方策の累積報酬の変化

4. まとめ

一般的な設定とモデル適用時を比較すると Thompson Sampling に大きな差があった. これは Thompson Sampling が当たりの回数など報酬の履歴に影響してアームを選択しているため, 変化が起こるとその対応に大幅な時間がかかるためだと考えられる. その他にモデル適用時にはバンディット問題の方策よりも強化学習の方策の累積報酬が高い結果となった. これはアームの確率が変更し報酬が受け取りにくくなった場合に, Q 学習の更新式による Q 値の減少率が高く, またソフトマックス関数によって確率的にアームを選択することで他のアームを選択しやすくなるためだと考える. また待機時間を減少さていくとすべての方策で累積報酬が大きく減少した. 状態遷移の回数が増加することで累積報酬の高かった方策であっても状態を推定しきれなくなっていると考えられる.

今後の課題として, 状態が連続で変化していく待機時間が短い場合には高い累積報酬を獲得した方策がなかった. そのためこのような場合でも高い累積報酬が獲得できるような方策を考える必要がある. また今回のモデルは状態数や次に遷移する状態などの設定を限定した中で状態遷移を行っている. しかし, 日常的な問題には様々なシチュエーションや状態が存在する. よって様々な問題設定のモデルで各方策の累積報酬を明らかにしていく必要がある. 以上の 2 つにおいてシミュレーションしていく.

5. 参考文献

[1] 総務省. (2015). インターネットショッピングの利用状況. 参照先: 総務省 | 平成 27 年版 情報通信白書: http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nc122400.html

[2] Lihong Li, Wei Chu, John Langford, Robert E.Schapire. (2012). A Contextual-Bandit Approach to Personalized News Article Recommendation.

[3] 本多淳也, 中村篤祥. (2016). バンディット問題の理論とアルゴリズム. 講談社.

[4] Peter Auer, Nicolò Cesa-Bianchi, Paul Fischer. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning.

[5] Olivier Chapelle, Lihong Li. (2011). An Empirical Evaluation of Thompson Sampling