連載

5分で分かる!? 有名論文ナナメ読み

黒木祐子（東京大学）

Chen, S. et al.:
Combinatorial Pure Exploration of Multi-armed Bandits

Advances in Neural Information Processing Systems, pp.379-387 (2014)

逐次的意思決定問題

ビア・バーの開業を考えているオーナーがいるとしよう．このオーナーは最高のクラフトビールを20種類ほど揃えて提供したいと考えている．ところがクラフトビールの醸造所は日本だけでもおおよそ500個所以上もあるという．そしてビールというものはどうやら繊細な生き物で，日光，保管期間，温度，輸送中の衝撃，酸素などさまざまな要因で風味が変化してしまう．一方で開業までの準備期間と予算は限られている．そこでオーナーは，まずは仕入れ価格などの条件から100種類に候補を絞り，この中でテイストテストをすることにした．ただし試飲の際は鮮度の問題で多少風味が変化することを踏まえ，100種類をすべて1回だけ試飲して決定するのではなく，テイストテストの実施状況に応じて物によっては1回以上試飲してデータを取る必要も出てき得る．なるべく早くオーナーの納得のいく決定をするにはどうしたらよいのか．オーナーの頭を悩ますこの逐次的な意思決定問題は，本稿で紹介する組合せバンディット問題の一種として捉えることができる．

組合せバンディット問題

「組合せ」がついていない通常の「バンディット問題☆1」の数式込みの詳細な解説についてはぜひとも文献1）などを参照されたいが，本稿ではダイレクトに組合せバンディット問題を導入する．先ほどのオーナーのテイストテストの問題は，報酬が未知である100個の候補から上位20個を探し出すのが目的である．直感的には試行（試飲）回数を大量に取れば，上位20個を確実に見つけ出すことに成功しそうであるが，予算や時間には限りがある．一方であまりにも少ない試行回数では，探索が不十分で必ずしもベストな選択を見つけられない．では，試行回数は最低限に抑えながら，上位20個を確実に識別する解法（アルゴリズム）は何か？　ここでアルゴリズムの評価指標は，出力までに必要とした試行回数であり専門的には標本複雑度と呼ぶ．この意思決定問題を特に「組合せ最適腕識別問題☆2」と最初に呼び，多くの組合せ的な意思決定を扱える統一的な枠組みを提案したのが，本稿で取り上げるChenらによる論文である．

本論文の背景をより理解するために実世界で現れる組合せ的な意思決定の例を見てみよう．先ほどのビア・バーの例では「20種類のビール」が意思決定を特徴付けていたが，一般に報酬が未知のn個のアイテムの中から上位個のアイテムを探索する問題と見ることができ，オンライン広告におけるキーワード集合の選択も同様の定式化になる．グラフ・

☆1　多腕をつけた「多腕バンディット問題」が正式な名称である．複数の行動候補をスロットマシンのアーム（腕）に見立てたギャンブラーのモデルからこのような呼び名になっている．

☆2　「最適腕識別」に対して「リグレット最小化」と呼ばれる設定では，期間Tに対して神のみが知っている最適な意思決定と，実際に選んだ意思決定との良さの差＝（リグレット）のT期間における累積和を最小にする方法を議論する．たとえばオーナーがすでに開業済みで今の意思決定がただちにお店の利益にかかわる場合は，純粋なビール探索に加えて目先の利益とのトレードオフを考慮する必要がある．組合せバンディットのリグレット最小化を扱う代表論文として文献2）を挙げておく．2つの設定の難しさは独立である．

ネットワーク上においてはさらに複雑な意思決定もあり得るだろう．たとえば新たに開設した道路ネットワーク上の最短経路を探索する問題や，通信ネットワークにおいて最小コストで伝達通信を可能にする全域木☆3を探索する問題，そしてクラウドソーシングにおける労働者と仕事の最適な割り当てを探索する問題など，「組合せ的な意思決定」を要する場面は実世界に多く存在しており，それらはしばしばコストや効用について不確実性を伴う．

所与の組合せ制約の元で，ある目的関数を最大（最小）にするような解を求める数理技術は「組合せ最適化」と呼ばれているが，伝統的な枠組みでは不確実性を逐次的に対処する方法論が確立されていなかった．組合せ最適腕識別問題では不確実性を解消するために，データの収集とそれを最適な組合せ的意思決定に反映させるための方法を議論しているのである．

問題の定式化とCLUCBアルゴリズム

n 個の行動候補が与えられる（最初の例では各ビールが各行動に対応）．各行動の報酬は未知の確率分布と対応し，期待報酬が確率分布の期待値となる．t 回目の試行で，ある行動 i_t を選ぶとそれに対応する確率分布から報酬がノイズ付きで観測される．たとえば通信ネットワーク上で選んだ区間の移動にかかったコストやクラウドソーシング上での割り当てた労働者と仕事における効用にあたる．注意しておきたいのが行動候補は n 個であるが最終的な意思決定の全候補 $\mathcal{X}$ のサイズは指数サイズである（たとえば n 個から k 個取り出す組合せは $nCk = O(n^k)$）．そこで効用・コストを固定した最適化問題を多項式時間で解くオラクルの利用を仮定する☆4．

Chen らは CLUCB（Combinatorial Lower Upper Confidence Bound）という手法を提案した（図 -1 参照）．図 -1 の例において各行動は各枝に対応し，その報酬は枝重みに対応する．各行動 i の報酬の推定値は観測に基づく平均をとり，その信頼区間 $\mathrm{rad}_t(i)$ を適切に考える．t 期までの各行動の報酬の推定値 $w_t(i)$ のもとで最適な意思決定 $M_t \in \mathcal{X}$ をオラクルで求め，M_t に含まれなかった行動 i についてはUCB，つまり $\widetilde{w}_t(i) = w_t(i) + \mathrm{rad}_t(i)$，$M_t$ に含まれている i についてはLCB，つまり $\widetilde{w}_t(i) = w_t(i) - \mathrm{rad}_t(i)$ としてこの重みのもとで再度オラクルを呼び出し

☆3 全域木とは，元のグラフのすべての頂点を含み，選んだ辺集合に閉路を含まない連結グラフ．

☆4 たとえば最小全域木問題，最大重み付きマッチング問題など多項式時間アルゴリズムが知られている．

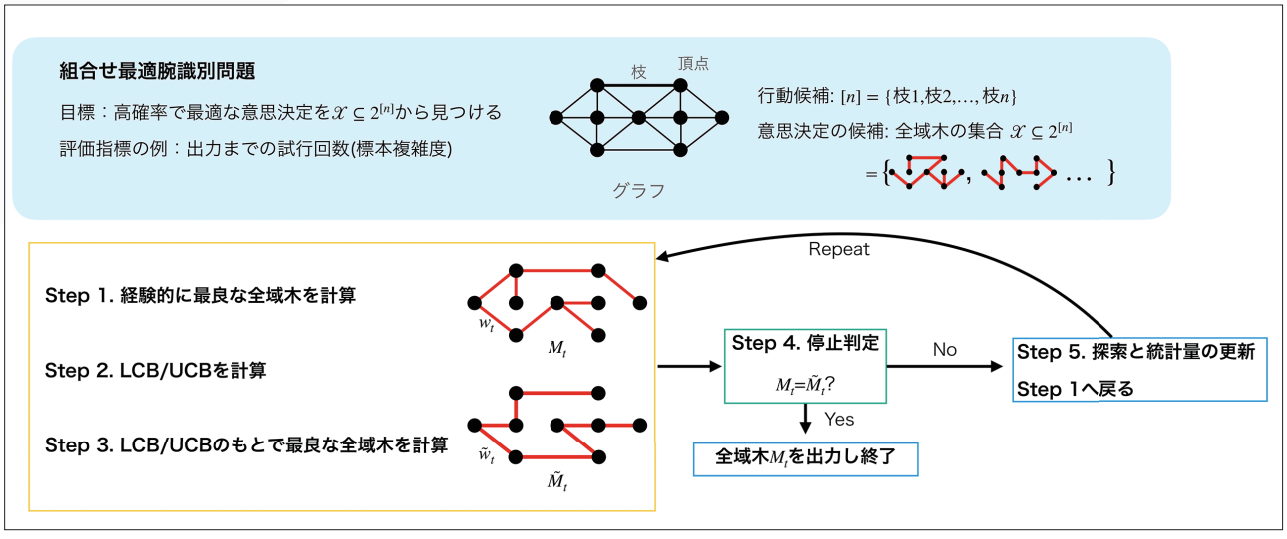

図 -1　CLUCB の流れ (全域木の場合)

$\widetilde{M}_t$ を得る．M_t と $\widetilde{M}_t$ の推定報酬がほぼ一致したら探索は十分と判断し，経験的に最適な M_t を出力する．そうでなければ2つの差集合の中で最も不確実性の高い i_t を選択し試行する．つまり信頼区間の上下界を考慮した解との比較により現時点で最良な解の良さと次に選択する行動を決定するアイディアである．

この手法についてChenらは標本複雑度の理論解析を背後の組合せ構造によって定まるパラメータに依存する形で与えた．さらに下界として，どんなに賢いアルゴリズムを用いてもこの組合せ最適腕識別問題で最低限必要となる試行回数を示した．また，試行回数の上限が予算として初めに与えられ，失敗する確率を最小にする別の設定でも同様の結果を与えている．いずれの設定でもその後の後続研究のベースとなる枠組みとなっている．

本論文以降の発展と課題

Chenらの論文を皮切りに，さまざまな問題設定の拡張が提案された．1つの方向に背後の組合せ構造に仮定を置き，よりタイトな標本複雑度の上界を証明するものがある．たとえば組合せ構造をマトロイド基と呼ばれるある種美しい数理構造を持ったものに限定した場合はよりタイトな結果が知られている．2つ目としては「限定された観測しか得られない」より挑戦的な設定への拡張がある．この場合指数的に大きい行動空間を扱う計算量的困難性が大きなハードルである．3つ目には「非線形な報酬関数」を扱う技術開発も重要である．というのも最初のビールの例では20種類の効用の単なる和ではなくその組合せによって全体の効用が定まるべきである．このようにまだ解決すべき課題は多くある．組合せ最適腕識別は「不確実性を考慮した組合せ最適化問題」と「組合せ的な逐次的学習問題」の両面を併せ持つ興味深いトピックであり，今後の理論発展にぜひ注目したい．

参考文献

1) 本多淳也，中村篤祥：バンディット問題の理論とアルゴリズム（機械学習プロフェッショナルシリーズ）．
2) Chen, W., Wang, Y., Yang, Y., : Combinatorial Multi-Armed Bandit : General Framework and Applications, Proceedings of the 30th International Conference on Machine Learning, PMLR 28, pp.151-159 (2013).

（2022年1月31日受付）

黒木 祐子（正会員）
yukok@is.s.u-tokyo.ac.jp
2021年東京大学大学院博士課程修了．博士（情報理工学）．2021年より同大学助教．理化学研究所革新知能統合研究センター客員研究員．統計的機械学習や組合せ最適化にかかわる数理的問題へのアルゴリズム研究に従事．