The exploration-exploitation dilemma for adaptive agents

Lilia Rejeb[1], Zahia Guessoum[1,2] and Rym M'Hallah[3]

[1] CReSTIC, MODECO Team, Rue des Crayères, Reims Cedex2, France
[2] Université de Paris-VI, LIP6, OASIS Team, 4 place Jussieu, 75252 cedex 5, France
[3] Kuwait University, Dep. of Statistics and Operations Research
P.O. Box 5969, Safat 13060

Abstract. Learning agents have to deal with the exploration-exploitation dilemma. The choice between exploration and exploitation is very difficult in dynamic systems; in particular in large scale ones such as economic systems. Recent research shows that there is neither an optimal nor a unique solution for this problem. In this paper, we propose an adaptive approach based on meta-rules to adapt the choice between exploration and exploitation. This new adaptive approach relies on the variations of the performance of the agents. To validate the approach, we apply it to economic systems and compare it to two adaptive methods: one local and one global. These methods which were originally proposed by Wilson are adapted herein to economic systems. Moreover, we compare different exploration strategies and focus on their influence on the performance of the agents.

1 Introduction

The exploration-exploitation dilemma, which is an important problem frequently encountered in reinforcement learning [10], is defined as follows. When an agent is faced with a state of the environment, it has to choose between two options: (i) exploration and (ii) exploitation. The agent can choose to (i) explore its environment and try new actions in search for better ones to be adopted in the future [6], or (ii) exploit already tested actions and adopt them. When opting to explore new actions, the agent is considering its long term performance whereas when opting to exploit tested actions, the agent is guaranteeing its short term performance. Of course, this choice is in many cases motivated by the estimated life span of the agent. The shorter its life span, the more inclined is the agent to improve its short term performances; subsequently, the more inclined is the agent to limit exploration [8].

Formally, the agent has to resolve two subproblems. The first subproblem consists of choosing an exploration method. The exploration can be either directed or undirected. The exploration is directed when the choice is based on the acquired knowledge whereas it is undirected when the choice is random. The second subproblem consists of identifying a method that switches the agent's mode between exploration and exploitation according to the state of the agent and the state of its environment. The two subproblems are important since they influence the learning speed, the performance and the actions of an agent. This influence is more critical when the agent's environment is dynamic -which is the case of economic systems-.

In this paper, we study the aforementioned two subproblems in the context of an economic system characterized by a set of firms in competition in a shared market. We propose an adaptive approach to the exploration-exploitation problem in a dynamic economic context where firms are modeled using the XCS-learning classifier system [3]. We show that firms' performance can be improved if we opt for directed exploration and if we use a meta-rules based approach to choose between exploration and exploitation. To assess our meta-rules based approach, we first adapt Wilson's [10] techniques, which were originally tested in non-dynamic environments, to a dynamic economic environment. We, then compare our meta-rules based approach to the adapted versions of Wilson's [10] techniques.

This paper is organized as follows. Section 2 presents the firm model and an overview of the learning classifier system XCS. Section 3 investigates exploration techniques. Section 4 presents the proposed meta-rules approach and the adaptation of Wilson's techniques to our context. Section 5 presents and analyzes the experimental results. Finally, Section 6 summarizes the contributions of this paper and provides possible future extensions.

2 Adaptive Firms

We study the exploration-exploitation dilemma in the context of a dynamic economic system where a set of firms are in direct interaction in a shared market. We model the firms as adaptive agents with the XCS-learning classifier system for their learning. In Section 2.1, we present the model of a firm while in Section 2.2, we detail the characteristics of the XCS classifier system, explain how agents learn using the XCS classifier system, and specify the reward qualitative function of an agent.

2.1 The firm model

We model the firms using a resource-based approach [4]. Indeed, we regard a firm as a collection of physical and human resources. We stipulate that the survival of a firm depends on the way it allocates its resources. A firm is characterized by

- a set X of resources,
- a set Y of performance indicators: $Y = (Y[1], Y[2])$ where $Y[1]$ is profitability and $Y[2]$ is performance,
- a capital K,
- a budget B (which when allocated updates the status of the firm's resources), and
- a set S of strategies available for the firm. Each strategy defines a method for allocating the budget to the different resources according to the firm's priorities.

A firm's behavior is dynamic over time. At each time period, a firm

- observes its environment and updates its competition model;
- updates its internal parameters (eg., its capital K and budget B);
- opts for a strategy; and
- updates its performance.

A firm chooses the strategy that best suits its current context. The context of a firm is determined by the firm's

- internal parameters (capital, budget, amount of resources and performances), and
- perception of the environment.

The environment of a firm is determined by the other firms present in the market. The environment is strongly competitive and non-stationary. At each time period, firms can either join or leave the market. While new firms offer opportunities to some firms, they threaten others. Firms leaving the market are generally unsuccessful ones. The aggregate of the performances and capital of the firms present in the environment is an adequate quantification of the current state of the environment.

A firm evaluates its current context based on its internal parameters and the dynamic variation of its environment. It then chooses a strategy that best suits this context.

The dynamic nature of the environment makes it difficult for a firm to anticipate all the possible outcomes of its strategy and/or to take into account its prior inadequate strategies. The firm needs to build its knowledge regarding its environment and its decision making process. It has to gradually construct its rule based inference mechanism. The firm has to enrich its decision rules based on its experience, and to evaluate these rules based on their outcome.

2.2 XCS and adaptive firms

We use XCS to model the decision process of adaptive firms. XCS-based firms are obtained by the integration of XCS and the agent representing a firm. XCS is a learning classifier system [9] where knowledge is represented by standard "condition-action" rules called classifiers. Each classifier is characterized by three parameters:

- the prediction p, which corresponds to the average estimated reward of a classifier;
- the prediction error e, which estimates the error in the prediction of the classifier, and
- the fitness F, which evaluates the prediction quality by measuring the average accuracy of the prediction p.

Based on F, XCS foresees the possible feedback of each action a_i[1], where the feedback is the reward the firm or agent gets from its environment when it adopts action a_i. XCS then uses the set of feedbacks to choose the action of its agent.

The condition part of classifier j, cl_j, is a representation of the perception of the environment. It represents the context of the firm.

The set of possible actions is determined by the economist who is aware of the strategies of the firm. An action a_i is characterized by its average prediction

$$PS_i = \frac{\sum F_{cl_j} p_{cl_j}}{\sum F_{cl_j}}, \tag{1}$$

[1] i varies between 1 and k, where k is the number of the possible actions

where F_{cl_j} and p_{cl_j} are respectively the fitness and the prediction of classifier j when undertaking action a_i.

The choice of an action is based either on exploration (random choice) or on exploitation (choice of the action having the largest PS_i). Because of the strong competition characterizing the environment, a firm should be careful when selecting exploration or exploitation. Exploration directs a firm either towards risk taking whereas exploitation directs it towards risk avoiding.

When a firm adopts an action or strategy, it gets a reward. The reinforcement learning component of XCS uses this reward to update the parameters p, e and F of the classifiers. The reward could be immediate. The reward at time t, r_t is modeled according to the performance variation. It is defined as an aggregation of the variations of the performances of the firm:

$$r_t = aggreg\left(\frac{Y_t[1] - Y_{t-1}[1]}{Y_{t-1}[1]}, \frac{Y_t[2] - Y_{t-1}[2]}{Y_{t-1}[2]}\right) \tag{2}$$

where $Y_t[1]$ is the firm's profitability, $Y_t[2]$ its market performance and *aggreg* the average aggregation operator.

The considered version of XCS uses the Q-Learning updating algorithm as a reinforcement learning component to update the classifiers parameters. The Q-Learning allows XCS to learn the qualities associated with the classifiers. At each time period, the quality or prediction of the classifier is updated:

$$Q(s_t, a_t) = Q(s_{t-1}, a_{t-1}) + \beta(r_t + \gamma max_{a \in A}[Q(s_t, a) - Q(s_{t-1}, a_{t-1})]) \tag{3}$$

where s_{t-1}, a_{t-1} correspond to the condition and action at time $t-1$.

This perception, prediction, action cycle undertaken by XCS at each decision period of the firm is detailed in the algorithm provided in Table 1. A more comprehensive description is available in [1].

1. Perceive the context of the firm;
2. define the set [M] of classifiers that match the environment state; if [M] is empty, apply covering;
3. define the system prediction array [PS];
4. choose the action "a "having the best PS value either by exploration (random choice) or by exploitation;
5. undertake action "a "and compute the reward r;
6. generate the action set [A] which is the set of classifiers in [M] having "a "as action;
7. allocate the reward r to the firm;
8. evaluate the classifiers and apply the genetic algorithm if possible.

Table 1. Perception, Prediction, and Action cycle of XCS

In the current version of XCS, the choice between exploration of new strategies and exploitation of strategies having the highest prediction is random. A number is randomly generated from the continuous Uniform (0,1), and compared to the exploration

probability fixed by the designer prior to the beginning of the simulation. In addition, the exploration of the strategies is undirected. In the following, we discuss the impact of the exploration strategy on a firm's performance and survival.

3 Exploration techniques

Exploration techniques are classified as undirected and directed [7]. Undirected techniques are random. Directed exploration techniques seek to improve the knowledge of the environment by adopting more informative actions. They include techniques such as recency-based exploration and frequency-based exploration.To compare the performance of firms under directed techniques to their performance under undirected ones, we need to integrate directed exploration techniques in XCS. In fact, in the absence of a decision methodology, XCS uses undirected exploration.

To apply the *recency-based technique*, we consider, for each action a, the corresponding matching classifiers $cl_j, j = 1, \ldots, n$ and determine the least recent activation-time

$$Rec(s,a) = \max_{j=1,n}(t - ActivationTime(cl_j(s,a))) \tag{4}$$

where $ActivationTime(cl_j(s,a))$ is the last activation date of classifier j, s is the condition cl_j, and t is the current date.
To apply the *frequency-based technique*, we consider, for each action, the corresponding matching classifiers and determine the number of classifiers previously rewarded

$$Freq(s,a) = \min(\sum_{j=1,n} cl_j(s,a)) \tag{5}$$

such that $experience(cl_j(s,a)) > 1$, where $experience$ corresponds to the activation of cl_j in a similar state of the environment.

Wiering [8] states that continuous exploration techniques are useful when the agent does not care about immediate reward and is rather interested in a long term policy [8]. However, when the firm is interested in immediate reward, it has to switch to exploitation and has to gradually increase its rate of switching to exploitation. An exploitation-exploration tradeoff is therefore needed.

4 Exploration-exploitation tradeoff

Finding a balance between exploration and exploitation is not an easy task. Most existing methods deal with small non-complex problems [7]. Methods that are applicable to complex contexts such as a multi-agent context are limited in number [5, 2]. The exploration techniques of [5] are applicable to the context of the bar problem which is a simple case of a multi-agent system. Peres [5] underlined the necessity to link the changing rate of exploration and the changing indicators of performance to the changing prediction, but proposed no method to achieve this link. Carmel [2] integrated an exploration technique in his learning-based model and applied it in the context of game theory with a small number of agents.

Wilson [10] proposed ten techniques that were tested on small simple test problems. Their performance is sensitive to the constant gain factor fixed by the designer. The behavior of these techniques in complex systems remains however an open issue. In the following, we propose to test the behavior of two of these techniques in more complex settings.

4.1 Wilson techniques

Wilson techniques focus on an "on-line "choice between exploration and exploitation in a varying environment. They are based on the rate of variation of the performance (prediction) or the prediction errors. In this section, we adapt two of Wilson's techniques. We choose two adaptive techniques, one global and one local. These two techniques are based on the prediction errors.

Local technique An adaptive local technique senses a condition that is a property of a niche in its interaction with the environment and chooses its action accordingly. It determines the exploration probability p_1 based on $\widehat{E}_i$, the moving or recency-weighted average of the difference between the current error and the error estimate of action a_i:

$$p_1 = min\left(1, f\left(\widehat{E}_i\right) \times Gf\right), \tag{6}$$

where Gf is a given gain factor, and $f\left(\widehat{E}_i\right)$ is the average of the $\widehat{E}_i$ with i varying between 1 and the number of the identified actions.

This technique is included in the XCS step as follows. When all the classifiers matching the current context are identified, the values of $\widehat{E}_i$ are computed for all actions a_i.

Global technique An adaptive global technique tracks the changes in the environment, and adapts its actions depending on these changes. It senses the condition of the whole system-environment interaction, and sets the system's general exploration probability accordingly. This technique estimates $\widehat{E}$ the average prediction error during exploration periods and determines the exploration probability p_1 accordingly. The exploration is continued till the average error does not change or changes very little. Thus, if the average prediction error changes, ***n*** other steps of exploration are executed prior to switching to exploitation. The rate of change $g\left(\widehat{E}\right)$ is the difference between the moving averages of $\widehat{E}$ before and after ***n*** periods of exploration (where ***n*** is usually set to 100). The probability p_1 is determined as:

$$p_1 = min\left(1, g\left(\widehat{E}\right) \times Gf\right), \tag{7}$$

where Gf is a given gain factor.

Both of Wilson's strategies -the local and the global- are sensitive to the gain factor. Their performance depends on the gain factor, on the number of periods, and on the exploration probability threshold fixed by XCS. To avoid the sensitivity of Wilson's techniques, we propose an approach based on meta-rules.

4.2 A meta-rules based approach

We propose to use meta-rules to control the activation of exploration and exploitation. These meta-rules adapt the choice between exploration and exploitation to the evolution of the firm's performance. They account for the new variations of the environment, once the firm has learned. They are simple and make the behavior of the classifier system close to that of the real decision maker. Contrary to the techniques of Wilson, they allow the return of a firm to exploration and do not use the gain factor.

The proposed meta-rules based approach is based on two parameters: ***n*** the number of periods of exploration, and ***m*** the number of periods of exploitation. After ***n*** periods of exploration and ***m*** periods of exploitation, the approach starts applying the meta rules as follows.

- If $Perf[t+n] > Perf[t+n+m]$, the system must continue learning. Subsequently, the number of periods of exploitation ***m*** must be decreased:
  $m = m * (1 - Exploitation\_Rate)$.
- If $Perf[t+n] \leq Perf[t+n+m]$, the system has achieved enough learning. Subsequently, the number of periods of exploitation ***m*** must be increased:
  $m = m * (1 + Exploitation\_Rate)$.

$Exploitation\_Rate$ represents the variation rate of ***m***. The exploration period ***n*** is herein set to a fixed value, but it could be allowed to vary. Once the system has acquired enough learning, the value of ***m*** becomes very large. The value of ***n*** is maintained positive to allow the system to adapt to small changes of the environment.

5 Experimental results

The objective of this experimentation is twofold. First, we investigate the impact of exploration techniques on the performance of a firm. Second, we study the behavior of the meta-rules based approach and compare it to other choice techniques of exploration and exploitation.

The XCS parameters are fixed as follows: the population size N = 6000 to allow the system to represent all the possible classifiers when the generalization is not used, the generalization probability #_probability = 0.5; the learning rate β = 0.001; the crossover rate = 0.8; the mutation rate = 0.02; the minimum error = 0.01; the genetic algorithm frequency $\theta_{gen} = 10$; and the exploration probability = 0.5. Each simulation is replicated 20 times.

5.1 Exploration techniques

The first series of experiments compares exploration techniques. The comparison is based on the results of the simulation of three populations involving 300 firms each. These populations use respectively recency, frequency, and random based exploration techniques. The first two techniques are directed exploration techniques while the last technique is undirected. The three populations use identical initial parameters and the same exploration-exploitation method.

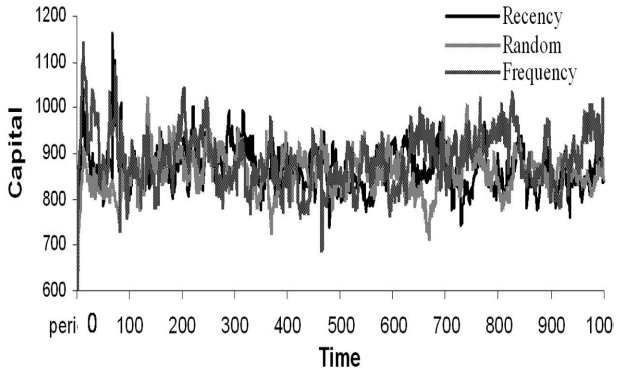

Fig. 1. Directed vs. Undirected exploration

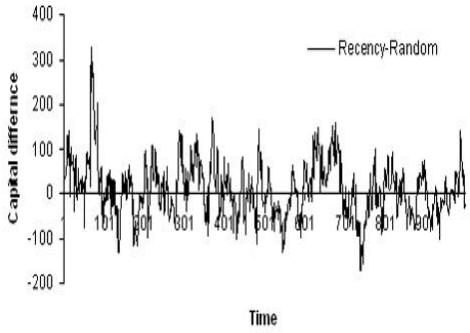

Fig. 2. Random vs. frequency-based exploration techniques

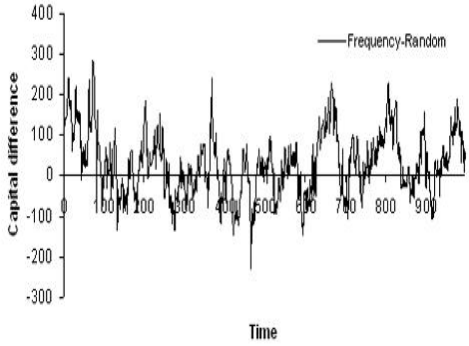

Fig. 3. Random vs. Recency-based exploration techniques

Figure 1 displays the impact of the exploration technique on the average capital of a firm. It shows that directed exploration is interesting at the beginning of the simulation period. It directs the exploration towards the use of new actions; which is not always the case for random exploration. It enriches the classifier population at the beginning better than random exploration, and results in a larger accumulation of the environment's knowledge. However, in the long run, directed exploration becomes equivalent to random-exploration. Figures 2, 3 show the difference between the directed and undirected Exploration techniques. They show that on average, directed exploration does not greatly improve the performance of a firm. The percent improvement is on average 3.4 % and reaches a maximum of 9.1% and a minimum of -7%). Table 2 displays the mean and standard deviation of the capital of firms from different simulation runs. The mean of the two techniques of directed exploration is greater than that of the random exploration but this difference is not statistically significant at the 0.0005 % level.

		Run				
		1	2	3	4	5
Random	Standard deviation	99.78	104.22	111.21	128.57	50.11
	Average	869.63	854.64	880.57	875.57	862.43
Frequency	Standard deviation	58.72	123.48	123.53	113.77	125.35
	Average	882.66	874.73	861.91	870.71	883.02
Recency	Standard deviation	123.87	118.44	131.30	122.32	55.35
	Average	872.38	869.61	880.31	886.91	877.57

Table 2. Summary statistics for the capital of firms

Nevertheless, directed exploration slightly improves the resistance of firms by allowing them to survive longer. The fittest of the firms survive better when they opt for direct exploration techniques, as illustrated by Figure 4. Directed exploration increases the firm's learning speed; making them converge more rapidly to a stable classifiers population. In conclusion, directed-exploration alone is not sufficient to improve the performance of a firm. A balance between exploration and exploitation remains needed.

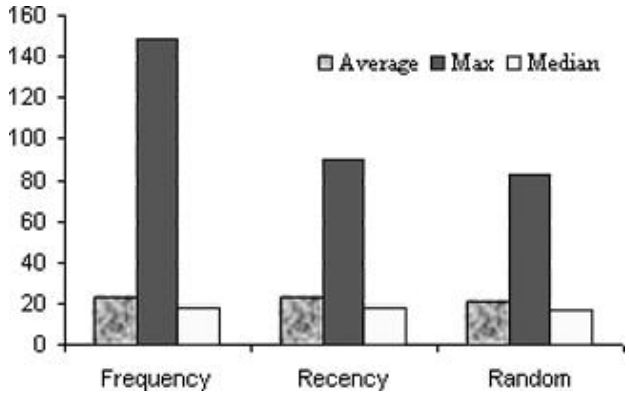

Fig. 4. Comparison of firms' resistance

5.2 Exploration-exploitation techniques

The second series of experiments compares the techniques of choice between exploration and exploitation. First, we compare the proposed meta-rules based approach to a random switch approach. Second, we compare the proposed meta-rules based approach to the adapted Wilson's techniques.

Meta-rules vs. random switch techniques To compare the proposed meta-rules approach to a random switch approach, we run a simulation involving random-XCS firms

and firms using meta-rules. The first population uses a random choice between exploration and exploitation whereas the second uses the meta-rules with an $exploitation\_Rate = 20\ \%$. The periods of exploration and exploitation, ***n*** and ***m*** are set to 10 and 30, respectively. To focus on the exploration-exploitation switch technique, we endow these populations by identical parameters and by the same exploration technique.

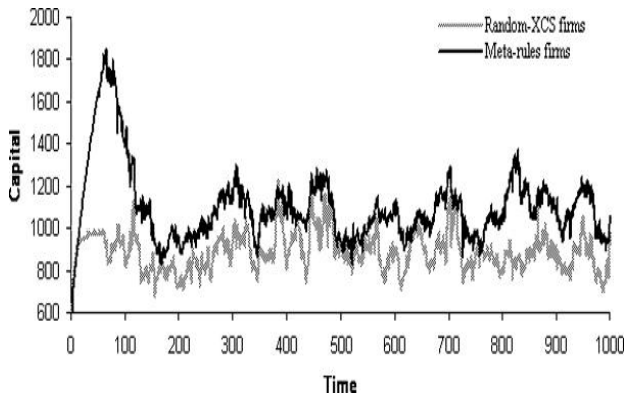

Fig. 5. Comparison of the capital of random XCS firms and meta-rules firms

Figure 5 shows that the use of meta-rules improves the performance of surviving firms. The comparison of the average life span for firms adopting meta-rules (112 periods) to the average life span for random-XCS firms (107 periods) shows that meta-rules improve the resistance of firms. The important degradation of the performance of firms when meta-rules are applied coincides with the beginning of the exploitation period. This degradation shows that firms should have pursued learning and that it was too early for them to consider exploitation.

Despite their positive impact on the performance of firms, the meta-rules are sensitive to the values of ***n*** and ***m***. Large periods of exploration are advantageous at the beginning when the firm has not learned enough. However, large ***n*** values could become hazardous when the firm has acquired enough learning. At the end of the simulation, shorter periods of exploration are preferred.

Meta-rules vs. Wilson's techniques The following results correspond to the simulation of a market including 3 populations: a population of 300 firms adopting the meta-rules based approach, a population of 300 firms using Wilson's adaptive local technique, and a population of 300 firms using Wilson's global adaptive technique. The gain factor for Wilson's techniques is set to 0.5. The three populations share the same parameters and strategies except for the exploration-exploitation switch strategy. For the sake of clarity, we compare Wilson's local strategy to the meta-rules strategy on Figure 6 while we compare Wilson's global strategy to the meta-rules strategy on Figure 7.

These graphics show that the use of meta-rules improves the performance of firms. This improvement is more pronounced with respect to Wilson's global adaptive tech-

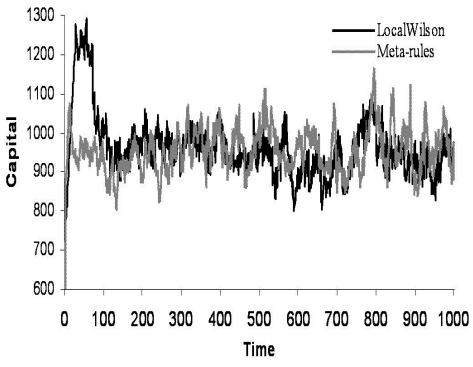

Fig. 6. a. Comparison of the meta-rules based approach to Wilson's local technique

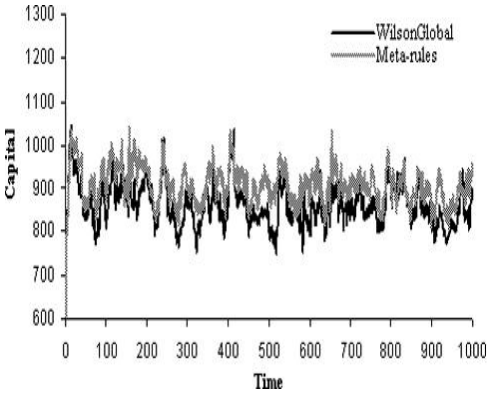

Fig. 7. Comparison of the meta-rules based approach to Wilson's global techniques

nique. In fact, this strategy does not allow firms to return to exploration once they have acquired enough knowledge and the environment has changed. The improvement is less pronounced when comparing the meta rules to Wilson's local technique as the latter reconsiders the choice for each period. The local Wilson strategy is clearly better at the beginning of the simulation period as the meta-rules based approach engages only in exploration for long periods, at the beginning of the simulation. However, in the long term, meta-rules outperform Wilson's local strategy. This better performance may be explained by the fact that basing the decision only on the current information is not wise on the long run. The meta-rules based approach is promising and could be improved by adapting the $exploitation\_Rate$ to the age of the firm.

These conclusions are further confirmed when we compare the standard deviation and mean of the capital of each population displayed in Table 3. A smaller standard deviation of the capitals reflects a more stable behavior of an approach; thus, the meta-rules strategy is more stable than either Wilson's Local or Global strategies. Even though its mean is the largest, Wilson's local strategy is not necessarily the best strategy because of its very high variation: it could cause a large drop of the capital due to successive erroneous choices between exploration and exploitation and subsequently cause the disappearance of the firm.

However, on the long run, the average capital doesn't greatly improve as all firms are simultaneously learning.

	Meta-rules	GlobalWilson	LocalWilson
Standard Deviation	44.65	47.97	85.07
Average	912,50	862.19	962.75

Table 3. Summary statistics for the average capital of firms under different exploitation-exploration strategies

6 Conclusion

In this paper, we studied the exploration-exploitation dilemma and learning in the context of large scale economic systems. The experiments demonstrated the interest of directed exploration techniques for firms: decreasing the learning time. However, these techniques are insufficient to improve the performance of the firm. We proposed an adaptive approach that determines the choice between exploration and exploitation. This approach is based on meta-rules that adapt the choice to the evolution of the performance and knowledge of the firm. We also compare it to two adaptive techniques proposed by Wilson. The obtained results show that this method is promising. However, the adaptation of the rate of change of the meta-rules to the life span of the firm is needed.

References

1. Butz, M. V., Wilson, S. W.: An algorithmic description of XCS. Journal of Soft Computing, **6** (2002) 144–153.
2. Carmel, D. and Markovitch, S.: Exploration Strategies for Model-Based Learning in Multi-agent Systems. Autonomous Agents and Multi-agent systems. Nicholas Jennings and Katia Sycara and Michael Georgeff (eds.). **2(2)** (1999) 141–172.
3. Guessoum, Z., Rejeb, L. and Sigaud, O.: Using XCS to build adaptive agents. In Proc. AAMAS Symposium 2004. Leeds (2004)
4. Penrose, E. T.: The theory of the growth of the firm. Basil Blackwell, (1959).
5. Peres-Uribe, A., Hirsbrunner, B.: The risk of Exploration in multi-agent learning systems: a case study. Proc. Agents-00 Joint workshop on learning agents, Barcelona, June 3–7, (2000) 33–37.
6. Sutton, R. S. and Barto, A.G.: Reinforcement learning, an introduction. The MIT Press, (1998).
7. Thrun S. B. : The role of exploration in learning control. In D A. Sofge (eds.). Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches. Florence, Kentucky:Van Nostrand Reinhold (1992).
8. Wiering, M.: Explorations in Efficient Reinforcement Learning. Ph.D. thesis. February (1999).
9. Wilson, S.W. : Classifiers Fitness Based on Accuracy. Evolutionary computation, Volume **3(2)**.(1995) 149-175
10. Wilson, S.W., : Explore/Exploit Strategies in Autonomy. In P. Maes, M. Mataric, J. Pollac, J.-A. Meyer and S. Wilson eds. From Animals to Animats 4, Proc. of the 4th International Conference of Adaptive Behavior, Cambridge (1996).