7B-8

モバイルエージェントにおけるクラスタリングを用いた階層的強化学習

下山 佐助† 末田 直道‡

大分大学大学院工学研究科知能情報システム工学専攻† 大分大学工学部‡

1. はじめに

我々の研究室ではスマーティブ(2005年までは「アクティブコンテンツ」の名称を使用)に関する研究をしている。スマーティブとは、受動的に起動される動作を仕込んだだけの、いわゆるカプセル化コンテンツとは異なり、個々のコンテンツが自ら周囲の状況の変化を検出して、その変化に適応するように判断し、必要に応じて交渉しあって相互に協調的に振舞う、能動的/自律的なコンテンツである[1]。

現在はスマーティブのアプリケーションとして、インターネット上を巡回し自律的にデジタルコンテンツ販売を行うエージェントの開発をしている。エージェントに自律的な行動を行わせる場合、何らかの方法で行動を制御する必要があるが、製作者が予め行動を定義すると、エージェントが環境の変化に追従できない可能性がある。そこで、自律学習の一手法である強化学習をエージェントに対して適用することで、ユーザーに応じた適切なコンテンツの配布を学習させ、柔軟なコンテンツ流通を実現する。

強化学習ではある状態においてどの行動を選択するべきかを学習する。今回の問題では、エージェントが持つコンテンツを状態、コンテンツの配布先となるユーザー（以下顧客）を行動と定義している。ここで問題となるのが行動空間の大きさである。エージェントの行動空間は離散の値で表されるが、膨大な離散の行動空間において学習を行う手法はあまり研究されていない。そこで本論文では、行動空間をクラスタリングによって分割することで階層的な強化学習を行う手法を提案する。

2. 問題設定

現在はシミュレータ上で実験を行っている。シミュレータは流通エージェントとコンテンツエージェント、そして顧客によって構成される。シミュレータの実行プロセスは以下のようになる。最初に流通エージェントに対してコンテンツの供給が行われる。流通エージェントは自分が持つコンテンツをどの顧客の下に配布すればよいかを判断し、コンテンツエージェントを乗せて顧客の下へと移動する。それぞれのコンテンツエージェントは、移動先の顧客に対して自分が保持するコンテンツの販売交渉を行う。顧客はそれぞれの嗜好に従ってコンテンツを購入するか否かを決定する。最終的には流通エージェントとコンテンツエージェントの両方が学習を行うシステムを作成する予定だが、今回は流通エージェントの学習のみに焦点を絞った研究を行っている。

Hierarchical Reinforcement Learning with Clustering in a Mobile Agent
Sasuke SHIMOYAMA†, and Naomichi SUEDA‡
†Graduate School of engineering, Oita University
‡Department of Computer Science and Intelligent Systems, Oita University

顧客にはプログラムにより自動生成したデータを使用している。顧客は職業、年齢などのパラメータを持っており、その値によってによってコンテンツに対する嗜好が変わり、購入確率が変動する。

3. 利用技術

① Q-Learning

Q-Learning[2]とは強化学習の代表的手法であり、報酬に至るエピソードのステップごとに、以下の式を用いて、時刻 t における状態 s_t での行動 a_t の行動価値 $Q(s_t, a_t)$ を更新する。

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

ここで、$\gamma(0 \le \gamma \le 1)$ は割引率、$\alpha(0 \le \alpha \le 1)$ は学習率である。この手法は、環境がマルコフ性を満たすときに状態行動の価値が最適解へと収束することを保証している。

② k-means

k-meansはクラスタリングの代表的な手法である。k-means法はクラスタの代表点とし、代表点に対して最も近い位置にある個体を割り当てる。k-meansのアルゴリズムは以下のようになる。

1：k個のクラスタの代表点を無作為に選ぶ。
2：それぞれの個体を最も近い代表点に割り当てる。
3：クラスタごとに代表点を計算し直す。全てのクラスタ代表点が変化しなければ終了。そうでなければ2に戻る。

4. 提案手法

① クラスタリング

顧客は名義属性(職業、商品の嗜好)、数値属性(年齢、年収)などを持っている。そこで、同じパラメータを持つ顧客が似たような嗜好を持つと仮定し、同一のクラスタに割り当てる。この時、数値属性をそのまま使用すると、クラスタリングの際に無駄が生じる。そこで、離散化により連続値を離散値に変換する。なお、離散化には様々な手法があるが、今回は一定のサイズで分割を行う単純な手法を用いている。こうして得られた行動階層に対して、更に行動の類似度によるクラスタリングを行う。行動 a_i, a_j の類似度は以下の式によって定義される。

$$\sigma(a_i, a_j) = \sum_{s \in S''} | Q(s, a_i) - Q(s, a_j) |$$

S'' は今までに経験した全ての状態を表す。類似度の逆数を行動間の距離とし、k-means によってクラスタリングを行う。その結果、三階層の行動空間が生成される。行動空間は上から上位グループ、下位グループ、顧客となる。流通エージェントは行動選択の際に、まず上位グループを選択し、次にその上位グループに属する下位グループを選択する。最後に下位グループに属する顧客を

選択し、その顧客の下へと移動し交渉を行う。なお行動選択の方策にはε-greedyを使用している。

② 学習

上位グループuの行動価値Q_uは以下の式によって表される。

$$Q_u(s_t, a_t) = \sum_{l \in L^u} Q_l(s_t, a_t) / numL^u$$

L^uは全ての上位グループuに属する全ての下位グループである。下位グループでは選択された行動に対しては通常のQ-Learningによる学習を行い、同一の上位グループに属する他の行動に対しても以下の式を用いて学習を行う。

$$Q_l(s_t, a_t) \leftarrow Q_l(s_t, a_t) + \alpha\zeta[r_{t+1} + \gamma \max_a Q_l(s_{t+1}, a) - Q(s_t, a_t)]$$

$\zeta(0 \le \zeta \le 1)$は影響度パラメータであり、この値が大きいほど、選択された行動が他の行動に対して与える影響が大きくなる

5. 実験

提案手法の効果を確認するために実験を行った。実験では行動層が一階層の場合(通常のQ-Learningによる学習),二階層の場合(パラメータによるクラスタリング,FirthestFirst[3],Cobweb[4],k-means)、そして三階層の場合(パラメータによるクラスタリングとk-meansの併用)により性能にどのような差が出るかを調べた。なおクラスタ数は遺伝的アルゴリズムを使用して決定した。

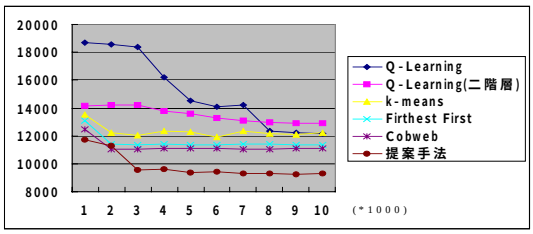

図２：実験結果 1(顧客数 5000,試行回数 10000)

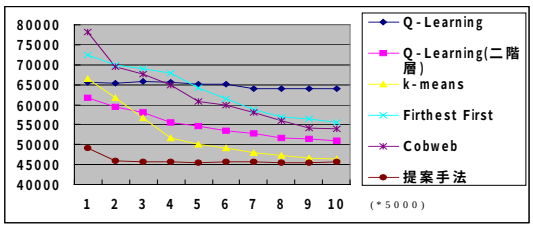

図３：実験結果 2(顧客数 10000,試行回数 50000)

グラフの縦軸は、全試行回数/10回ごとに商品を売り終わるまでにかかった試行回数の合計を表しており、グラフの横軸が試行回数を10分割したものである。グラフが下側にあるほど、商品の販売終了にかかった時間が短く、学習の性能が良いということになる。

結果を見ると実験結果1,実験結果2ともに提案手法が最も収束が早く、また学習性能が良いことがわかる。

また、顧客パラメータのみによるクラスタリングを行った場合も一応学習が可能であるが、提案手法に比べると多少性能が劣っている。クラスタリングを行わなかった場合、実験結果1ではある程度学習しているが実験結果2ではほとんど学習を行えていない。これは、データ数が増加することにより状態行動空間が拡大し、学習がそれに追随できなかったためであると考えられる。それに対して、行動空間を階層化した場合はどの手法においてもある程度学習ができている。このことから、行動数が多い場合には、単純なクラスタリングを行いて階層化を行うことによるメリットはあると考えられる。Firthest-FirstとCobWebを用いた学習手法はどちらも似たような性能を示しているが、データの数が増えると学習性能が悪くなっている。これはデータが増加したことによりクラスタリングの精度が低下していることが原因であると考えられる。逆にk-means法を使用した場合はデータの数が増えると学習性能が向上しており、データの数がクラスタリングの精度に大きく影響していることがわかる。

6. 考察

今回の結果により、膨大な離散の行動空間においても、クラスタリングを行い行動空間を階層化をすることで学習が可能であることがわかった。また、今回使用した環境は不確実性が大きくQ-Learningによる学習は困難であると思われたが、クラスタリングを行った場合には学習が行えている。つまり、行動空間の階層化によりある程度のロバスト性が得られたのではないかと考えられる。

今回の実験では行動間の類似度を用いたクラスタリングにk-means法を用いたが、距離尺度によるクラスタリングを行うものであれば他の手法も使用できるので、Firthest-First法などを使用することによってどのような結果が得られるかを調べてみたいと考えている。

また、行動の類似度の逆数を距離として使用したが、状態を要素、行動を属性として使用することによりCobwebやEMアルゴリズムなどのクラスタリングアルゴリズムも使用することができるのではないかと考えている。

今後は、今回使用したユーザーモデルが妥当かの検討を行う必要があると考えられる。また設定をより現実世界に近づけていく必要がある。一つのアイディアとしては時間とともにユーザーモデルを変化させていくことが考えられる。これは流行などのユーザー嗜好の変化をシミュレータに取り込むためである。この場合ユーザーモデルが変化することに対して、エージェントが追従できるかが大きな問題となる。

[1]吉岡信和・田原康之・本位田真一:モバイルエージェントによる柔軟なコンテンツ流通を実現するアクティブコンテンツ,情報処理学会論文誌:データベース,Vol44,No.SIG18(TOD20)(2003)

[2]Watkins, C. J. C. H. and Dayan, P.:Technical Note: Q-Learning, Machine Learning 8, pp. 279- 292(1992)

[3]Hochbaum, S.D. and B.D. Shmoys,A Best Possible Heuristic for the K-Center Problem. Mathematics of Operational Research(1985). 10(2):p. 180-184

[4] Douglas H. Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering" Machine Learning 2: 139-172(1987)