学習失敗リスクを考慮した学習用教材選択のリスク考慮型強化学習によるモデル化

江原 遥
Yo EHARA
静岡理工科大学情報学部
Department of Informatics, Shizuoka Institute of Science and Technology
Email: ehara.yo@sist.ac.jp

あらまし： 本稿では、外国語語彙学習に適したテキスト選択のように、学習者に適した教材を自動的に選択し推薦する問題を考える。難度の高い教材は、意欲を損なうなど学習に失敗するリスクが高い一方、既習事項が少なく、学習に成功すれば高い学習効果が得られると期待され、ハイリスクハイリターンである。難度の低い教材は、その逆である。本稿では、適切な教材を複数選ぶ問題が、リスク考慮型強化学習（分布強化学習）という機械学習の問題として自然にモデル化できることを示す。

1 はじめに

本稿では、学習者に適した教材を自動的に選択し学習者に逐次的に推薦していく問題を考える。具体的には、下記のプロセスを繰り返すことを考える。

1. システムが学習者の現在の能力に合わせて所与の教材の中から自動的に教材を選択し、学習者に対して推薦する。
2. 学習者はシステムが選択した教材に挑戦し、学習に成功すれば能力が向上することもあり得るが、学習に失敗することもあり得る。
3. システムの目的は、直近の回だけでなく将来的な能力の総向上量を最大化する事である。

上記のプロセスは、相当に一般的な枠組みであり、多くの教材推薦がこの枠組みを用いて記述される。例えば、語彙学習の例でいえば、読解教材用のテキストを選択する多読支援システムが挙げられる。この場合、学習の「成功」はテキストを読みこなせた（内容の理解にまで至った）、学習の「失敗」はテキストの読解に失敗したことに対応し、「能力の向上」は例えば読解によって獲得した（と推定される）語彙量などが相当する。あるいは、語学学習ではなく数学やプログラミングなどの教育においても、「教材」の単位や学習の成功/失敗の定義を明確にすれば、上記の枠組みで教材推薦の過程をモデル化していくことが可能である。

また、上記のプロセスは、個人化学習支援と、クラスなどの複数の学習者の単位（学習者集合）への推薦の両方をモデル化できていると考えられる。1. のステップで学習者個人のテスト結果などの特性を考慮して教材を選択する場合は個人化学習支援のモデルと捉えられるし、学習者の属性（学年など）だけを考慮するようにして、「学習者」をクラスなどの「学習者の集合」に置き換えて考えれば、各クラスの進捗に合わせた教材推薦をモデル化しているともとらえられる。また、本稿では一貫性のため「教材」という言葉で統一しているが、推薦する単位を「教材」と呼んでいるだけであり、例えば用語の説明や適切な計算問題、指示・指導なども一種の教材と考え同じ枠組みでモデル化できる。

上記の枠組みは強化学習（reinforcement learning）としてみることができる。強化学習は幅広いモデルを内包する枠組みだが、一般に、意思決定を行うエージェント（agent）と環境（environment）からなる。環境には状態（state）があり、逐次的に環境が変化する。エージェントが逐次的に行動（action）を選択すると、環境から報酬（reward）が返され、報酬を通じて行動の適切性を判断することができる。強化学習では、次の時点だけではなく将来にわたる報酬の累積値を価値観数（Value Function）としてモデル化し、この価値観数を最大化するようにエージェントを学習する。上記の例における「システム」がエージェント（意思決定主体）、「学習者」が強化学習における環境に対応し、「学習者の（現在の）能力」が状態に対応し、教材選択が行動（action）に対応し、学習者の能力向上が報酬に対応する。

学習者に対する教材推薦の過程を強化学習でモデル化する事には、いくつかの利点がある。まず、強化学習は、直近の学習効果だけを最大化しようとするのではなく、将来的な学習効果（報酬）の累積総和の期待値をモデル化し、これを最大化しようとする。「将来的な学習効果」は測定が難しいが、教育において本質的に重要であり、これを陽にモデル化していることは重要である。また、直近の学習効果を最大化する問題も、将来的な報酬を大きく割り引くことで表現できるため、強化学習によるモデル化の方がより一般的なモデル化になっていると言える。

次に、強化学習によるモデルは機械学習分野で理論的な性質が研究され、モデルがよく分類されている。特に、教育における教材推薦においては、「リスク考慮型強化学習」という種類の強化学習モデルが従来の教材推薦の場面ではあまり考えられてこなかった「リスク」をうまくモデル化できているように思われる。これは、単純に将

来的な学習効果（報酬）の累積総和の期待値を最大化しようとするのではなく、学習効果がどの程度安定的に得られるか（リスク）をも考慮して、適切な行動選択（教材推薦）を行おうとするモデルである。

例えば、前述の読解教材選択のタスクであれば、「難しく大抵の場合読めないが、まれに読解に成功すれば語彙量を一気に増やすことが可能なテキスト」というものが存在するかもしれない。学習効果（報酬）の期待値だけ考慮するのであれば、そうしたテキストでも期待値が高ければ選択されうる。しかし、学習においてそうしたテキストを積極的に選択したいかといえば、通常はそうはならないと考えられる。実際に学習に成功したごく一部の学習者については高い学習効果が認められたとしても、大半の学習者にとって学習効果が認められない教材であるならば、公平性の観点から、そうした教材は通常避けられる。このような「この学習者にはまだ早すぎる」、「成功したら素晴らしいが危険すぎる」という「リスク」を考慮して教材を選択することは、従来は教員の直観に頼るところが多かったと思われる。強化学習を用いて、この直観を適切にモデル化できれば、これまで人間の教員が行ってきたような学習者の力量を総合的に考慮した高度な教材選択を自動化することが可能になるかもしれない。

そこで、本稿では、リスク考慮型強化学習を用いて教育分野における推薦過程をモデル化できることを示す。強化学習においては、このように、期待値だけを考慮する方法は安定的に良い結果を得られない場合があることが知られている。このような場合に、価値観数として期待値だけではなくリスクを考慮する「リスク考慮型強化学習」が知られている [5]。リスク考慮型強化学習は「分布強化学習」の一種であり、累積報酬の確率分布の期待値だけではなく確率分布全体の形を考慮して行動を決定する手法とみなせる。

2　関連研究

筆者は [2, 1] において、読解用教材選択の応用において、機械学習の識別器の結果を統合 (aggregate) して、簡単な単語テスト結果から学習者がテキストの読解に成功する確率分布を与える手法を提案した。この研究では、確率値を考慮することによって、学習者がテキストの読解に成功するかどうかを単語テスト結果だけから予測する予測精度が実際に向上することを確かめたが、この読解成功確率の確率分布を精度向上以外の方向性で有効活用する方法については述べられていなかった。

3　提案するモデル化

[5] に従って表記を定義する。学習者の状態集合を $\mathcal{S}$、取りえる行動の集合を $\mathcal{A}$ とし、学習者が状態 s にある時にどのような行動を取ればよいかという方策を確率的に表す関数を $\pi(a|s)$ とする。方策 π に基づくマルコフ決定過程を $M(\pi)$ とする。時刻 t の学習者の状態を S_t、割引率 γ での割引累積報酬（リターン）を $C_t := \lim_{K\to\infty} \sum_{k=0}^{K} \gamma^k R_{t+k}$ とする。この時、方策 π のもとでのリターン分布は次の形で定義される。

$$P_C^\pi(c|s) := \Pr(C_t \le c | S_t = s, M(\pi)) \tag{1}$$

今、リターン分布の推定は可能であるとする。例えば、学習者の単語テストから推定される状態から、学習者が所与のテキストの読解に成功する確率分布は、リターン分布の一種とみなせ、[1] の方法で推定可能である。この時、リターン分布の期待値の最大化だけを目的とするのがリスクを考慮しない強化学習である。[5] でも述べられているように、何を「リスク」として定義するかは様々なモデル化の方法がある。例えば、リターン分布の分散をリスクとしてとらえる方法は、現代ポートフォリオ理論 [4] として知られている。現代ポートフォリオ理論とリスク考慮型強化学習の関連性については [3] が詳しい。

リスクをどのように定義したとしても、同じリスクであればリターンが大きい方が望ましいことは明白である。このように、許容できるリスクを最初に定め、許容できるリスクのもとでのリターン（Value-at-Risk, VaR）を最大化するモデル化を行えば、リスクを考慮した意思決定を行ったと言える。[5] では、VaR を最大化する具体的な方法として、リターン分布の q 分位点（確率 q で起こりえる最小リターン値）を最大化する方法を紹介している。具体的には、(2) で表されるリターン分布の q 分位点を、(3) のように最大化する方策を探すモデル化と、これを実際に解くための論文が紹介されている。ただし、Π は考えられる方策の集合である。

$$\mathcal{Q}_q[P_C^\pi|s] := \inf_{c\in\mathbb{R}} \{P_C^\pi(c|s) \ge q\} \tag{2}$$

$$\max_{\pi\in\Pi} \sum_{s\in\mathcal{S}} \mathcal{Q}_q[P_C^\pi|s] \tag{3}$$

4　おわりに

本稿では、「この学習者には、まだこの教材は難しいかもしれない」といった、人間の教員が日常的に行っているリスクを考慮した教材選択の過程が、リスク考慮型強化学習の枠組みで自然にモデル化できることを示した。今後の研究としては、具体的な教材推薦の実験設定でシミュレーションを交えるなどして実験し、リスク考慮型強化学習がどの程度適切な教材推薦を行えるかを定量的・定性的に計測することが挙げられる。

参考文献

(1) Yo Ehara. Uncertainty-Aware Personalized Readability Assessments for Second Language Learners. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1909–1916, December 2019.
(2) Yo Ehara. テキストカバー率の確率的拡張に基づく語彙テストのみからの個人化読解判定. 2019.
(3) Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, Vol. 16, No. 1, pp. 1437–1480, 2015.
(4) H.M Markowitz. Portfolio selection. *The Journal of Finance*, Vol. 7, No. 1, p. 77–91, 1952.
(5) Tetsuro Morimura. 強化学習 (機械学習プロフェッショナルシリーズ). 講談社, 2019.