1B2-2

U-Martにおける Q学習エージェントの設計と評価

A design and evaluation of Q-learning agents on U-Mart

松本光弘[*1] 福井健一[*2] 森山甲一[*2] 栗原聡[*2] 沼尾正行[*2]
Mitsuhiro Matsumoto Ken-ichi Fukui Koichi Moriyama Satoshi Kurihara Masayuki Numao

[*1]大阪大学大学院 情報科学研究科 情報数理学専攻
Department of Information and Physical Sciences, Graduate School of Informartion Science and Technology, Osaka University

[*2]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

In late years an artificial market attracts attention because it can explain complex behavior of market economy. U-Mart is an open-type test bed of such an artificial market. In this study, we design a Q-learning agent that autonomously makes a trading decision to always make a profit on a transaction. We construct three types of agent and examine which is the best on U-mart.

1. 序章

近年，従来の経済理論では解析できなかったさまざまな現象を解明する可能性があるとして，人工市場が注目されている．人工市場はエージェント同士がそれぞれ取引を行うことで現実的な市場の構築を目指している．その仮想的市場で株価の高騰や急落といった経済的な現象を観測し，その原因を解明しようと試みている．

その人工市場を研究するプロジェクトの一つとして U-Mart プロジェクト [1] がある．U-Mart とは，エージェントや時には人間がエージェントの代わりに株価指数の先物を取引することで市場を構築している仮想先物市場シミュレータである．U-Mart プロジェクトは社会科学に必要な共通テストベッドとして多くの研究者が参加し活動している．

本研究では，強化学習の手法の一つであるQ学習を用いて，先物価格の予測を行うエージェントを設計，評価した．

2. 仮想先物市場 U-Mart

2.1 先物市場

先物市場とは，先物取引が行われる市場のことである [2][3]．先物取引とは，ある商品について将来のある時点で，予め取り決めた価格（先物価格）で取引することを約束する契約のことである．例えば，「金 1kg について，一年後に 130 万円で買うことを約束する.」といった取引が先物取引である．もし，金 1kg が一年後に 150 万円になっている場合でも，130 万円で買うことが出来るため，20 万円儲けることが出来る．

2.2 U-Mart

U-Mart はエージェントにより構成された，コンピュータ上の人工的な市場である．U-Mart で取り扱う商品は株価指数 J30 だけである．J30 は日本株 30 の愛称であり，毎日新聞社が計算・公表していた株価指数である．U-Mart では，J30 を現在の価格（現物価格）とし，それぞれのエージェントがその先物を取引することで，市場が成立する．各エージェントは今期の先物価格が決定される前にそれを予想し，その予想にもとづいて先物価格決定後に自分の資産合計を最大にする行動を決定する．

連絡先: 松本光弘, 大阪大学産業科学研究所, 沼尾研究室,
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1,
Tel:06-6879-8426 Fax:06-6879-8428
E-mail:m_mit@ai.sanken.osaka-u.ac.jp

3. エージェントの売買戦略

エージェントが利益を得るためには，先物価格が将来どのように推移していくのか予測しなくてはならない．U-Mart では，それぞれのエージェントが売買注文を出すことで先物価格が決定する．そのため，時刻 t に各エージェントが売買注文を出すことで，時刻 $t+1$ で先物価格が決定し，その価格より高い買い注文か安い売り注文を出したエージェントがその価格で取引する．利益を出すためには，買った価格より高く売るか，売った価格より安く買い戻さなければならない．したがって，時刻 t に買い（売り）注文を出し，時刻 $t+1$ に買い（売り）注文が成立した場合，時刻 $t+2$ で価格が上（下）がっていれば，利益を得ることが出来る．つまり，時刻 $t+2$ の先物価格が時刻 $t+1$ の先物価格より高くなるのか，安くなるのかを予測できればエージェントは利益を得ることが出来る．本研究では，その予測を強化学習手法を用いて行う．

3.1 強化学習

強化学習の手法の一つにQ学習がある．Q学習はエージェントが試行錯誤的に行動を行いながら，Q値を逐次更新することでQ値を学習し，最適行動価値関数に近づけようとする手法である [4]．

エージェントがある状態 s で，行動 a を取り，状態 s' に遷移したとき，報酬 r が得られたと仮定する．このとき，Q値 $Q(s,a)$ の更新式は以下のとおりとなる．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r_{t+1} + \gamma \max Q(s',a) - Q(s,a)] \quad (1)$$

α は学習率と呼ばれ，一回の学習で修正する度合いを表す．γ は割引率と呼ばれ，将来の報酬が現在においてどれだけの価値があるかを決定する．

本研究では，直前の価格変化（**状態**）を観測し，Q値から株価の昇降を予測し注文を出す（**行動**）．その予測が正しければ報酬を与え，間違っていれば罰を与えることで最適な行動を取るようにQ値を更新していく．

3.2 エージェントの学習戦略

Q学習では状態空間の取り方が重要である．本研究では，大きく分けて 3 つの状態空間のとり方を比較検討する．1 つ目は

先物価格の昇降を状態に取る方法であり，2 つ目は現物価格を状態に取る方法であり，3 つ目は先物価格と現物価格の差を状態に取る方法である．また，価格の昇降を純粋に予測することを目的とするため，得られる利益に関係なく予測が当たった場合は報酬 1 を与え，逆に予測が外れた場合は罰として報酬 -1 を与える．なお、エージェントの行動は買うか売るかの 2 値である．

3.2.1 先物価格の昇降を状態に取る戦略（戦略 1-1,1-2）

この戦略は先物価格が現在（時刻 t）に至るまで，どのように価格変化してきたのかを状態に取る戦略である．例えば，先物価格が時刻 $t-1$ で上昇し，時刻 t で下降していれば，上昇→下降を状態にとり，時刻 $t+2$ の先物価格を予測する．本研究では時刻 $t+2$ の先物価格を予測するために，2 つの予測方法を提案した．それを以下に示す．

時刻 $t+1$ の先物価格を予測し，その予測を元に時刻 $t+2$ の先物価格を予測する（戦略 1-1）

まず時刻 $t+1$ の先物価格を予測する．次に時刻 $t+1$ の先物価格を状態の一部に取り，その状態から上と同様に，時刻 $t+2$ の先物価格を予測する．具体例を図 1 に示す．

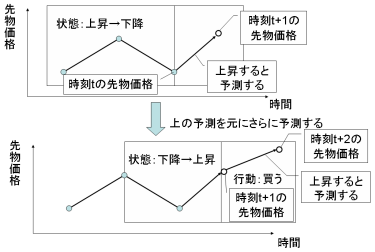

図 1: 戦略 1-1 の予測方法

図 1 では，上昇→下降という状態から時刻 $t+1$ の先物価格が上昇すると予測し，その予測から次状態では下降→上昇となるため，時刻 $t+2$ の先物価格は上昇すると予測し，時刻 t で買い注文を出している．

時刻 $t+1$ の先物価格の変化を考慮せず，時刻 $t+2$ の先物価格を予測する（戦略 1-2）

時刻 $t+1$ の先物価格を予測せず，時刻 $t+2$ の先物価格だけを予測する．具体例を図 2 に示す．

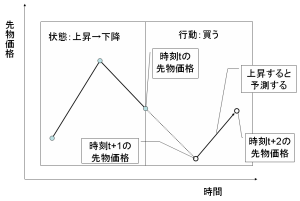

図 2: 戦略 1-2 の予測方法

図 2 では，上昇→下降という状態から時刻 $t+2$ の先物価格が上昇すると予測し，時刻 t で買い注文を出している．

3.2.2 現物価格の昇降を状態に取る戦略（戦略 2）

この戦略は現物価格が現在（時刻 t）に至るまで，どのように価格変化してきたのかを状態に取る戦略である．時刻 $t+1$ の現物価格を予測し，その予測から以下の売買注文決定ルールを用いて売買注文を決定する．

- 時刻 $t+1$ の現物価格が上昇すると予測し，かつ時刻 t の先物価格が現物価格より安ければ買い注文を出す．
- 時刻 $t+1$ の現物価格が下降すると予測し，かつ時刻 t の先物価格が現物価格より高ければ売り注文を出す．
- 上記以外の場合は取引を行わない．

具体例を図 3 に示す．

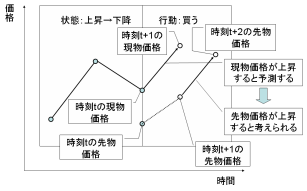

図 3: 戦略 2 の予測方法

図 3 では，現物価格が上昇→下降という状態であるため，時刻 $t+1$ の現物価格は上昇すると予測している．売買注文決定ルールより，買い注文を出している．

3.2.3 先物価格と現物価格の差を状態に取る戦略（戦略 3-1, 3-2）

この戦略は時刻 t での先物価格と現物価格の差を状態に取る戦略である．例えば，先物価格 − 現物価格 =5 円であるならば，5 円を状態とし時刻 $t+2$ の先物価格を予測する．本研究では時刻 $t+2$ の先物価格を予測するために，2 つの予測方法を提案した．それを以下に示す．

時刻 $t+1$ の先物価格の変化を考慮せず，時刻 $t+2$ の先物価格を予測する（戦略 3-1）

先物価格と現物価格の差から，時刻 $t+1$ の先物価格を予測せずに，時刻 $t+2$ の先物価格だけを予測する．具体例を図 4 に示す．

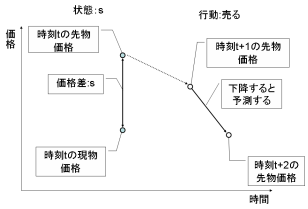

図 4: 戦略 3-1 の予測方法

図 4 では，時刻 t での先物価格と現物価格の価格差が s であることから，時刻 $t+2$ の先物価格は下降すると予測し，売り注文を出している．

時刻 $t+1$ の先物価格と現物価格の大小関係を予測し，売買注文を決定する (戦略 3-2）

先物価格と現物価格の差から，時刻 $t+1$ の先物価格と現物価格の大小関係を予測する。先物価格が高いと予測すれば売り注文を出し，先物価格が安いと予測すれば買い注文を出すという売買注文決定ルールにしたがって，注文を決定する．具体例を図 5 に示す．

図 5 では，時刻 t での先物価格と現物価格の価格差が s であることから，時刻 $t+1$ で先物価格が現物価格より高くなる

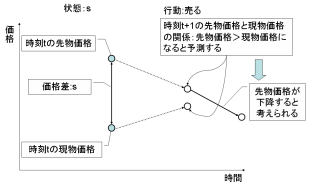

図 5: 戦略 3-2 の予測方法

と予測し，時刻 $t+2$ での先物価格は下降すると考えられ，売り注文を出している．

4. 実験

4.1 実験環境

1. 取引に参加するエージェントは比較検討するQ学習エージェント 1 体と U-Mart に組み込まれているサンプルエージェント 19 体の計 20 体のエージェントである．なお，サンプルエージェントは学習を行わない戦略が固定されたエージェントである．
2. エージェント 1 体の初期資金は 10 億円である．
3. 1 日の取引回数を 8 回とし，30 日 (取引回数 240 回) を 1 セットとして取引を行う．

比較検討するQ学習エージェントは実験を行う前に，100 セットの取引により学習する．エージェントの有能性を評価するために，図 6 のような 3 つの現物価格のパターン（上昇系列，下降系列，振動系列）を用いて実験を行う．どの現物価格のパターンにおいても、利益を得ることができるエージェントを優秀なエージェントとする．

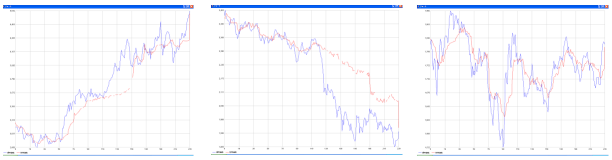

図 6: 左から上昇系列、下降系列、振動系列

4.2 実験結果

4.2.1 比較検討するエージェントの利益

それぞれの現物価格の系列に対して，対象とするエージェントの利益をそれぞれ表 1 に示す．

表 1: エージェントの各利益（× 百万円）

	上昇系列	下降系列	振動系列
戦略 1-1	−268	−50	−34
戦略 1-2	192	−3	−172
戦略 2	472	87	−30
戦略 3-1	264	12	23
戦略 3-2	553	93	35

戦略 1-1 のエージェントは全ての系列で損失を出している．戦略 1-2，2 ではひとつの系列で損失を出している．戦略 3-1，3-2 では全ての系列で利益を得ることが出来た．これより，比較検討するエージェントの中では，戦略 3-1，3-2 のエージェントが最も有能なエージェントである．

4.3 実験結果の考察

各エージェントの利益が表 1 になる原因を考察するため，各エージェントの報酬和（すなわち予測が的中した回数 − 予測が外れた回数）を表 2 に示す．

表 2: 比較検討するエージェントの報酬和

	上昇系列	下降系列	振動系列
戦略 1-1	39	43	54
戦略 1-2	8	−26	−23
戦略 2	−16	−29	−41
戦略 3-1	27	−52	52
戦略 3-2	80	59	53

表 2 より，戦略 1-1 の予測は的中している．戦略 1-1 では時刻 $t+1$ の先物価格を予測するため，時刻 $t+1$ の先物価格を予測することは出来たことになる．そこで，戦略 1-1 の時刻 $t+2$ の先物価格の報酬和（すなわち時刻 $t+2$ の先物価格の予測が的中した回数 − 予測が外れた回数）を表 3 に示す．

表 3: 戦略 1-1 のエージェントの時刻 t+2 の報酬和

	上昇系列	下降系列	振動系列
時刻 $t+2$ の報酬和	−12	10	−13

表 3 より，時刻 $t+2$ の先物価格の予測が外れているため，利益を得ることが出来ていない．

戦略 1-2 も同様に表 2 から，時刻 $t+2$ の予測が出来ていないため，安定して利益を得ることは出来ていない．

戦略 2 では，表 2 より，全ての系列で時刻 $t+1$ の現物価格を予測出来ていない．しかし，上昇系列と下降系列で利益を得ることが出来ている．これは，予測に関係なく，戦略 2 の売買注文決定ルールにより，上昇系列と下降系列で，うまく売買出来たからである．しかし，現物価格の昇降が激しい振動系列では利益を上げることが出来ていない．

戦略 3-1 では，表 1, 2 より，下降系列で予測を大きく外しているが，下降系列でも利益を得ていることが分かる．下降系列での利益の推移を図 7 に示す．

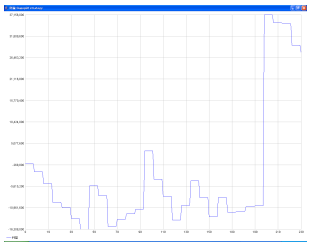

図 7: 下降系列での戦略 3-1 のエージェントの利益の推移

図 7 より，予測を外したときの損失が少なく，予測が的中したときの利益が多いことが分かる．

戦略 3-2 では，表 2 より時刻 $t+1$ の先物価格と現物価格の大小関係を予測出来ていることが分かる．そこで，時刻 $t+2$ の先物価格の報酬和（すなわち時刻 $t+2$ の先物価格の予測が的中した回数 − 予測が外れた回数）を示す．

表 4: 戦略 3-2 のエージェントの時刻 t+2 の報酬和

	上昇系列	下降系列	振動系列
時刻 $t+2$ の報酬和	67	40	58

時刻 $t+2$ の報酬和が戦略 3-1 の予測の報酬和より大きいため，戦略 3-1 のエージェントより時刻 $t+2$ の先物価格を予測できている．また，利益に関しても全ての系列で戦略 3-1 よりも上回っているため，比較検討するエージェントの中で戦略 3-2 のエージェントが最も優秀である．

4.4 他の優秀なエージェントとの比較

U-Mart プロジェクトには研究者のモチベーションを高めるために，エージェント同士を対戦させる大会が毎年開かれている．国内大会の U-Mart と国際大会の UMIE がある．その UMIE で昨年優勝したエージェント 3 体 (Osako_pivotStrategy(OpS)，Nakamura_spreadStrategy(NsS)，Trend_swiftStrategy(TsS) と一昨年の UMIE で優勝したエージェント 2 体 (OPUFuzzyStrategyB(OFSB)，TriDiceP(TDP)) を市場に参加させて，自作エージェントを評価した．参加させた自作エージェントは戦略 3-2 を用いたエージェントである．さらに，19 のサンプルエージェントも取引に参加させ，計 25 体のエージェントで取引を行った．戦略 3-2 のエージェントは，この実験のために改めて学習させず，前の実験の学習結果を用いて実験を行った．取引の結果を表 5 に示す．

表 5: 各エージェントの利益（×百万円）

	上昇系列	下降系列	振動系列
戦略 3-2	395	61	5
OpS	13	−67	−34
NsS	576	46	18
TsS	406	96	65
OFSB	−672	551	−37
TDP	2197	−273	−3

表 5 より，最も安定して利益を出したエージェントは戦略 3-2 と NsS と TsS であることが分かる．戦略 3-2 のエージェントは，他のエージェントが加わった市場においても，安定して利益を出すことが出来た．

5. まとめ

本論文では U-Mart 上で安定して利益を得られるQ学習エージェントを設計，評価するために，状態空間や予測する対象の違うエージェントを比較検討した．検討した中では，先物価格と現物価格の差を状態空間に取り時刻 t+1 の先物価格と現物価格の大小関係を予測するエージェントが最も優れていることが判明した．

今後は，U-Mart2006[1] が今年開かれるため，この大会に出場する予定である．

参考文献

[1] U-Mart project http://www.u-mart.org/html/

[2] 商品先物取引の基礎 http://www.toushin.com/managed/sakimono/futures.htm

[3] 商品先物取引入門 http://www.fuji-ft.co.jp/story/index.html

[4] Richard S.Sutton and Andrew G. Barto (三上貞芳, 皆川雅章 訳), ”強化学習”, 森北出版, 2000