F-034

強化学習を用いたデータ提供者戦略最適化とデータ取引市場への影響評価
Optimization of Data Provider Strategies Using Reinforcement Learning and Evaluation of Its Impact on the Data Trading Market

山本 健太† 早矢仕 晃章†
Kenta Yamamoto Teruaki Hayashi

1. はじめに

近年、組織に蓄積されたデータの交換・取引・流通が新たなイノベーションの源泉として注目を集め、異分野間でデータ売買を行うデータ取引市場の整備が進展している[1]。しかし、取引対象データの品質保証や流通プロセスを規律する制度的枠組みは未だ十分ではない。そこで本研究では、データ取引市場に流通するデータ品質向上を目的としたレビュー制度導入の有用性を検証する。さらにその状況下でデータ提供者の最適な戦略について強化学習を用いて分析を行う。市場にレビュー機能が導入されると、提供者は自身の評判や利益の最大化のため、高品質かつ高価格なデータを市場に供給する戦略を選択すると仮定し、強化学習を用いて市場参加者の戦略形成プロセスをシミュレーションした。実験の結果、レビュー制度下においてデータ提供者が長期的報酬を最適化する行動を学習し、個別の取引利益の向上と併せて市場全体の活性化が促進されることが分かった。これにより、高品質なデータの流通を促進するレビュー制度の有用性ならびに、データ取引市場シミュレーションと強化学習の親和性についての示唆を得られた。

2. 関連研究と本研究のアプローチ

近年、インターネットの普及に伴い、口コミ情報がビジネスや市場に大きな影響を与えることが分かってきている。特に市場内の提供者の評判や購入者との信頼関係が重要と報告されており[3]、データ取引市場においても同様に、データ提供者が自身の評判を意識し、購入者との信頼構築を促進するメカニズムの構築が不可欠であると考えられる。

データ取引市場は萌芽的段階にあり、実運用下で観測可能な取引事象や市場動向の把握は著しく限定されている。このような場合、市場内の参加者をエージェントとしてモデル化した市場シミュレータの開発とその利用による解析が有効となる[2]。関連研究として、金融市場を対象に強化学習を用いた市場シミュレーションが報告されている。これらの研究では、金融取引を逐次意思決定問題として捉え、市場参加者に利益最大化行動を強化学習により獲得させることで、その有効性を検証している。データ取引市場においても、参加者は自身の長期的利益を最大化する戦略を取ることが想定される。そこで本研究では、価値ベースの Q 学習手法を導入し、市場参加者が長期的効用を最大化する戦略を学習するシミュレーション実験を行った。なお、データ取引市場における参加者の役割や行動は未だ十分に解明されていないため、本研究では市場構造を簡略化し、提供者と購入者の 2 者モデルを採用することで両者の戦略的相互作用が市場ダイナミクスに与える影響を分析した。

† 東京大学大学院工学系研究科システム創成学専攻
Department of Systems innovation, School of Engineering, The University of Tokyo

3. 提案手法と実験概要

3.1 提案手法

強化学習は価値ベースと方策ベースに大別され、本研究では価値ベースの Q 学習を採用する。価値ベース手法はモデルフリーであり、未知の市場ダイナミクス下でも汎用的に適用可能であるため、観測可能事象が限定的なデータ取引市場に適合すると判断した[4]。さらに、本研究では市場にレビュー制度を導入し、市場に提供されるデータ品質に応じて提供者のレビュー値が更新される仕組みを構築した。市場参加者はその値を考慮して行動選択するよう設計した。

3.2 市場参加者の戦略モデル

本シミュレーションでは、取引対象データの価格を 40~80、品質を 0.4~1.0 の範囲に設定した。市場参加者は提供者と購入者に大別され、提供者は長期戦略と短期戦略の 2つを選択する。長期戦略では、高品質(0.6~1.0)かつ高価格(60~100)なデータを供給し、購入者との長期的な信頼関係構築を重視する。短期戦略では、低品質(0.4~0.8)かつ低価格(40~80)を提供し、短期的利益獲得を優先する。データ購入者は、高品質かつ低価格データを購入する戦略と、低品質かつ高価格データも許容して購入する戦略を選択する。購入者は提供者のレビュー値が高い場合に限り、品質と価格のトレードオフをある程度許容し、データが高くても購入条件を緩和する。

3.3 強化学習アルゴリズム

提供者の状態は平均利益、平均取引量、レビュー値とし、長期と短期の行動戦略を取り得る。報酬は提供者の効用、購入者の効用、そして取引成立の有無で与える。購入者の状態は、購入成功率と、購入者の効用を設定し、3.2 に記載の 2 つの行動戦略を取り得る。報酬は購入者の効用で与えている。提供者と購入者の効用関数は以下の式で与える。

$$U_P = P(\alpha Q + \beta R) - C \quad (1)$$

$$U_b = Q - 0.01P + \gamma R \quad (2)$$

ここで、P: 価格、Q: 品質、R: レビュー値、$C(=Q \times Q)$: コストである。係数$\alpha = 0.47, \beta = 0.35, \gamma = 0.76$は Kaggle プラットフォーム上のメータデータの分析結果をもとに経験的に設定した。

また、強化学習のベルマン更新式は以下の式(3)とした。

$$Q^{new}(s,a) \leftarrow Q(s,a) - \alpha[r + \gamma max Q(s',a') - Q(s,a)] \quad (3)$$

ここで、学習率は初期値を実験的に安定した学習が行われるよう先行研究より$\alpha = 0.1, \gamma = 0.9$とし、探索率は$\varepsilon$グリーディ探索を用いて、初期状態から指数的に減衰させた[5]。

3.4 レビュー値 (評判概念)

レビュー値は次の手順で更新する。各提供者はまず正規分布からサンプリングした値を取得し、0.25 刻みに丸めた上で正規化し、初期レビュー値を設定する。その後、各ラウンドで提供者が販売したデータの品質値と、過去のレビュー値を用いて、正規化したスコアを得る。続いて、前ラウンドの値と暫定スコアの折衷を係数 C(本実験では 0.5)を用いて再度正規化し、新しいレビュー値を得る[6]。

3.5 シミュレーション設定

初めに各データ提供者・購入者にランダムに初期戦略を割り当てる。提供者は、自身の戦略に基づき、価格と品質の組を探索的に最適化し、データを生成する。購入者は Q 学習に基づいて購買戦略を選択し、提供者の全オファーから自身の効用を最大化し、かつ閾値条件を満たす提供者を選択して取引を成立される。そして、提供者と購入者は強化学習を用いて学習を更新し、レビュー値を更新される。

4. 結果と考察

図 1 および 2 は実験で得られた各ラウンドの平均市場収益性と取引量の推移の結果を表す。この結果から、長期戦略をとったデータ提供者の方が短期戦略をとった提供者よりも市場収益性も取引量のどちらも高いことが分かる。

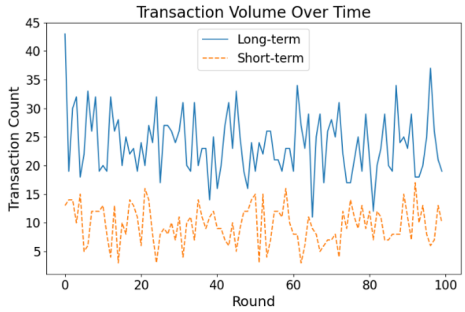

図 1 収益性推移

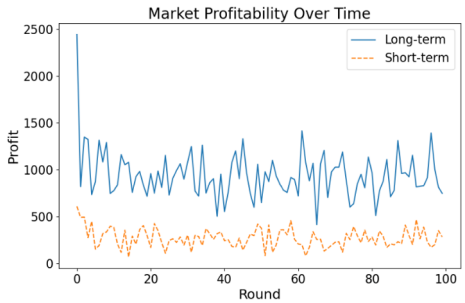

図 2 取引量推移

続いて、市場に流通するデータの品質の各ラウンドの平均値の推移について比較を行った。図 3 はレビュー制度が存在する市場のデータ品質値の推移であり、図 4 は存在しない市場の推移を表す。この結果から、レビュー制度が存在する場合の方が、市場に流通するデータの品質平均は存在しない市場よりもデータ品質値が上回っていることが分かる。

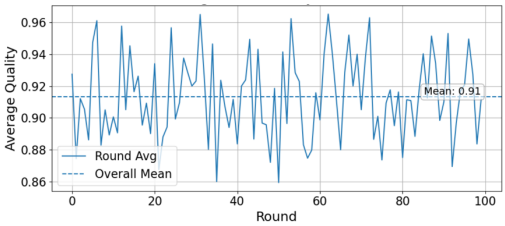

図 3 データ品質値の推移 (レビュー制度あり)

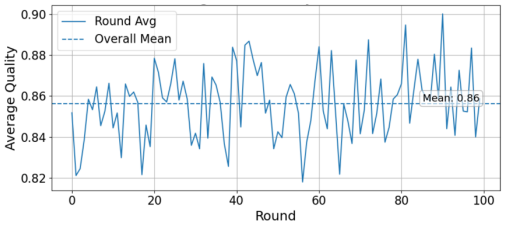

図 4 データ品質値の推移(レビュー制度なし)

長期戦略を選択した提供者は、市場収益性が短期戦略者と比較して 2 倍以上高く、取引量も多い傾向を示した。これは、長期戦略者が高品質かつ高価格のデータ提供を模索し、一度獲得した高いレビュー値を維持・向上させながら最適解を漸進的に学習することで長期的報酬を最大化したためと考えられる。購入者は品質と価格に加え、レビュー情報も考慮して戦略選択を行うため、高評価を獲得した提供者に需要が集中する好循環を形成している。シミュレーション中盤以降、長期・短期戦略ともに収益や取引量が比較的一定の振幅を保ちつつ安定していた。これは、両戦略が環境に適応した学習均衡に達したことを示唆している。

5. 結論

本研究では、データ取引市場におけるレビュー制度の有効性の検証をシミュレーション実験によって行った。実験の結果、レビュー制度を導入した市場では、購入者が品質とレビュー値を併せて評価するため、低品質データは選択されにくい。一方、レビュー制度なしの市場は、品質と価格のみで評価されるため品質のばらつきが拡大し、低品質データの流通が増加する傾向がある。これにより、レビュー制度は提供者の品質改善動機を高めるとともに、購入者の意思決定バイアスを是正し、市場全体に流通するデータ品質を向上させる可能性が高いことが分かった。

謝辞

本研究の一部は JST さきがけ JPMJPR2369 の成果です。

参考文献

[1] Magdalena Balazinska, Bill Howe, and Dan Suciu, “Data markets in the cloud”, VLDB endowment, pp.1482-1485, 2011.
[2] J. Doyne Farmer, Duncan Foley, “The Economy Needs Agent-Based Modelling,” Nature, pp.685-686, 2009.
[3] Chrysanthos Dellarocas, “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms,” Management Science, pp.1407-1424, 2003.
[4] Christopher J. C. H. Watkins, Peter Dayan, “Q-Learning,” Machine Learning, 8, pp 279-292, 1992.
[5] Gerald Tesauro, Jeffrey O. Kephart, “Pricing in Agent Economies Using Multi-Agent Q-Learning”, Autonomous Agents and Multi-Agent Systems, Vol.5, pp 289-304, 2002.
[6] Anton Kolonin, Ben Goertzel, Cassio Pennachin, Deborah Duong, Matt Ikle, Nejc Znidar, Marco Argentieri, “A Reputation System for Multi-Agent Marketplaces,” arXiv:1905.08036, 2019.