Application of reinforcement learning to control a multi-agent system

François Klein, Christine Bourjot, Vincent Chevrier

APPLICATION OF REINFORCEMENT LEARNING TO CONTROL A MULTI-AGENT SYSTEM

François Klein, Christine Bourjot, Vincent Chevrier
LORIA – Nancy University, Campus scientifique BP 239, 54506 Vandoeuvre-lès-Nancy Cedex
Francois.Klein@loria.fr, Christine.Bourjot@loria.fr, Vincent.Chevrier@loria.fr

Keywords: Multi-agent system, control, experimental approach, emergence, global behaviour, reinforcement learning.

Abstract: This study takes place in the context of multi-agent systems (MAS), and especially reactive ones. In such a system, interactions are essential, and trigger a collective behaviour that is not directly linked to the individual ones. Whereas the evolution of the system is unknown if not tried, the regularity of emergent structures in the system is observable and forms a global behaviour. In this paper, we propose to control the global behaviour of a MAS thanks to reinforcement learning tools applied at its global level. We also highlight the choice of the features taken into account to achieve this control, that is the information considered to decide which action to perform.

1 INTRODUCTION

This study takes place in the context of multi-agent systems (MAS) and especially reactive ones (Ferber, 1999). In such a system, the behaviours of the agents at the local level define the system's dynamics. In a reactive MAS these individual behaviours are simple. But the interactions between the agents lead to complex collective behaviours which are not directly linked to the individual ones and which are difficult to predict.

The collective behaviour of the system is observed at the global level. It often involves regular emergent structures or phenomena with a higher time scale than the local evolution of the system. This makes the global behaviour appear stable.

Different global behaviours can be observed for the same MAS depending on perturbations of the system, on changes in initial conditions or changes in parameter values. Indeed, the behaviour of a MAS is highly sensitive to the current state because of the multiple interactions between the agents. It is often strongly non-linear. It has been proved (Wegner, 1997) that such a system cannot be analytically modelled in a general case and that the only way to know precisely its evolution is to experimentally try and run it.

The goal of this study is to control the MAS, that is to make it present one desired global behaviour called the target. To do this, the controller can act on some degrees of freedom, e.g. undefined parameter values, which affect the local level of the MAS.

As it is not possible to determine the effect of local changes on the global behaviour, determining good actions requires an experimental approach (Edmonds, 2004, Edmonds & Bryson, 2004, DeWolf & Holvoet, 2005).

In this paper, we propose to use the global regularities of the MAS to predict its global behaviour, and to control it thanks to reinforcement learning (RL) tools applied at its global level. We highlight the choice of the features considered to achieve this control, that is the information used to decide which action to perform.

In order to assess the proposition, we will compare it to reference approaches on a toy example. We expose the different approaches in the next section. The toy example modelling pedestrians is presented in section 3. An experimental evaluation of the approaches is done in section 4 and their performences are compared.

2 MAS CONTROL SOLUTIONS

In order to make the MAS present the target behaviour, the degrees of freedom can be set only at the beginning of the simulation or they can be dynamically changed during the simulation. In the first case, a classical solution is parameter setting

(Sauter & al., 2001, Sierra & al., 2002). The principle of parameter setting is to explore the parameters space in order to find optimal parameter values, typically thanks to genetic algorithms (Calvez & al., 2005). Once these values are found the controller does not act anymore on the MAS. The main limitation of this solution is its static nature. Specifically, if the MAS undergoes perturbations, the solution is no longer optimised.

A dynamical approach consists in tweaking the degrees of freedom depending on a current state of the system. It can be found in the classical use of decentralised Markov decision processes (DEC-MDP) where each agent is modelled by a MDP (Bernstein & al., 2002). The state of the MAS is the combination of the states of the agents. RL tools can be used to control the system (Sutton & Barto, 1998). But the determination of an optimal policy is a complex problem (NEXP-complete) when the number of agents increases (Goldman & Zilberstein, 2004), usually studied with only a few agents, and cannot be technically applied to large MAS. Another issue is the sensibility of the optimality of the policy to perturbations (Kretchmar & al., 2001). Finally, a local modelling implies that the possible actions are only local. In a DEC-POMDP, we assume that any agent can take a personal decision at any time, and one cannot command a shared resource like the size of the environment.

We propose to apply RL tools at the global level of the MAS (Klein & al., 2008). The system is modelled as a simple MDP with global states and actions. Good features must be chosen in order to differenciate the states. A state of the MDP gathers many local situations of the MAS. For two given states S and S' and an action *a*, we can consider as a transition function the proportion T(S, a, S') of situations in S that stabilise in S' when *a* is performed. The states of the MDP correspond to the regularities of the MAS. We can use for instance a description of the emergent phenomena. Since the exact evolution of the system is unpredictable, the past situations are insignificant, and the Markov property is a good approximation.

Finally, we want to compare our dynamical solution to the static parameter setting one, and to another dynamical reference solution. In this reference the actions are decided randomly. It is used to know if the optimisation of RL tools is useful or if the dynamicity is sufficient.

To assess these approaches we compare them on two control problems using the MAS described in the next section.

3 APPLICATION EXAMPLE

Our application example is a MAS which roughly models pedestrians walking in a circular corridor. Realism of the model is not prime concern here, since we wish to illustrate how to apply the proposition. Like in a flocking system (Reynolds, 1987), agents are leaded by a sum of forces that come from their own goals and from the repulsion with other agents and with the walls (figure 1).

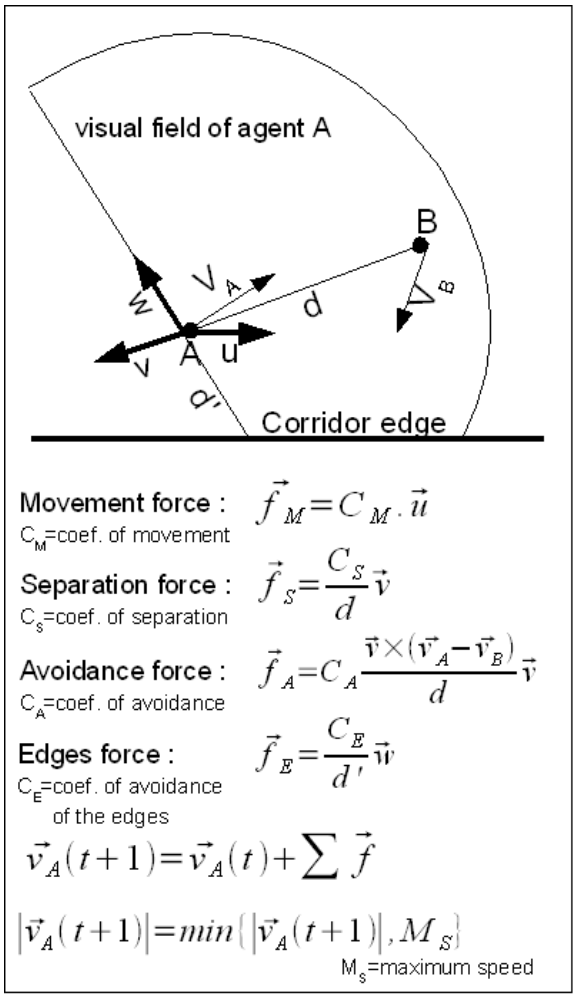

Figure 1: Dynamics of the system. The computation of the new speed of an agent A depends on the sum of forces. 5 parameters define the behaviour of an agent : Cm, Cs, Ca, Ce and Ms.

Five parameters define the behaviour of an agent: a maximum speed norm Ms and 4 coefficients corresponding to the 4 forces applied to the agent. We assume that the degrees of freedom are some of these parameters.

3.1 Definition of the Control Problems

When simulated, the system shows up emergent groups of agents: lines of same direction agents and blocks of opposite agents (see figure 2). We assume that we intent to control the number of each kind of groups, and we use lines and blocks to define the global behaviour.

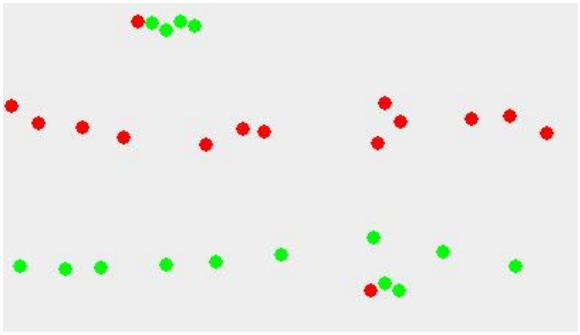

Figure 2: Emergent structures. 2 blocks and 2 lines of agents emerging in the pedestrians system. Red (dark) agents go to the right and green (light-coloured) agents to the left.

We consider two control problems. In the first one, the target is one block and two lines, and in the second problem, no block and two lines.

An action is the decision of the value of each controllable parameter. These values are applied to sets of agents, depending on the state definition. Each parameter can take 5 values. There are 25 possible actions in the first problem, and 125 in the second one. The control problems are summarised in table 1.

Table 1: Summary of both the studied problems.

Problem	Controlled parameters	Number of actions	Target	
			blocks	lines
1	Cm Cs	25	1	2
2	Cm Cs Ms	125	0	2

3.2 Implementation of the Proposition

We propose three states descriptions denoted D1, D2 and D3, which correspond to three sets of features to take into account when it comes to choose an action to perform.

In D1, each state corresponds to a global behaviour (e.g. 1 line and 2 blocks). The number of considered lines and blocks are respectively limited to 5 and 2, so that there are 18 possible states (from 0 to 5 lines and from 0 to 2 blocks).

The description D2 differentiates the control of the agents belonging to lines and the agents belonging to blocks. Two populations of agents are dynamically identified, so there are only two states. Each agent chooses an action depending only on the population it belongs to.

D3 derives from D1 and D2: we differentiate the populations, but we take the global behaviour into account. In this case, the choice of an action depends on the current global state and the population the agents belongs to. Since a state is the combination the states of the sub-populations, there are 18^2.

For all these implementations, a Sarsa algorithm is used to learn a stochastic policy after 3000 simulations. When we model the MAS with the D3 solution, we compute the near-optimal policy in alternating the optimisation of each population while letting the other population following its last optimised policy (Chades, 2006).

4 ASSESSMENT

In this study, the efficiency of an approach is represented by the rate of convergence toward the target, that is, the number of times the system is able to reach it. The approach is applied on 300 simulations and the potential reaching of the target is stored. An approximation of the rate of convergence toward the target can be computed.

A simulation begins by setting the MAS in a random situation. The controller then decides which action to realize, and the MAS is let run with the specified parameters, until a global behaviour is identified. This step is repeated until a stop criterion: either the target behaviour is reached, or a maximum number of steps k has been performed and we consider that the MAS will never reach the target. We always took k=50.

Table 2. Evaluation of the rates of convergence for different approaches on two problems (in %)

Approach	**First problem**	**Second problem**
Parameter setting	15,3	25,2
Dynamical stochastic reference	68,7	23
D1	94,2	66,6
D2	80	51,6
D3	92,4	59,4

For the parameter setting approach, only one initial action is performed and we just verify if the target is reached after this single step. The best parameter values are found by testing each action 500 times. The action which gives the best rate of convergence corresponds to the values to set.

The table 2 summarises the control performances of the different approaches on the two problems presented in §3.1. Different values for ε were tried in D1 and the best ones were kept.

Above all, we see that our proposition gives better results than the two other control methods in terms of proportion of convergence to the target.

The most surprising result is that the situation description D1 triggers better control performances than D3. Indeed, D3 allows to take more information into account than D1 thus the optimal policy with D1 is acceptable for D3, which should theoretically give the best results. It would be true if the optimal policy was actually reached for each model. But the more complex the model, the more difficult the policy to compute. We notice that for limited resources, especially simulation time, a simplest model can show the best performances.

Finally, if global features are not available for the control, as in D2, the proposed method gives promising results. We can consider to control the system at a mesoscopic level, between the local (classical use of MDP) and global (D1 and D3) ones.

5 CONCLUSIONS

In this paper, we demonstrate that it is possible to control the global behaviour of a multi-agent system by considering its global properties and regularities. We achieved a dynamical control, better than the static solution of parameter setting.

We highlight the importance of choosing good global features of the MAS to take into account in the action decision: a simple but well-chosen feature is easier to implement and gives better results than a more complex model if the computing resources are limited.

In further works, we plan to improve the choice of the features considered, and to find a balance between global but useful features and decentralised ones that triggers a weaker control. For instance, instead of creating a global controller that knows everything, we consider to control the system thanks to some luring agents with limited perception and local actions. Finally, we could optimise the tools used and apply them on other systems.

REFERENCES

Bernstein, D.S. Givan, R., Immerman, N., Zilberstein, S. The complexity of decentralized control of markov decision processes. Mathematics of Operations Research, 2002

Calvez, B., Hutzler, G., 2005. Automatic tuning of agent-based models using genetic algorithms. In Proceedings of the 6th International Workshop MABS'05: 39-50.

Chadès I., 2006. Algorithmes de co-évolution simultanée pour la résolution approchée de PDM Multi-agent. Revue d'intelligence artificielle, vol. 20, pp 345—382.

DeWolf, T., Holvoet, T., 2005. Towards a Methodology for Engineering Self-Organising Emergent Systems. In: Proceedings of SOAS 2005, Glasgow, Scotland.

Edmonds, B., 2004. Using the Experimental Method to Produce Reliable Self-Organised Systems. In Engineering Self Organising Sytems: Methodologies and Applications, Springer, pp 84-99.

Edmonds, B., Bryson, J., 2004. The Insufficiency of Formal Design Methods - the necessity of an experimental approach for the understanding and control of complex MAS. AAMAS 2004: 938-945.

Ferber, J., 1999. Multi-Agent System: An Introduction to Distributed Artificial Intelligence. Harlow: Addison Wesley Longman.

Goldman, C.V., Zilberstein, S., 2004. Decentralized Control of Cooperative Systems: Categorization and Complexity Analysis. Journal of Artificial Intelligence Research, 22:143-174.

Klein, F., Bourjot, C., Chevrier, V., 2008: *Controlling the global behaviour of a reactive MAS: Reinforcement learning tools*. In Proceedings of ESAW, 2008.

Kretchmar, R.M., Young, P., Anderson, C., Hittle, D., Anderson, M., Jilin Tu, Delnero, C., 2001: Robust Reinforcement Learning Control. *American Control Conference.*

Reynolds, C. W., 1987. Flocks, Herds, and Schools: A Distributed Behavioral Model. In Computer Graphics, 21(4) (SIGGRAPH '87 Conference Proceedings) pages 25-34.

Sauter, J.A., Parunak, H.V.D., Brueckner, S., Matthews, R., 2001. Tuning Synthetic Pheromones With Evolutionary Computing. In GECCO 2001, San Fransisco, CA.

Sierra, C., Sabater, J., Agusti, J., Garcia, P., 2002. Evolutionary Computation in MAS Design. In Proceedings ECAI, pp188-192.

Sutton, R., Barto, A., 1998. Reinforcement Learning : an introduction. MIT Press, Cambridge.

Wegner, P., 1997. Why interaction is more powerful than algorithms. In Communications of the ACM, Volume 40, pp. 80 – 91, New York