私のブックマーク

多腕バンディット問題[†1]

小宮山　純平（東京大学生産技術研究所）

1. はじめに

多腕バンディット問題（バンディット問題，multi-armed bandit problem）は，複数のアームと呼ばれる候補から最も良いものを逐次的に探す問題である．アームという奇妙な単語はこの問題のもとになったスロットマシン（バンディットマシン）の比喩から来ている．予測者はいくつかのスロットマシンを与えられ，それぞれのスロットマシンを引くと対応した報酬が得られる．繰り返す試行（アームの選択）を通じて得られる報酬を最大化するのが，予測者の目標である．報酬を最大化するという点で，バンディット問題は強化学習のカテゴリーに属する．実際，Suttonらによる強化学習のクラシックな教科書 [7] でも，バンディット問題は小節を割き説明されている．アームは，強化学習の分野ではアクションもしくはコントロールと呼ばれることがある．バンディット問題の予測者は，限られた試行回数において得られる総報酬を最大化したい．このとき，強化学習の普遍的なテーマである探索と活用のトレードオフが発生する．短期的には，現時点での経験期待報酬が高いアームを引くのが良い（情報の活用）．しかし，真の期待報酬が高いアームが，これまでたまたま運悪く少ない期待報酬しか得られていなかったということが低確率で発生する．これを防ぐためには，すべてのアームを全体的に引く必要がある（情報の探索）．予測者は，良いアルゴリズムに従って情報の活用と探索をバランスする必要がある．バンディット問題に関する研究は，主にこのアルゴリズムに関する研究と，どのような実問題をバンディット問題の枠組みに落とし込むかという研究に分かれる．本稿では，前者のアルゴリズム的な研究を主に説明する．これは，実問題をバンディット問題の枠組みに落とし込むとしても，アルゴリズム的にうまくいくかという直観は役に立つと考えるためである．

2. バンディット問題の定式化

バンディット問題には主に三つの定式化がある．つまり，割引定式化，確率的定式化，および敵対的定式化である．それぞれの定式化で，報酬に関する設定が異なり，枠組みが大きく変わる．バンディット問題の研究は非常に多いが，ほとんどの問題はこのいずれかの枠組みに収まる．

2·1　割引定式化（割引バンディット問題）

割引定式化においては，多くの強化学習問題と同じく未来の報酬を指数的に割引した総報酬を最大化することを目的とする．この定式化では，各アームは状態未知，遷移確率既知のマルコフ決定過程（Markov decision process, MDP）と考える．最初の状態に関する事前確率分布を与え，そのうえでの期待報酬の最大化を考える．割引定式化は，他の多くの文献ではベイズ的定式化もしくはベイズ的バンディット問題（Bayesian bandits）と呼ばれることが多い．しかし，ベイズ的であっても未来の報酬に割引因子を入れない場合の定式化ではむしろ後述する確率的バンディット問題に近いため，ここでは割引定式化という単語を使用する．この問題に対する最適なアルゴリズムはGittinsらによる指数によって特徴付けされる．つまり，各アームそれぞれの価値をGittins指数 [8],[31] で特徴付け，各ラウンドに最大のGittins指数をもつアームを選択することによって報酬を最大化することができる．そのため，このGittins指数が計算可能かどうかが割引バンディット問題の中心的な興味となる．

2·2　確率的定式化（確率的バンディット問題）

確率的バンディット問題では，割引率を使わず，現在と未来の報酬を同じ価値で扱う．そのため，どの程度の

†1 http://www.ai-gakkai.or.jp/my-bookmark_vol31-no5

ラウンド数が平均的に期待できるか不明な場合に適した設定となっている．この定式化では，各アームを分布既知，パラメータ未知の確率分布として扱う．この定式化では，パラメータに対するロバスト性（一貫性，strong consistency）をもつクラスのアルゴリズム [32] を主に考える．アルゴリズムの方針として，信頼上界による方法（Upper Confidence Bound：UCB 法 [32],[33]），事後確率サンプルによる方法（Thompson サンプリング [34]），そして経験尤度による方法（Minimum Empirical Divergence：MED 法 [35],[36]）が知られている．多くのアルゴリズムは，これらのいずれかをベースにしている．

2·3 敵対的定式化（敵対的バンディット問題）

敵対的バンディット問題では，確率的バンディット問題と同じく割引率がない総報酬の最大化を目指す．この問題では，敵対者（adversary）が決められた範囲内で報酬を自由に決めることができる．敵対者は予測者のアルゴリズムがどのようなものであるかを知った後に報酬を適応的に選択するため，良いアルゴリズムの提案は一見不可能であるように思える．しかし，Auer らによる指数重み（exponential weight）をベースとした乱択アルゴリズム（Exp3 [37]）は，最も総報酬の高いアームを引き続けた場合とほぼ同じ報酬が得られることが知られている．この問題は，エキスパートのあるオンライン学習 [11] と深い関係がある．

3. 最近の研究動向

著者のバックグラウンドは機械学習であるため，その他の分野（統計，オペレーションズリサーチ，経済学など）での動向についての調査は不十分であることをお断りする．以下は，資料を順番にあげていく．スライド→書籍→学会の順にリンクを紹介しているのは，とっつきやすい順という理由である．最新の研究について参照するには，会議のプロシーディングスを追う必要がある．しかし，数学的定式化に関してはバンディット問題に関するすべての論文を読む必要はなく，先ほどの三つの定式化の大枠を理解できれば十分である．それぞれの項目で，日本語の文献を最初に紹介し，次に英語の文献を紹介している．

3·1 紹介スライド・チュートリアルなど

スライド・チュートリアルは概観しやすいため，最初に紹介する．より正確な定式化などに関しては，次項に紹介する書籍などを参照されたい．

- 多腕バンディット問題の理論とアルゴリズム（PDF）[1]
  本多淳也による情報論的学習理論ワークショップ（IBIS 2014）でのチュートリアル．主に確率的バンディット問題のアルゴリズムについて解説している．確率的バンディット問題に関する数学的枠組みの説明に主眼が置かれている．
- バンディットの理論と応用（PDF）[2]
  中村篤祥による IBIS 2011 でのチュートリアル．三つの定式化すべてに少しずつ触れている．また，関連研究などに詳しい．
- Introduction to Bandits：Algorithms and Theory [3]
  J. Y. Audibert らによる ICML 2011 でのチュートリアル．主に確率的バンディットおよび敵対的バンディット問題に関して解説している．全体的に理論寄りであり，アーム数が多いときの定式化などに詳しい．解説動画 [4] も見られる．
- Bandit Processes and Index Policies（PDF）[5]
  R. Weber による割引定式化のチュートリアル．Weber は Gittins 指数の研究・チュートリアルを数多く出版している．

3·2 書籍・サーベイ論文など

- バンディット問題の理論とアルゴリズム [6]
  本多淳也，中村篤祥による日本語（でおそらく唯一？）のバンディット問題に関する書籍．確率的定式化および敵対的定式化について説明がされている．
- 強化学習 [7]
  R. Sutton らによる強化学習全般に関する書籍（三上貞芳らによる和訳書）．数学的精密さより強化学習の精神を説明することを重要視している．著名だが，1980 年初版の書籍なので現代ではやや古いかもしれない．

- Multi-armed Bandit Allocation Indices [8]
  J. Gittins，K. Glazebrook，R. Weber らによる割引定式化の書籍．
- Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems [9]
  S. Bubeck と N. C. Bianchi らによるバンディット問題のレビュー．理論寄りで，確率的定式化および敵対的定式化について説明がされている．S. Bubeck のホームページでほぼ同内容の PDF が入手可能である．
- Analyse de stratégies Bayésiennes et fréquentistes pour l’allocation séquentielle de ressources（PDF，一部フランス語）[10]
  E. Kaufmann の博士論文．確率的定式化のかなり理論寄りの内容．また，割引率のないベイズ的な設定に詳しい．
- Prediction，Learning，and Games [11]
  N. Cesa-Bianch らによるオンライン学習に関する書籍．難解，また表記などが読みづらい印象だが，これらのトピックを扱う本では最も良くまとまっている．オンライン学習のうち，エキスパートの助言が得られる問題設定は敵対的設定におけるバンディット問題と深い関係がある．
- Bandit problems：Sequential Allocation of Experiments [12]
  D. A. Berry らによるバンディット問題に関する書籍．1985 初版から内容が更新されていないのでやや古いように見える．
- A Survey of Monte Carlo Tree Search Methods [13]
  C. B. Browne らによるモンテカルロ木探索に関するサーベイ．モンテカルロ木探索はバンディット問題の最も有力な応用先の一つであり，囲碁などのゲーム AI で成功を収めている．また，モンテカルロ木探索は竹内聖悟による「私のブックマーク：ゲームプログラミング（将棋を中心に）」[14] でも紹介されている．

3·3 会　　議

バンディット問題に関する最新の研究は，基本的に国際会議を調べる必要がある．しかし，基礎的な枠組みを知るには下記「重要な論文」の項と同等の内容を理解すれば十分であり，国際会議での論文の多くはこれらの論文によって確立された枠組みを踏襲している．

- Neural Information Processing System（NIPS）[15]
- International Conference on Machine Learning（ICML）[16]
- International Conference on Artificial Intelligence and Statistics（AISTATS）[17]
  機械学習に関する国際会議．機械学習の会議は，概してデータからの学習手法に関する研究が多いが，実問題を解いた論文なども少なくない．近年，バンディット問題に関するアルゴリズム提案や理論解析の論文が多く提案されている．バンディット問題は，オンライン学習・学習理論などのセッションに位置付けられるか，それ自身のセッションが設置されることが多い．
- Conference On Learing Theory（COLT）[18]
  機械学習・学習理論に関する国際会議．NIPS/ICML より理論的な内容が好まれる傾向にあり，証明のない論文は扱わない．オンライン学習・バンディット問題の論文は多い．
- Journal of Machine Learing Research（JMLR）[19]
- Machine Learning（Journal）[20]
  機械学習に関する国際ジャーナル．上記の機械学習系国際会議で採択された論文のジャーナル版が採択されることも多い．
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining（KDD）[21]
- IEEE International Conference on Data Mining（ICDM）[22]
- International Conference on Web Search and Data Mining（WSDM）[23]
  データマイニング・知識発見に関する国際会議．バンディット問題のアルゴリズムをデータマイニングの問題へと応用する研究が多く発表されている．機械学習とデータマイニングは強く関連した分野である．機械学習との違いは，機械学習はデータや既存のモデルに対する学習可能性に関して主に興味があることに対し，データマイニングでは与えられた問題に対する興味をいかにして解決するかということに主眼が置かれていることである．
- ACM Recommender Systems Conference（RecSys）[24]
  推薦システムに関する国際会議．バンディットアルゴリズムは，推薦における重要な課題であるコールドスタート（情報が少ないときに推薦の精度が低くなってしまう問題）に対する一つの対応策として位置付けられている．

その他，人工知能に関する国際会議（AAAI [25]，IJCAI [26]）や情報検索に関する国際会議（SIGIR[27]）でも，バンディット問題に関する論文が近年見られる．人工知能に関する国際会議では，主に応用可能性などが重視されている印象がある．

3·4 ソフトウェアなど

バンディット問題に関するソフトウェアは，著者の知る限りそれほど多くない．以下にいくつかのライブラリをあげる．もともと，バンディット問題に関するアルゴリズムそのものは比較的単純で，たかだか数十〜数千行程度のプログラムで実装可能であるので，自分でUCBなどのアルゴリズムを書いてみるのも良いかもしれない．実際，多くの研究者は自分でライブラリを実装しているのではないかと考えられる．

- BanditLib [28]
  著者による確率的バンディット問題のライブラリ．C++で書かれており，コンパクトである．
- pymaBandits [29]
  フランス国立情報学自動制御研究所（INRIA）の研究者による確率的バンディット問題のライブラリ．Python/Matlabで書かれており，コンパクトである．
- Jubatus [30]
  PFN/NTT社によって開発されているオンライン分散学習に関するライブラリ．バンディット問題のモジュールがあり，UCBやThompsonサンプリングなどのアルゴリズムが実装されている．

3·5 重 要 な 論 文

バンディット問題の論文は数多いが，以下の論文相当の内容を理解すれば最低限の理解は得られる．割引定式化に関しては，Gittins指数の概念を理解すればよい．例えば，Weberによる論文 [31] はこの点についてまとまっている．確率的定式化に関しては，Laiらの論文 [32] によって導入された一貫性の概念が本質的である．もっとも，小難しいRegretの証明などが不要な場合は，UCB1 [32] などのアルゴリズムを実装して試してみると良いかもしれない．敵対的バンディット問題の枠組みは，Exp3アルゴリズム [37] が，エキスパートのあるオンライン学習における指数重みアルゴリズム＋報酬の不偏推定量であるという仕組みを理解できればよい．本多淳也・中村篤祥による日本語書籍 [6] は，確率的定式化および敵対的定式化をカバーしている．

- On the Gittins Index for Multiarmed Bandits [31]
  R. Weberによる Gittins 指数に関する論文．
- Asymptotically efficient adaptive allocation rules [32]
  T. L. Laiらによる確率的バンディット問題に関する最も基本的な論文．確率的バンディット問題に関するアルゴリズムはこの論文の枠組みをベースにしている．
- Finite-time Analysis of the Multiarmed Bandit Problem [33]
  P. Auerらによる UCB1 アルゴリズムなどの提案論文．確率的バンディット問題に関するアルゴリズムはUCB1をベースとしているものが多い．
- On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples [34]
  W. R. Thompsonらによる Thompson サンプリングアルゴリズムのもとになったアイディアのある論文．この論文自体は非常に古いが，Thompsonサンプリングは現代でも良く使われている．
- An Asymptotically Optimal Bandit Algorithm for Bounded Support Models [35]
- Non-Asymptotic Analysis of a New Bandit Algorithm for Semi-Bounded Rewards [36]
  本多淳也らによる Minimum Empirical Divergence（MED）アルゴリズムの論文．確率的バンディット問題に関するアルゴリズムはUCB，Thompsonサンプリング，MEDのいずれかをベースにしていることが多い．
- The Nonstochastic Multiarmed Bandit Problem [37]
  P. Auerらによる敵対的バンディット問題に関する最も基本的な論文．敵対的バンディット問題に関するアルゴリズムの多くはこの論文で提案されたExp3と同じく指数重みと報酬に関する不偏推定量を利用している．

4. お わ り に

本稿では，バンディット問題に関する論文を紹介した．バンディット問題は統計・オペレーションズリサーチの分

野で研究されてきた非常に基本的な問題であり，近年は機械学習の分野でも Web 広告などへの応用を想定して広く研究が行われている．現代でもその理論面は十分考察に値する．また, 応用面でもモンテカルロ木探索や広告のクリック率最適化など，いくつかの重要な適用先がある．本稿がこの問題に関する理解の一助となれば幸いである．

[1] http://ibisml.org/archive/ibis2014/ibis2014_bandit.pdf
[2] http://ibisml.org/archive/ibis2011/ibis2011-nakamura.pdf
[3] https://sites.google.com/site/banditstutorial/
[4] http://techtalks.tv/talks/tutorial-introduction-to-bandits-algorithms-and-theory/54451/
[5] http://www.statslab.cam.ac.uk/~rrw1/talks/YETQweber2013.pdf
[6] https://www.amazon.co.jp/dp/406152917X/
[7] https://www.amazon.co.jp/dp/4627826613
[8] https://www.amazon.co.jp/dp/0470670029/
[9] http://www.nowpublishers.com/article/Details/MAL-024
[10] http://chercheurs.lille.inria.fr/ekaufman/TheseEmilie.pdf
[11] http://dl.acm.org/citation.cfm?id=1137817
[12] https://www.amazon.co.jp/dp/9401537135
[13] http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6145622
[14] http://www.ai-gakkai.or.jp/my-bookmark_vol30-no2/
[15] https://nips.cc/
[16] http://icml.cc/
[17] http://www.aistats.org/
[18] http://www.learningtheory.org/
[19] http://www.jmlr.org/
[20] http://link.springer.com/journal/10994
[21] http://www.kdd.org/
[22] http://www.cs.uvm.edu/~icdm/
[23] http://www.wsdm-conference.org/
[24] https://recsys.acm.org/
[25] http://www.aaai.org/home.html
[26] http://ijcai.org/
[27] http://sigir.org/
[28] https://github.com/jkomiyama/banditlib
[29] http://mloss.org/software/view/415/
[30] http://jubat.us/
[31] https://projecteuclid.org/euclid.aoap/1177005588
[32] http://www.sciencedirect.com/science/article/pii/0196885885900028
[33] http://link.springer.com/article/10.1023/A:1013689704352
[34] https://www.jstor.org/stable/2332286
[35] http://colt2010.haifa.il.ibm.com/papers/29honda.pdf
[36] http://jmlr.org/papers/v16/honda15a.html
[37] http://epubs.siam.org/doi/abs/10.1137/S0097539701398375