An Efficient Knowledge Transfer Solution to a Novel SMDP Formalization of a Broker's Decision Problem

(Extended Abstract)

Rodrigue T. Kuate, Maria Chli, Hai H. Wang
School of Engineering and Applied Sciences, Aston University, Birmingham, United Kingdom
{tallakur, m.chli, h.wang10}@aston.ac.uk

ABSTRACT

Retail and wholesale broker's decision problems have been optimized separately, ignoring the probable existence of a globally optimal trading strategy. To address this, we propose a novel formalization, based on a semi-Markov decision process (SMDP) and solved using hierarchical reinforcement learning (HRL) in multi-agent environments. Furthermore, to mitigate the curse of dimensionality, which arises when applying SMDP and HRL to complex decision problems, we propose an efficient knowledge transfer approach. An analysis of our controlled experiments in two well-established multi-agent simulation environments within the Trading Agent Competition (TAC) community shows that this broker can outperform the top TAC-brokers and is able to reuse the trading knowledge acquired in previously experienced settings.

Categories and Subject Descriptors

I.2.6 [**ARTIFICIAL INTELLIGENCE**]: Learning

Keywords

Broker Agent; SMDP; Knowledge Transfer

1. INTRODUCTION

The Trading Agent Competition (TAC) community offers a number of multi-agent simulation environments to promote the development of autonomous trading agents. Among the TAC environments provided, many support the development of broker agents that make profit by minimizing the procurement cost in the wholesale market and by maximizing market share and retail revenue in the retail market. Many studies separately optimize the wholesale and the retail strategies and consider the global optimization of the broker strategy as intractable [4]. The brokers resulting from these studies work well in individually optimizing each strategy, but they never explore the possibility of using a global strategy to maximize their overall profit.

Against this background, we propose an SMDP formalization of the broker's decision problem which enables the simultaneous optimization of its main goal (to maximize the profit) and sub-goals such as minimizing procurement costs and maximizing retail returns. Furthermore, we put forward a knowledge transfer approach which addresses the curse of dimensionality resulting from the SMDP formalization, by reducing the learning time required for the broker to act appropriately in newly encountered markets.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

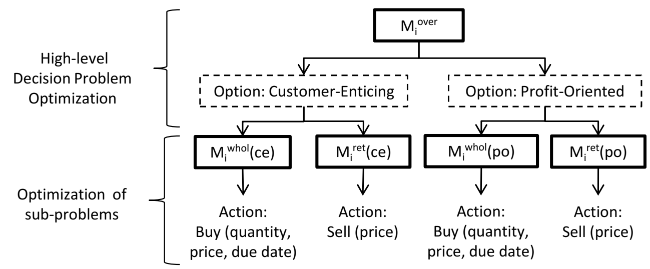

Figure 1: (S)MDP Hierarchy

2. SMDP AND HRL FORMALIZATION

Generally, the architectures of broker agents, which are largely influenced by its activities in the environment, are composed of two key components: one for retail strategy and one for the wholesale strategy [1]. The retail strategy consists of identifying the retail prices that could be accepted by most of the customers and of forecasting the short- and long-term retail demand, whereas the procurement strategy aims to reduce the procurement cost by buying the appropriate products in time, at low prices.

In order to perform well, the broker needs to optimize all its decision making problems both at a global and an individual level. Each decision problem can be modeled as an (S)MDP so that a hierarchy of (S)MDPs can be structured as illustrated in Figure 1. Consider $M_i^j = \langle S_i^j, A_i^j, P_i^j, R_i^j \rangle$, the MDP task for optimizing the decision problem $j \epsilon \mathbb{N}$ of the hierarchy, in simulation environment $i \epsilon \mathbb{N}$. Specifically, the overall SMDP, M_i^{over}, decides the hierarchical, concurrent option to follow: customer-enticing(ce) or profit-oriented(po) option. Based on the selected option, the wholesale MDP, M_i^{whol}, decides the quantity of product to buy, given the projected wholesale price. Concurrently, based on the same option information, the retail MDP M_i^{ret} decides the retail price that can simultaneously increase the profit and the market share. Each of the standard MDPs (M_i^{whol} and M_i^{ret}) can stochastically select many primitive actions before the option of M_i^{over} that is being executed terminates. We applied the any-termination condition for the M_i^{over} multi-options, as it is convenient to implement and preserves the Markov (or semi-Markov) property of the model [3].

The additional transfer framework (a recent survey on transfer learning is provided by [2]) enables the agent to transfer previously acquired knowledge to a new market and subsequently hone its trading skills to the characteristics of that specific market. In this work, we denote the *task domain*, $D_i^j = \langle S_i^j, A_i^j \rangle$ to be the state and action spaces of M_i^j. At the beginning of the learning $T_i^j = \langle P_i^j, R_i^j \rangle$ is defined as the *task objectives* of M_i^j and at the end of the learning it represents the *task skills*. To transfer knowledge, we propose to use an invariant abstract task representation $M_c^j = \langle S_c^j, A_c^j, P_k^j, R_k^j \rangle$ that is common to all tasks M_i^j and is composed of an invariant domain $D_c^j = \langle S_c^j, A_c^j \rangle$ and a portable skills $T_k^j = \langle P_k^j, R_k^j \rangle$. When starting to solve the task M_i^j in a new environment i, a mapping h_i^j is provided by the designer and is used to map the task domain D_i^j to the common domain D_c^j. Mapping h_i^j is defined by the tuple $\langle f_i^j, g_i^j, z_i^j \rangle$ of surjective functions so that:

$f_i^j : S_i^j \rightarrow S_c^j$ maps the state space (or state components) of M_i^j to that (or those) of M_c^j.

$g_i^j : A_i^j \rightarrow A_c^j$ maps the action space of M_i^j to the action space of M_c^j.

$z_i^j : y_i^j \rightarrow y_c^j$ maps the reward intervals defined by y_i^j of M_i^j to common reward intervals defined by y_c^j of M_c^j. Function y_i^j maps each reward r_i^j in the range of the reward function R_i^j to a unique interval $[a, b] = \{x \,\epsilon\, \mathbb{R} | a \leq x \leq b\}$.

Mapping h_i^j makes it possible for the different MDP domains D_i^j to seem the same to the agent. When solving a new MDP task M_i^j, which has a task-specific domain and task-specific objectives, D_i^j is mapped to the invariant domain D_c^j to enable transfer of the portable task skills, while T_i^j is still to be solved. This results in a new reduced task model $M_{i'}^j = \langle S_c^j, A_c^j, P_i^j, R_i^j \rangle$ that has an invariant task domain D_c^j and task-specific objectives T_i^j. Having defined $M_{i'}^j$ for each task, transferred knowledge T_k^j is provided to the agent at the beginning of the training in a new environment i. The aim of the learning is to improve the transferred skills T_k^j to solve $M_{i'}^j$. Let L_j be the learning algorithm used to solve $M_{i'}^j$. $L_j : T_k^j \rightarrow T_i^j$.

In each new environment, T_k^j is initially used and subsequently improved to approximate the skill required for achieving the task objectives T_i^j. We applied n-step TD methods to learn M_c^{over}. To solve M_c^{whol}, we use Monte Carlo (MC) methods, which have been shown to be appropriate for learning this model of the wholesale market MDP, whereas M_c^{ret} is solved using SARSA(λ).

3. EVALUATION

For the purposes of evaluation we use two internationally established multi-agent environments as test beds: the Power Trading Agent Competition (Power TAC) environment, which is an energy market simulation environment, and the TAC Supply Chain Management (SCM) environment - a PC market simulation setting. To evaluate our SMDP approach, we compare the performance of our broker, termed AstonTACPlus with the performance of the top TAC-brokers, AstonTAC and TacTex. Considering the average total cash received by each agent, the more the markets (retail and wholesale) are interdependent, the better is the AstonTACPlus's performance and the worse the performance of AstonTAC and TacTex. In the wholesale market, AstonTACPlus outperforms AstonTAC and TacTex in optimizing the order price and the energy imbalance by having the lowest average order price and lowest energy imbalance, whereas in the retail market, it outperforms AstonTAC and TacTex in optimizing the retail price by having the highest retail average revenue. Our approach performs well irrespective of the level of interdependence between the two markets.

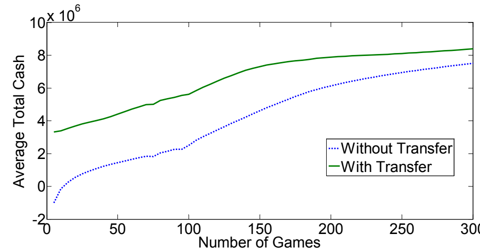

Figure 2: Learning Curves in Power TAC : this figure compares the performance of the trader with and without transfer over 300 games.

Figure 2 shows the average total cash gained by the broker over 300 Power TAC games of 1080 time steps. The curve annotated *with transfer* illustrates the performance in Power TAC of a test broker agent when trained in TAC SCM with 100 games and placed in Power TAC for further training. The curve termed *without transfer* shows the performance of the same test broker when trained in Power TAC without transfer. The performance of the broker with knowledge transfer is better to its performance when no transfer is considered, throughout with the difference in performance being more significant the less games the agent has experienced.

4. CONCLUSION

In this work, we address the broker's decision-making problem by proposing a novel formalization of it as a semi-Markov decision process (SMDP), which enables the broker to simultaneously optimize its retail and wholesale strategies without compromising its global strategy. To reduce the training time that is needed to learn and solve the SMDP, we also propose an efficient agent-centric knowledge transfer approach, which enables knowledge transfer between MDP tasks with different state and action spaces, as well as different reward functions and state transition models.

REFERENCES

[1] M. He, A. Rogers, X. Luo, and N. R. Jennings. Designing a successful trading agent for supply chain management. In *AAMAS*, pages 1159–1166. ACM, 2006.

[2] A. Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012.

[3] K. Rohanimanesh and S. Mahadevan. Decision-theoretic planning with concurrent temporally extended actions. In *UAI*, pages 472–479. Morgan Kaufmann Publishers Inc., 2001.

[4] D. Urieli and P. Stone. Tactex'13: a champion adaptive power trading agent. In *AAMAS*, pages 1447–1448. IFAAMAS, 2014.