Combinando Modelos de Interação para Melhorar a Coordenação em Sistemas Multiagente

Richardson Ribeiro [1]

Adriano F. Ronszcka [2]

André P. Borges [3]

Bráulio C. Ávila [3]

Edson E. Scalabrin [3]

Fabrício Enembreck [3]

Resumo. A contribuição principal deste artigo é a implementação de um método híbrido de coordenação a partir da combinação de modelos de interação desenvolvidos anteriormente. Os modelos de interação são baseados no compartilhamento de recompensas para aprendizagem com múltiplos agentes, no intuito de descobrir de maneira interativa políticas de boa qualidade. A troca de recompensas entre os agentes durante a interação é uma tarefa complexa e se realizada de forma inadequada pode ocasionar atrasos no aprendizado ou até mesmo causar comportamentos inesperados, tornando a cooperação ineficiente e convergindo para uma política não-satisfatória. A partir desses conceitos, o método híbrido utiliza as particularidades de cada modelo, reduzindo possíveis conflitos entre ações com recompensas de políticas diferentes, melhorando a coordenação dos agentes em problemas de aprendizagem por reforço. Resultados experimentais mostram que o

[1] Departamento de Informática (COINF) da Universidade Tecnológica Federal do Paraná (UTFPR), Pato Branco - Pr, 85503-390
richardsonr@utfpr.edu.br

[2] Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI) da UTFPR, Curitiba - Pr, 80230-901
ronszcka@gmail.com

[3] Programa de Pós-Graduação em Informática (PPGIa) da Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba - Pr, 80215-901
{andre.borges, avila, scalabrin, fabricio}@ppgia.pucpr.br

método híbrido é capaz de acelerar a convergência, conquistando rapidamente políticas ótimas mesmo em grandes espaços de estados, superando os resultados de abordagens clássicas de aprendizagem por reforço.

Abstract: The main purpose of this paper is to implement a hybrid method of coordination from the combination of interaction models previously developed. The interaction models are based on rewards sharing for multi-agent learning, in order to discover interactively good action policies. The exchange of rewards among the agents during their interaction is a complex task and if it is not carried out properly can cause delays in learning or even cause unexpected behaviors, making the cooperation inefficient and converging to a non-satisfactory policy. From these models, the hybrid method uses the advantages of each model, reducing conflicts between actions with different policies rewards, improving the agents' coordination in Reinforcement Learning problems. Experimental results show that the proposed hybrid method has been able to accelerate the convergence, achieving optimal policies quickly even in large state spaces, overcoming the results of classic Reinforcement Learning approaches.

1 Introdução

Sistemas multiagente podem ser constituídos por agentes adaptativos que interagem e cooperam para a resolução de determinadas tarefas utilizando técnicas de aprendizagem por reforço. Aprendizagem por reforço é um paradigma computacional de aprendizagem em que um algoritmo tenta maximizar uma medida de desempenho baseado nos reforços (recompensas ou punições) que recebe ao interagir com um ambiente [25]. Além da aprendizagem, os agentes de um sistema multiagente devem ter capacidade de cooperação e coordenação no intuito de resolver problemas que demandam habilidades inexistentes localmente ou poderiam comprometer a sua performance. Dessa forma, a coordenação dos agentes e a fragmentação de seus conhecimentos (normalmente representados por políticas de ação) tornam-se importantes para a convergência de um comportamento globalmente coerente, ou simplesmente, para a solução de um problema.

Algoritmos de aprendizagem por reforço tal como o *Q*-learning [29] podem ser utilizados para a descoberta de uma política ótima por um único agente quando este explora repetidamente o espaço estado-ação. Uma política de ação é o mapeamento entre estados e ações,

que determina qual ação *a* deve ser executada em um estado *s*. Uma política de ação tem um estado inicial, um estado objetivo e um conjunto de transições de estados determinadas por diferentes ações, gerando um processo decisório de *Markov*. Entretanto, esta abordagem tende a ser ineficiente em problemas onde o espaço estado-ação é grande. Nestes casos, aprendizagem por reforço com múltiplos agentes tem se mostrado uma alternativa interessante [18]. Em aprendizagem por reforço com múltiplos agentes, cada agente pode ter uma política parcial, *i.e.* cada agente pode ter acesso a parte do conhecimento global. O objetivo é os agentes interagirem entre si para melhorarem seus conhecimentos sobre o ambiente e também melhorem a qualidade global da solução do problema representado na forma de estados e ações.

O processo de descoberta de políticas locais tem como objetivo final a descoberta de uma política global gerada a partir da combinação de conhecimentos dos indivíduos. Quando essa política é a melhor possível, denominada política ótima (Q^*), a mesma maximiza o total de recompensas recebidas pelos agentes. Portanto, os agentes devem compartilhar um modelo de interação cooperativo e coordenado para a troca de informação, visando melhores políticas Q.

A contribuição principal deste trabalho é conceber um método híbrido de coordenação para aprendizagem por reforço com múltiplos agentes que utiliza modelos explícitos macro e micro para a coordenação dos agentes, integrando as características dos modelos de interação estudados anteriormente. É mostrado que a interação entre os agentes empregando o método híbrido melhora a utilidade da política quando as recompensas são compartilhadas, favorecendo o modelo de aprendizagem por reforço. O método proposto acelera a convergência dos agentes que interagem com diferentes modelos de interação, produzindo resultados mais consistentes. Nos resultados preliminares, nós comparamos o desempenho do agente com cada modelo de interação e observamos que o agente pode alterar sua convergência devido às particularidades do ambiente, aprendizado e de modelo.

Este artigo está organizado da seguinte maneira: na Seção 2 é apresentado o estado da arte, no qual (i) são descritos alguns métodos de coordenação para aprendizagem em sistemas multiagente; (ii) revisão do algoritmo Q-learning; e (iii) aprendizagem por recompensas usando modelos de interação que compartilham recompensas. Na Seção 3 é apresentado o método híbrido para coordenação, o ambiente de avaliação e os principais resultados comparando o método híbrido com os modelos de interação. O trabalho é finalizado na seção 4 com algumas conclusões e perspectivas para trabalhos futuros.

2 Revisão da Literatura

A aprendizagem em sistemas multiagente, diferentemente da aprendizagem em ambiente com um único agente, supõe que o conhecimento relevante não está disponível localmente em um único agente. Na aprendizagem multiagente [32], os agentes aprendem a realizar uma tarefa que envolve mais do que um agente na sua execução. Segundo Weiss e Sen (1996) [30], a aprendizagem pode ser dividida de duas formas: a aprendizagem isolada e a aprendiza-

gem coletiva. Na aprendizagem isolada, o processo de aquisição do conhecimento pelo agente ocorre sem a influência dos demais agentes ou qualquer outro elemento da sua sociedade. Já na aprendizagem coletiva, o processo de aquisição do conhecimento tem influência direta de todos os elementos da sociedade em que o agente está inserido.

Stone e Veloso (1996) [24] completam ainda que, se um agente está aprendendo a conquistar habilidades para interagir com outros agentes em seu ambiente, e independentemente se os outros agentes estão ou não aprendendo simultaneamente, esta aprendizagem é considerada aprendizagem multiagente. Dessa forma, aprendizagem multiagente inclui algumas situações na qual o agente aprende interagindo com outros agentes, alterando e evoluindo o próprio modelo de coordenação.

2.1 Coordenação por Interação

A interação propicia a combinação de esforços entre um conjunto de agentes na busca de soluções para problemas globais, pressupondo ações de coordenação entre os agentes [6]. Alguns aspectos podem ser considerados no processo de interação dos agentes: (i) quais agentes devem interagir; (ii) em qual momento ocorrerá a interação; (iii) qual o conteúdo da interação ou comunicação; (iv) como será realizada a interação, definindo os processos e recursos a serem utilizados; (v) definir se a interação é necessária; e (vi) empregando algum mecanismo, de que maneira será estabelecida a compreensão mútua (linguagem comum, interpretação baseada no contexto, etc.).

Uma situação de interação é um conjunto de comportamentos resultantes de um grupo de agentes que agem para satisfazer seus objetivos, e que levam em conta as restrições devidas a limitações de recursos e à limitação de suas competências individuais [7]. Considerando que a interação entre agentes pode ocorrer através de ações para atingir seus objetivos, agentes realizam ações, que podem eventualmente utilizar recursos, consumíveis ou não, que se encontram disponíveis no ambiente. Tais situações de interação podem ser classificadas de acordo com as dimensões distintas [23]:

> i) compatibilidade de objetivos: os objetivos dos agentes são considerados compatíveis/incompatíveis quando o fato de atingir um deles não acarretar/acarretar necessariamente a impossibilidade de atingir o outro;
>
> ii) quantidade de recursos: os recursos são considerados suficientes/insuficientes quando os agentes puderem/não puderem realizar suas tarefas simultaneamente;
>
> iii) competência dos agentes: a competência de um agente é considerada suficiente/insuficiente quando ele for capaz/incapaz de realizar sua tarefa, de modo a atingir seu objetivo.

Muitas vezes, a interação entre os agentes está diretamente relacionada a um mecanismo de aprendizagem. Em sistemas multiagente, a aprendizagem está diretamente relacionada

com a maneira que os agentes interagem, podendo comprometer a convergência parcial ou total da aprendizagem dos agentes ou até mesmo causar situações inexplicáveis nas suas ações, devido a conflitos entre comportamentos e objetivos e limitação de recursos e habilidades. Por exemplo, em cenários clássicos como monitoramento com sensores distribuídos [3], alocação distribuída de tarefas [1], e formação de coalizão [2], cada agente percebe uma parte do estado global do cenário e toma medidas que modificam alguma parte deste estado, na intenção de maximizar uma função de utilidade local [27].

Dessa maneira, os agentes devem ser capazes de interagir em um ambiente comum, trocando informações relevantes e cooperando com os indivíduos que podem contribuir para conquistar um determinado objetivo. Na literatura são encontrados diversos trabalhos que descrevem diferentes formas de aprendizagem a partir da interação [7, 31], aprendizagem coletiva ou social [23].

Um modelo de aprendizagem por interação possui um conjunto de comportamentos resultantes do grupo de agentes que agem para satisfazer seus objetivos e ainda consideram as restrições impostas pela limitação de recursos e pelas competências individuais. Em problemas de aprendizagem usando métodos de aprendizagem por reforço, a interação depende de um modelo que possibilita a troca das melhores recompensas acumuladas e dos reforços imediatos da transição. Com esse objetivo, Chapelle *et al.* (2002) [4] propuseram um modelo por interações onde o valor da recompensa é calculado usando a satisfação individual dos agentes vizinhos. No processo de aprendizagem, os agentes continuadamente emitem um nível de satisfação pessoal. Por exemplo, se a ação do agente *A* no ambiente *E* pode ajudar o agente *B*, o nível da satisfação de *B* também aumenta. O processo de aprendizagem ocorre até que todos os agentes vizinhos alcancem um nível satisfatório para as recompensas recebidas.

Em um trabalho anterior [156] foi desenvolvida uma estratégia de aprendizagem na qual o agente inserido no sistema é capaz de manter na memória políticas aprendidas para serem reusadas nas políticas futuras, evitando atrasos ou falta de convergência na aprendizagem do agente com a dinâmica no ambiente. O método proposto, denominado por nós de política adaptativa baseada em recompensas passadas (*k*-learning), foi testado em cenários com características de trânsito com diferentes níveis de congestionamento com até 64 estados. Resultados experimentais mostraram que o método proposto é melhor do que o algoritmo *Q*-learning padrão, pois estima valores e encontra soluções usando políticas passadas. O método também foi testado em ambientes dinâmicos e com espaço de estados de tamanho da ordem de centenas [20].

Mataric (1998) [14] propôs um método onde os agentes podem, de forma cooperativa, transmitir para outros agentes a situação atual do estado alterado, após a realização de uma tarefa. Neste caso, um agente somente poderá dividir seu aprendizado com o agente mais próximo do seu estado atual, a fim de economizar recursos e evitar troca de informações incorretas.

DeLoach e Valenzuela (2007) [6] propuseram um mecanismo chamado de modelo de capacidade. O modelo visava demonstrar como os agentes interagiam em um dado ambiente,

colocando em evidência o uso de suas capacidades. Esse modelo é composto pelos seguintes elementos: um modelo de capacidade, um ambiente e um conjunto de interações entre os objetos do modelo de capacidade e do ambiente. Neste caso, cada agente tem a capacidade de perceber e manipular os objetos do ambiente por meio de interações, aprendendo/refinando estratégias para alcançar seu objetivo. O modelo de capacidade define as ações possíveis que cada agente pode realizar com o objetivo de manipular os objetos do ambiente. Quando executada uma ação, essa atividade recebe uma recompensa do ambiente. Se a ação modifica o objeto, o ambiente consequentemente é alterado com as recompensas recebidas.

Na maioria das vezes é difícil adaptar os métodos propostos em um modelo genérico de coordenação, devido à diversidade das classes de problemas existentes e o demasiado conhecimento do domínio exigido. Além disso, essas interações podem não compartilhar as melhores informações de algoritmos baseados em recompensas, pois não consideram a reputação de cada agente e acabam ocasionando a troca de informações não satisfatórias.

Além da interação, cada agente deve ser capaz de aprender e cooperar no ambiente. Em ambientes complexos, um único agente só pode, ao longo do tempo, adquirir experiências suficientes que convergem para uma política ótima, se e somente se uma grande quantidade de episódios é possível, bem como se estratégias específicas são utilizadas para evitar máximos locais. No entanto, em um sistema com vários agentes, valores contraditórios para recompensas acumuladas podem ser gerados, à medida que cada agente utiliza apenas valores locais de aprendizagem. Dessa forma, a aprendizagem coletiva, diferentemente da aprendizagem com um único agente, pressupõe que o conhecimento relevante ocorre quando recompensas são compartilhadas, intensificando a relação entre os agentes.

2.2 Algoritmo *Q*-learning

A aprendizagem por reforço tenta solucionar problemas onde um agente recebe do ambiente um retorno (recompensas ou punições) por suas ações. Este tipo de aprendizagem vem sendo utilizado em diversos trabalhos [5, 9, 10, 12, 16, 26, 28], a fim de encontrar soluções para problemas complexos com o uso de agentes. Na aprendizagem por reforço, os agentes atuam em um ambiente formado por um conjunto de estados e podem selecionar ações dentro de um conjunto de ações possíveis, indicando o valor imediato da transição de estado resultante [13]. Agentes dotados de algoritmos de aprendizagem por reforço devem aprender uma política de ação que maximize a soma esperada desses reforços, descontando as recompensas ou punições proporcionais ao seu atraso temporal.

Um método bastante conhecido para resolver problemas de aprendizagem por reforço em sistemas multiagente é o algoritmo *Q*-learning [29]. Antes de discutirmos tal algoritmo, é apresentado brevemente o processo decisório de *Markov*, usado para formalizar a AR. Um processo decisório de *Markov* é uma tupla (S, A, $T^{a}_{s,s'}$, $R^{a}_{s,s'}$), onde S é um conjunto finito de estados do ambiente, que pode ser composto por uma sequência variável de estados = $< x_1, x_2,..., x_v >$; A é um conjunto finito de ações, no qual um episódio é uma sequência de ações $a \in$

A que leva o agente de um estado $s_{inicial}$ a um estado $s_{objetivo}$, $T^{a}_{s,s'}$ é a função de transição de estado $T : S \times A \rightarrow [0,1]$, que indica a probabilidade do agente chegar em um estado s' quando uma ação é aplicada no estado s. Similarmente, $R^{a}_{s,s'}$ é a função de recompensa $R : S \times A \rightarrow \Re$ recebida toda vez que a transição $T^{a}_{s,s'}$ ocorre.

O objetivo de um agente de aprendizagem por reforço é aprender uma política $\pi : S \times A$ que mapeia o estado atual s em uma ação desejada, de forma a maximizar a recompensa acumulada ao longo do tempo, descrevendo o comportamento do agente [11]. Para alcançar uma política ótima, o algoritmo de aprendizagem por reforço deve explorar iterativamente o espaço de estado $S \times A$, atualizando as recompensas acumuladas e armazenando essas recompensas em Q(s,a) (Equação 1).

A ideia básica do Q-learning é que o algoritmo de aprendizagem aprenda uma função de avaliação ótima sobre todo o espaço de pares estado-ação $S \times A$. A função Q fornece um mapeamento da forma Q: $S \times A \rightarrow V$, onde V é o valor de utilidade esperada ao se executar uma ação a no estado s. Desde que o particionamento do espaço de estados do agente e o particionamento do espaço de ações não omitam informações relevantes, uma vez que a função ótima seja aprendida, o agente saberá que ação resultará na maior recompensa futura em uma situação particular s.

A função Q(s,a), da recompensa futura esperada ao se escolher a ação a no estado s, é aprendida através de tentativa e erro, conforme Equação 1:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R + (y \times V)] \quad (1)$$

no qual $\alpha \in [0,1]$ é a taxa de aprendizagem, R é a recompensa, ou custo, resultante de tomar a ação a no estado s, y é o fator de desconto e V é obtido utilizando a função Q aprendida até o presente. A função Q representa a recompensa descontada esperada ao se tomar uma ação a quando visitando o estado s, e seguindo-se uma política ótima desde então. A descrição detalhada do algoritmo Q-learning é encontrada em [29].

Ghavamzadeh e Mahadevan (2002) [8] utilizaram o Q-learning para desenvolver um algoritmo de aprendizagem por diferença temporal, chamado MAPLE (*Multi-Agent Policy Learning*). O MAPLE usa o Q-learning para gerar soluções globais eficientes para problemas multiagente a partir de soluções de PDM individuais.

Nós empregamos o algoritmo Q-learning para gerar e avaliar políticas de ação parciais e globais. Aplicando o Q-learning é possível encontrar uma política para um agente. No entanto, se agentes similares interagem no mesmo ambiente, cada agente possui seu próprio processo decisório de *Markov* no qual o comportamento ótimo pode não estar definido. Dessa forma, em um ambiente formado por diversos agentes, o objetivo é selecionar as ações de cada processo decisório de *Markov* no tempo t, de forma que o total de recompensas esperadas para todos os agentes envolvidos seja maximizado.

2.3 Aprendizagem por Recompensas Compartilhadas

A aprendizagem por reforço com múltiplos agentes baseada em recompensas compartilhadas pode produzir um conjunto refinado de comportamentos obtidos a partir das ações tomadas. Parte do conjunto de comportamentos (*i.e.*, uma política global) é compartilhada pelos agentes por meio de uma política de ação parcial (Q_i). Geralmente, tais políticas parciais contêm informações (valores de aprendizagem) incompletas sobre o ambiente, mas, com um modelo para compartilhar as recompensas, essas podem ser integradas para maximizar a soma das recompensas parciais obtidas ao longo da aprendizagem. Quando políticas $Q_1,\ldots,Q_x$ são integradas, é possível formar uma nova política, denominada de política de ação baseada em recompensas partilhadas $\partial = \{Q_1,....,Q_x\}$, na qual ∂(s,a) é uma tabela que denota as melhores recompensas adquiridas pelos agentes durante o processo de aprendizagem.

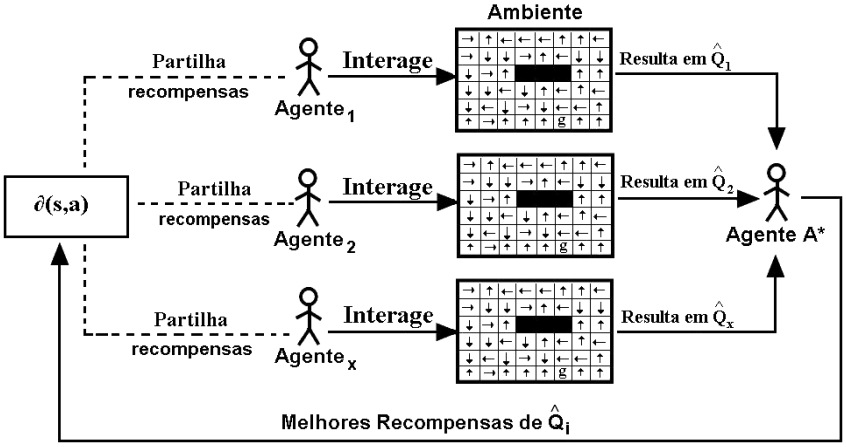

Figura 1. Interação com recompensas partilhadas [18]

Na Figura 1 é mostrado como os agentes trocam informações ao longo das interações. Empregando o algoritmo Q-learning, um Agente$_i$ gera e armazena as recompensas em $\hat{Q}_i$. Quando o agente A* recebe as recompensas, o seguinte procedimento é realizado: quando o Agente$_i$ alcança o estado objetivo g a partir do estado inicial $s \neq g$ com um caminho de menor custo, o agente usa um modelo para compartilhar as recompensas com os demais agentes. Os valores de aprendizagem de uma política parcial Q_i podem ser utilizados para atualizar a política global $\hat{\partial}$(s,a) disponível, interferindo posteriormente na forma como os demais agentes atualizam seus conhecimentos e interagem com o ambiente.

No Algoritmo 1 é apresentada a função que compartilha as recompensas dos agentes. Essa tarefa pode ser realizada de três formas e são discutidas na subseção 2.3.1, sendo que

todas elas utilizam internamente um algoritmo de aprendizagem por reforço (Q-learning). As melhores recompensas de cada agente são enviadas para a ∂(s,a), formando uma nova política com as melhores recompensas adquiridas pelos agentes $I = \{i_1,...i_x\}$, que na sequência podem ser socializadas com os demais agentes. Para estimar ∂(s,a), é usada uma função custo, que encontra o caminho de menor custo do estado inicial ao estado objetivo em uma dada política. A descoberta desse caminho é realizada com o algoritmo A* que produz um modelo generativo para administrar a política que maximiza a recompensa total esperada, *i.e.*, política ótima, de acordo com a metodologia apresentada em [20].

Os seguintes elementos são necessários para a compreensão e formalização dos modelos de aprendizagem por reforço propostos:

- Um conjunto de estados $S = \{s_1,...,s_m\}$;
- Um conjunto de agentes $I= \{i_1,...,i_x\}$;
- Políticas de ação parcial $\{Q_1,....,Q_x\}$, onde Q_i representa a política parcial do agente i;
- Uma função de recompensa $st(S) \rightarrow ST$, onde $ST= \{-1; -0,4; -0,3; -0,2; -0,1\}$;
- Um instante de tempo de valor discreto $t= 1,2,3,...,n$;
- Um episódio de tempo c, onde $c < n$;
- Um conjunto de ações $A= \{a_1,...,a_k\}$, onde cada ação é executada no tempo t;
- Uma tabela de aprendizagem $\hat{Q}: (S \times A) \rightarrow \Re$, que define uma política Q;
- Um conjunto de modelos de compartilhamento de recompensas $M=$ {discreto, contínuo, dirigido por objetivo};
- Uma função custo: $custo(s,g) = \sum_{s \in S}^{g} 0.1 + \sum_{s \in S}^{g} st(S)$ usada para calcular o custo de um episódio (caminho do estado inicial s até o estado objetivo g) baseado na política atual;
- Uma função que define o modelo de partilha f: $(t \times M \times I \times S \times A) \rightarrow \partial$(s,a), onde t é a condição de parada do modelo;
- Uma política ótima objetivo Q^* estimada com um algoritmo supervisor (A*);

O Algoritmo 1 deve ser interpretado da seguinte forma:

- Linhas 2-8: Inicialização da $\hat{Q}_i(s,a)$;

- Linha 10: Interação dos agentes $i \in I$;
- Linha 11: A função *f* seleciona um modelo de partilha; no qual (t, modelo, Q_i, s, a) são os parâmetros, onde *t* é a iteração corrente, *modelo* ∈ {discreto, contínuo, dirigido por objetivo}, *s* e *a* são o estado e a ação escolhidos respectivamente da política Q_i;
- Linha 14: Para cada par estado-ação é empregada a regra de atualização que calcula os valores de recompensas;
- Linhas 17-18: $\hat{Q}_i$ do agente $i \in I$ é atualizado com ∂(s,a);

Algoritmo AR_cooperativo (I, modelo)

Tabela de aprendizagem: Q_i, Q^*, ∂

01 $\partial(s,a) \leftarrow 0$;
02 **Para** cada agente i ∈ I **faça:**
03 **Para** cada estado s ∈ S **faça:**
04 //inicialização dos valores de aprendizagem
Para cada ação a ∈ A faça:
05 $\hat{Q}_i(s,a) \leftarrow 0$;
06 **Fimpara**
07 **Fimpara**
08 **Fimpara**
09 t ← 0;
10 **Para** cada agente i ∈ I **faça:**
11 **Enquanto não** f(t, modelo, Q^*, Q_i, ∂(s,a)) **repita:**
12 T ← t + 1;
13 **Escolha** estado s ∈ *S*, ação *a* ∈ *A*
Atualize:
14 $$V \leftarrow \gamma \max_{a_{t+1}} \hat{Q}_t(s_{t+1}, a_{t+1}) - \hat{Q}_t(s_t, a_t);$$
$$\hat{Q}_{t+1}(s_t, a_t) \leftarrow \hat{Q}_t(s_t, a_t) + \alpha\left[R(s_t, a_t) + V\right];$$
15 **Fimenquanto**
16 **Fimpara**
17 **Para** cada agente i ∈ I **faça:**
18 $\hat{Q}_i(s,a) \leftarrow \partial(s,a)$;
19 **Fimpara**
20 **Fim**

Algoritmo 1. Algoritmo de aprendizagem por reforço cooperativo [18]

Algoritmo f (t, modelo, Q^*, Q_i, ∂(s,a))

c: número de ciclos
01 **Escolha** (modelo):
02 **Caso** "discreto":
03 **Se** t **mod** *c* == 0 **então**
04 ∂(s,a) ← atualiza_política(Q^*, Q_i, ∂(s,a))

```
05        Fimse
06      Caso "contínuo":
07        Se s ≠ g então
08          r ← Σ_{i=1}^{x} Q̂_i(s,a);
09          Q̂_i(s,a) ← r;
10          ∂(s,a) ← atualiza_política(Q*, Q_i, ∂(s,a))
11        Fimse
12      Caso "dirigido_por_objetivo":
13        Se s == g então
14          r ← Σ_{i=1}^{x} Q̂_i(s,a);
15          Q̂_i(s,a) ← r;
16          ∂(s,a) ← atualiza_política(Q*, Q_i, ∂(s,a))
17        Fimse
18      Fimescolha
19      Retorne(∂(s,a))
```

Algoritmo 2. Modelos de interação

```
Algoritmo atualiza_política(Q_i, Q*, ∂(s,a))
01      Para cada estado s ∈ S faça:
02        Se custo(Q_i,s) ≤ custo(Q*,s) então
03          ∂(s,a) ← Q̂_i(s,a);
04        Fimse
05      Fimpara
06    Retorne(∂(s,a))
```

Algoritmo 3. Atualiza política

Para auxiliar a compreensão dos pseudocódigos, é ilustrado na Figura 2 o diagrama de atividades que utiliza os Algoritmos 1, 2 e 3.

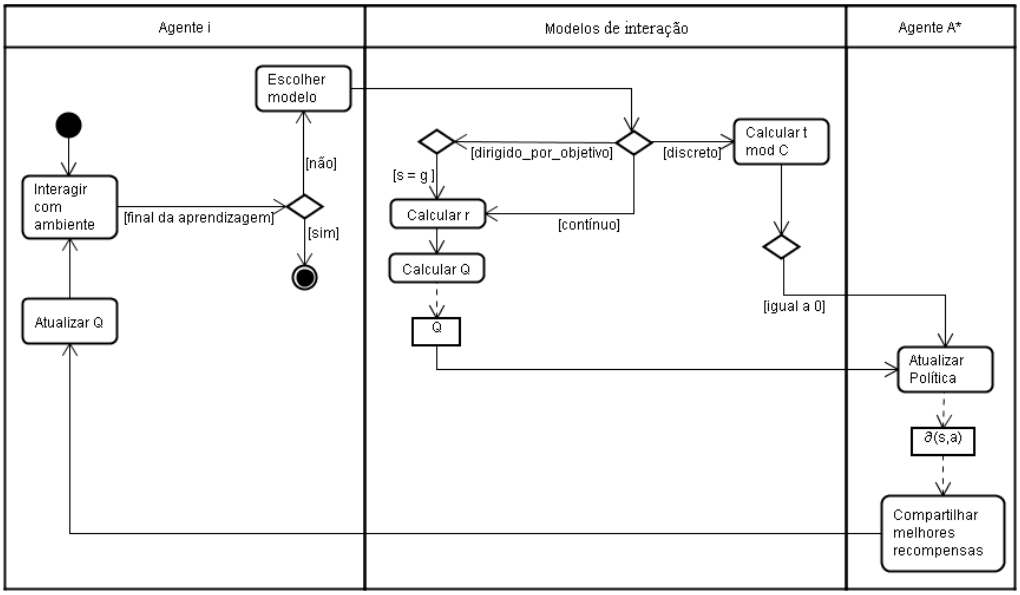

Figura 2. Diagrama de atividade do processo de aprendizagem com os modelos de interação

2.3.1 Modelos de Interação

Em um trabalho anterior, nós estudamos modelos de interação para aprendizagem por reforço com múltiplos agentes. Nesta subseção, nós sumarizamos os modelos e os resultados apresentados em [18].

Modelo Discreto: Os agentes cooperam compartilhando seu aprendizado em um ciclo predefinido de passos *c*. A cooperação no *modelo discreto* ocorre da seguinte maneira: o agente acumula as recompensas obtidas de suas ações durante o ciclo de aprendizagem. No final do ciclo, cada agente envia os valores da $\hat{Q}_i$ para a ∂(s,a). Se o valor da recompensa é adequado, *i.e.*, se ele melhora a eficiência dos demais agentes para o mesmo estado (linha 2 do Algoritmo 3), os agentes compartilham essas recompensas (linha 18 do Algoritmo 1).

Modelo Contínuo: Os agentes cooperam compartilhando o valor do reforço obtido em cada transição $T^{a}_{s,s'}$. A cooperação no *modelo contínuo* ocorre da seguinte maneira: se $s \neq g$, então toda ação executada pelo agente gera um valor de reforço (positivo ou negativo), que é a soma dos reforços acumulados para todos os agentes na ação *a* em um estado *s* (linha 8 Algoritmo 2). O objetivo é acumular recompensas maiores na $\hat{Q}_i$ que pode ser compartilhada a cada iteração (linha 18 Algoritmo 1). Isso mostra que recompensas adquiridas de várias políticas geradas separadamente podem gerar uma política Q^*.

Modelo Dirigido a Objetivo: A cooperação no *modelo dirigido a objetivo* acontece da seguinte maneira: diferentemente do *modelo discreto*, a cooperação ocorre quando o agente alcança o estado objetivo, *i.e.*, $s == g$ (linha 13 Algoritmo 2). Neste caso, o agente interage acumulando valores de recompensas. Isso é necessário porque nesse modelo o agente compartilha suas recompensas em número variável de passos, ou seja, somente quando o estado objetivo é alcançado. Quando o agente conquista o estado objetivo, o valor das recompensas adquiridas é enviado para a ∂(s,a). Se o valor da recompensa do estado melhora a eficiência global, então os agentes compartilham tais recompensas. Isso mostra que mesmo compartilhando recompensas baixas e não satisfatórias no início da aprendizagem, o agente é capaz de se adaptar, sem prejudicar a convergência global.

Os experimentos comparando os modelos foram testados em ambientes de até 400 estados, usando de 3 a 10 agentes. Resultados experimentais mostram que os modelos propostos são melhores do que o algoritmo *Q*-learning padrão: (i) empregando o modelo discreto, as políticas de ação descobertas pelos agentes são próximas da ótima (Q^*), consequência dos valores de reforços acumulados pelos agentes em iterações pré-definidas para uma boa convergência; (ii) com o modelo contínuo, a convergência ocorre com poucas iterações, devido aos valores de reforços compartilhados no início da aprendizagem. É observado com esse modelo que estados mais próximos do estado-objetivo acumulam reforços com valores altos, o que resulta em um máximo local; e (iii) com o modelo dirigido por objetivo, os agentes compartilham o aprendizado em um número variável de iterações, compartilhando bons valores de reforços desde que haja uma quantidade de iterações suficiente para acumular valores de recompensas satisfatórias.

Em um trabalho relacionado, Miyazaki e Kobayashi (2005) [15], propuseram um modelo de coordenação baseado na teoria da racionalidade de recompensas compartilhadas. O modelo foi testado em problemas de aprendizagem por reforço multiagente. Foi usado um conjunto de regras que determinava o sinal da recompensa pelas ações dos agentes. Quando um agente recebe recompensas maiores que zero (recompensas diretas), era atribuída aos demais agentes uma recompensa indireta, computada conforme o número de agentes no sistema, conflitos nas regras e a racionalidade de cada regra usada pelos agentes. A teoria utilizada mostra alguns aspectos eficientes no compartilhamento de recompensas, como a melhoria na qualidade e a aceleração da aprendizagem.

3 Um Método Híbrido para Melhorar a Coordenação Multiagente

Em aprendizagem por reforço baseada em recompensas compartilhadas, a descoberta de políticas não satisfatórias pode ocorrer em um dado momento, pois a troca de conhecimento entre os agentes pode gerar novas políticas intermediárias incompatíveis com uma rápida convergência (máximos locais). Como há mudanças no aprendizado de cada agente, é necessário que todos os agentes estejam atualizando e trocando suas recompensas enquanto interagem.

Não há garantia de convergência da política de ação baseada em recompensas compartilhadas usando os modelos descritos anteriormente. Observou-se empiricamente que políticas com estados e valores de recompensas inadequados podem sofrer modificações com recompensas informadas por outras políticas parciais, melhorando a $\partial(s,a)$. No entanto, pode ocorrer o efeito contrário, onde políticas com estados com altas recompensas podem se tornar menos interessantes para a política corrente, pois estados que produziam acertos passam a produzir erros.

Para superar esse problema (máximos locais) foi desenvolvido um Método Híbrido de Coordenação. Essa abordagem surgiu a partir dos modelos discreto, contínuo e dirigido por objetivo e de constatações observadas nos experimentos. Pode-se notar que o comportamento da $\partial(s,a)$ com os modelos se altera em função da quantidade de iterações, de episódios e da quantidade de agentes. Nas Figuras 6-8 são ilustrados os modelos em ambientes com 400 estados.

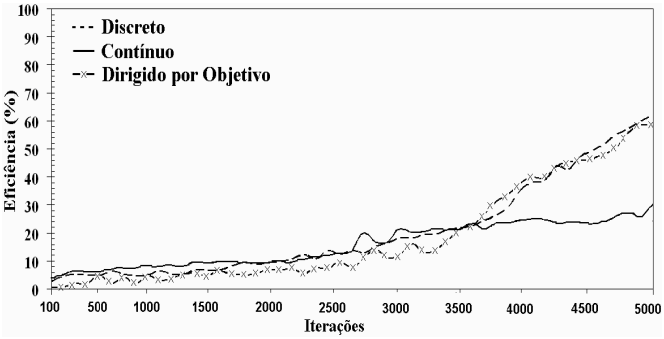

Figura 6. 400 estados, 3 agentes [18,22]

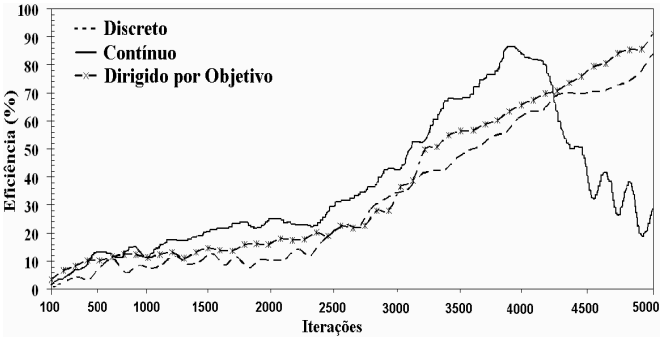

Figura 7. 400 estados, 5 agentes [18, 22]

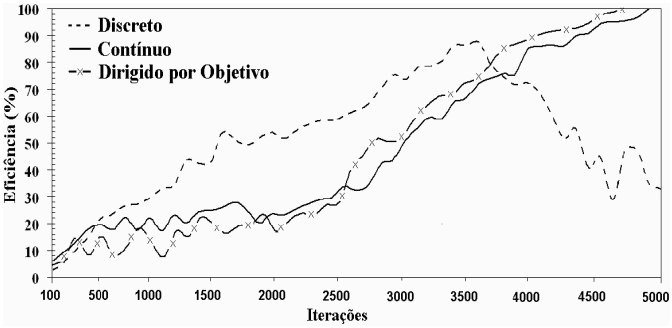

Figura 8. 400 estados, 10 agentes [18, 22]

O método híbrido utiliza as particularidades de cada modelo, permitindo a utilização das melhores características. Essa abordagem descobre novas políticas de ação sem causar atrasos na aprendizagem, reduzindo possíveis conflitos entre ações com recompensas de políticas diferentes, melhorando a convergência. O método híbrido funciona da seguinte forma: a cada iteração do algoritmo *Q*-learning com um modelo de interação, o desempenho do agente com os modelos é comparado, gerando uma nova tabela de aprendizagem, de nome MH-∂(s,a) (Método Híbrido). Quando a condição de atualização do modelo é alcançada, o agente inicia o aprendizado utilizando o modelo de melhor desempenho e a aprendizagem é transferida para a MH-∂(s,a). Portanto, a MH-∂(s,a) terá as melhores recompensas adquiridas dos modelos discreto, contínuo e dirigido por objetivo. No Algoritmo 4 é apresentado o pseudocódigo que integra os modelos de cooperação discreto, contínuo e dirigido a objetivo.

```
Algoritmo Método_Híbrido()
M : discreto, contínuo, dirigido a objetivo // modelos
Qi : QM_discreto, QM_contínuo, QM_dirigido_objetivo //tabelas de aprendizagem
1.    Para cada Qi ∈ M faça
2.      Se QM_discreto *> QM_contínuo E
3.         QM_discreto *> QM_dirigido_objetivo Então
4.         MH-∂(s,a) ← QM_discreto
5.      Fimse
6.      Se QM_contínuo *> QM_discreto E
7.         QM_contínuo *> QM_dirigido_objetivo Então
8.         MH-∂(s,a) ← QM_contínuo
```

```
9.      Senão MH-∂(s,a) ← Q_M_dirigido_objetivo
10.    Fimse
11.  Fimpara
12. Retorne (MH-∂(s,a))
```

Algoritmo 4. Algoritmo do método híbrido

É possível observar no Algoritmo 4 que, para todo ciclo de aprendizagem, os desempenhos do agente com os modelos de interação são comparados (através do operador '*>', que compara duas tabelas de aprendizagem). Quando o desempenho de um modelo é superior, essa aprendizagem é transferida para a MH-∂(s,a), que representa a política de ação corrente com o método híbrido. Antes de discutirmos os resultados com o método híbrido, é apresentado, na Subseção 3.1, o ambiente de simulação.

3.1 Ambiente de Simulação

Em um ambiente onde a aprendizagem é dinâmica, os valores de recompensas são alterados continuamente pela política ou pelo modelo de coordenação, tornando o processo dinâmico. Em ambientes com múltiplos agentes, os agentes podem ser influenciados pelas atitudes ou ações dos demais agentes. Nesse caso, a ausência de mecanismos de coordenação pode influenciar a aprendizagem, causando situações inexplicáveis nas suas ações, devido a interdependência dos agentes e conflitos entre objetivos e habilidades.

Os modelos de interação propostos neste trabalho vão ao encontro dessas observações, e são portanto utilizados, para coordenar as ações dos agentes em um sistema multiagente. Os modelos foram avaliados a partir de um ambiente artificial implementado com este objetivo. O ambiente de simulação[4] utilizado para avaliação dos modelos é constituído por um espaço de estados onde há um estado inicial ($s_{inicial}$), um estado objetivo (g) e um conjunto de ações A= {↑ (para frente), → (para a direita), ↓ (para trás), ← (para a esquerda)}. Um estado s é um par (X,Y) com coordenadas de posições nos eixos X e Y respectivamente. No ambiente há uma função status $st : S \rightarrow ST$ que mapeia os estados e situações de trânsito (recompensas) onde ST = {–0,1 (livre); –0,2 (pouco congestionado); –0,3 (congestionado ou desconhecido); –0,4 (muito congestionado); –1 (bloqueado); 1,0 (g)}. Os agentes são utilizados para simular movimentações de motoristas em um cenário de trânsito. O objetivo é encontrar um mapeamento entre estados e ações (políticas de ação), utilizadas para determinar os melhores caminhos entre um estado inicial e o estado objetivo, *i.e.*, qual ação a deve ser executada em um estado s.

Após cada movimento do agente (transição) de um estado s para o estado s', o agente sabe se sua ação foi positiva ou negativa por meio das recompensas atribuídas e/ou compartilhadas por outros modelos. A recompensa para a transição $T^{a}_{s,s'}$ é $st(s')$. Na Figura 9 são apresentados os cenários utilizados nas simulações e na Figura 10 é mostrada uma representação simplificada de um ambiente com uma política π. Os cenários foram gerados arbitrariamente,

[4] A descrição detalhado do simulador pode ser encontrada em Ribeiro (2006) [20].

na intenção de gerar situações próximas do mundo real. Os agentes são posicionados aleatoriamente no ambiente, e quando algum agente encontra o estado objetivo, esse é posicionado similarmente em outro estado do conjunto *S*.

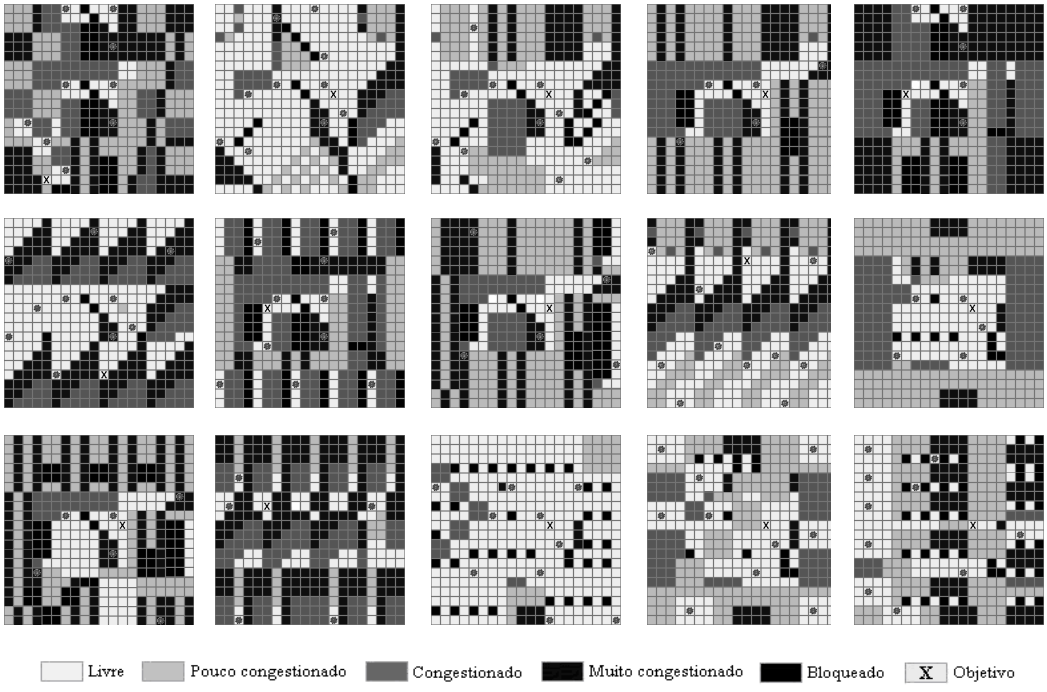

Figura 9. Cenários simulados

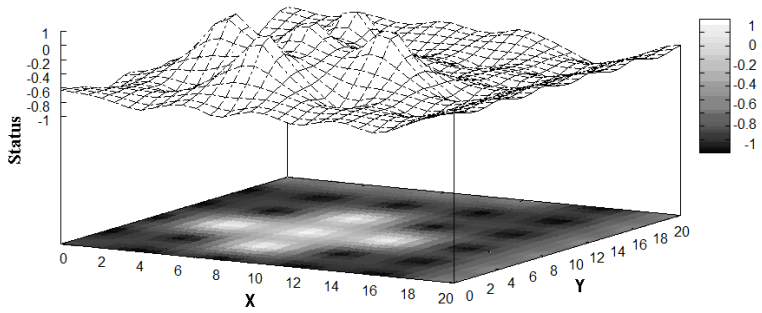

Figura 10. Exemplo de um cenário com 400 estados. Os agentes são posicionados aleatoriamente no ambiente e possuem campo de profundidade visual de 1

3.2 Método Híbrido *vs*. Modelos Contínuo, Discreto e Dirigido por Objetivo

Nesta subseção são apresentados os principais resultados comparando o método híbrido com os modelos discutidos anteriormente. Os parâmetros utilizados com o método híbrido são os mesmos dos demais modelos. A variação desses parâmetros pode ser encontrada em [18]. Os experimentos foram realizados em ambientes que variam entre 100 (10 × 10) e 400 (20 × 20) estados. Os resultados apresentados nas Figuras 11-19 comparam o método híbrido e os demais modelos com diferentes quantidades de agentes e estados.

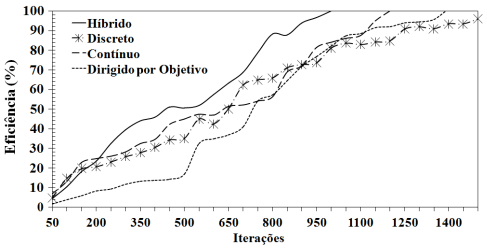

Figura 11. Ambiente de 100 estados; 3 agentes

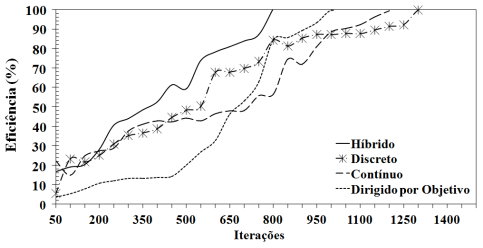

Figura 12. Ambiente de 100 estados; 5 agentes

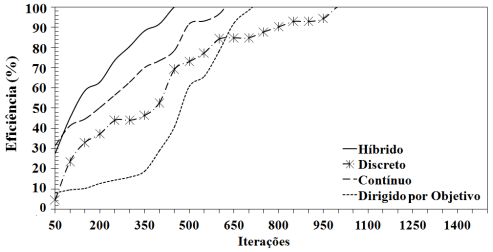

Figura 13. Ambiente de 100 estados; 10 agentes

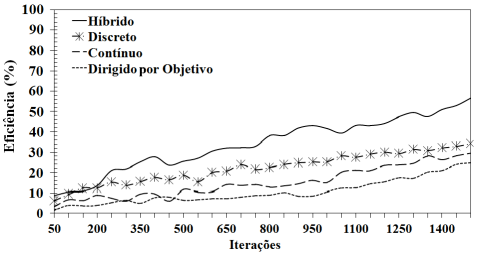

Figura 14. Ambiente de 250 estados; 3 agentes

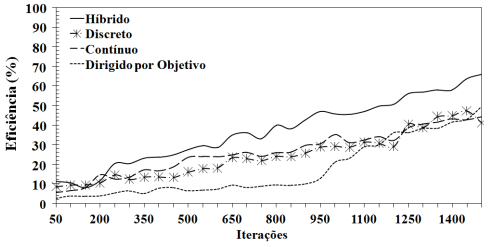

Figura 15. Ambiente de 250 estados; 5 agentes

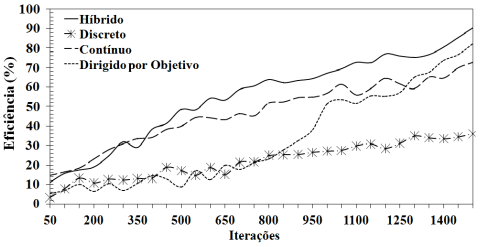

Figura 16. Ambiente de 250 estados; 10 agentes

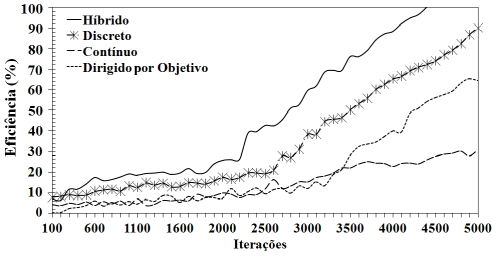

Figura 17. Ambiente de 400 estados; 3 agentes

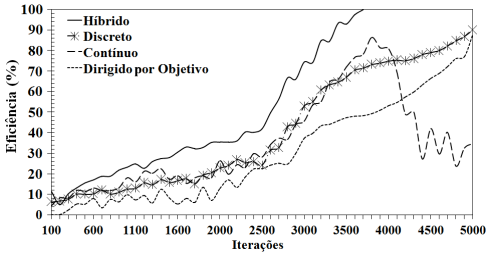

Figura 18. Ambiente de 400 estados; 5 agentes

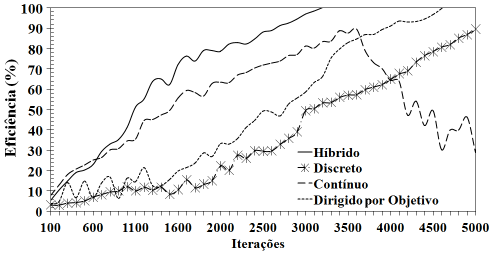

Figura 19. Ambiente de 400 estados; 10 agentes

Nas Figuras 11-19 é mostrado que o método híbrido apresenta eficiência superior em relação às políticas geradas com os modelos discreto, contínuo e dirigido por objetivo. Geralmente, o método híbrido obtém desempenho superior em qualquer fase das iterações. Isso fez com que diminuísse significativamente o número de iterações necessárias para encontrar uma boa política de ação. Em ambientes com 100 estados o número de iterações diminuiu aproximadamente 23,4% quando utilizados 3 agentes; 27,1% com 5 agentes e 40,4% com 10 agentes. Nos ambientes com 250 estados o método híbrido consegue diminuir o número de iterações em 20% com 3 agentes; 22,8% com 5 agentes e 33,9% com 10 agentes. Já nos ambientes com 400 estados o número de iterações diminuiu 18,6% com 3 agentes; 23,3% com 5 agentes e 28,7% com 10 agentes. Na Tabela 1 é sumarizada a superioridade média da eficiência do método híbrido comparado com o melhor dentre os demais modelos.

Tabela 1. Superioridade média do método híbrido em relação aos modelos discreto, contínuo e dirigido por objetivo

	Quantidade de Agentes		
# Estados	**3**	**5**	**10**
100	17.6%	69.8%	37.2%
250	50.3%	31.4%	33.7%
400	45.1%	40.9%	37.6%

É possível observar que o desempenho dos agentes com o método híbrido é melhor do que o desempenho das políticas de ação baseadas em recompensas individuais. O bom desempenho dos agentes que cooperam utilizando o método híbrido é decorrente de $\partial(s,a)$ gerada a partir de valores de aprendizagem descobertos de forma colaborativa. Assim, os agentes conseguem gerar políticas com boas recompensas para aproximar os valores aprendidos de uma boa política. Ademais, o número de iterações diminuiu significativamente com o método híbrido. No entanto, deve-se considerar que o número de mensagens trocadas pelos agentes durante o processo de aprendizagem é exponencial. Tal discussão não está no escopo desse trabalho.

4 Conclusões e Discussões

Este artigo concebeu um método híbrido para melhorar a coordenação em sistemas multiagente a partir da combinação de modelos de interação apresentados em um trabalho anterior. Em problemas de decisão seqüencial um agente interage repetidamente no ambiente e tenta otimizar seu desempenho baseado nas recompensas recebidas. Assim, é difícil determinar as melhores ações em cada situação, pois uma decisão específica pode ter um efeito prolongado, em função da influência sobre ações futuras. O método híbrido proposto é capaz de

levar o agente a uma aceleração na convergência de sua política de ação, superando as particularidades encontradas nos modelos de interação apresentados em [18]. A política desses modelos pode apresentar características diferentes em ambientes com configurações diferentes, fazendo com que o agente deixe de convergir em determinados ciclos da aprendizagem. Diante disso, o método híbrido de cooperação utiliza as melhores características dos modelos de interação (sumarizados na Subseção 2.3.1). As políticas do método híbrido superam a política de cada modelo, auxiliando na troca das melhores recompensas para formar uma boa política de ação global. Isso é possível porque o método híbrido consegue estimar as melhores recompensas adquiridas ao longo da aprendizagem, descobrindo um arranjo com os melhores reforços encontrados nas políticas de ação parcial de cada modelo.

Acumular recompensas geradas por diferentes políticas de ação é uma alternativa para auxiliar um agente adaptativo na busca de boas políticas de ação. Os experimentos realizados com o método híbrido mostram que mesmo tendo um custo computacional superior, seus resultados são satisfatórios, pois a abordagem melhora em termos de convergência o algoritmo *Q*-learning padrão. Apesar dos resultados satisfatórios, experimentos adicionais são necessários para responder algumas questões em aberto. Por exemplo: o uso de algoritmos com paradigmas diferentes poderia ser utilizado para explorar os estados com regiões de congestionamento mais elevado? Observamos que nessas regiões as recompensas dos estados são menores. Um sistema de aprendizagem por reforço com algoritmos de diferentes paradigmas foi proposto por [19, 21], onde os resultados parecem ser encorajadores. Pretende-se também, utilizar diferentes métodos de aprendizagem para analisar situações como: i) explorar os estados mais distantes do objetivo, onde a exploração é menos intensa e; ii) empregar uma função heurística, onde *a priori* poderia acelerar a convergência dos algoritmos, onde uma política heurística indicaria a escolha da ação tomada e limitaria, assim, o espaço de busca do agente. Funções heurísticas foram usadas em [17]. É pretendido ainda avaliar o método em ambientes dinâmicos e com maiores variações de estados. Essas hipóteses e direcionamentos serão objetos de estudo em pesquisas futuras.

Agradecimentos

Nós agradecemos aos revisores anônimos por seus comentários úteis e aos colegas do Laboratório de Agentes de Software da Pontifícia Universidade Católica do Paraná e do Laboratório de Redes Convergentes da Universidade Tecnológica Federal do Paraná, Pato Branco - Pr. Esta pesquisa é apoiada pela Financiadora de Estudos e Projetos (FINEP), sob o número 3560/06, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico), processo 470325/2008-9 e Universidade Tecnológica Federal do Paraná.

Referências

1. Abdallah, S., Lesser, V. R.: Modeling Task Allocation Using a Decision Theoretic Model. In fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05), Netherlands, pp. 719-726, 2005.
2. Abdallah, S., Lesser, V. R: Organization-Based Cooperative Coalition Formation. 2004 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'04), Beijing, China, pp. 162-168, 2004.
3. Bryan, H., Mailler, R., Lesser, V. R: A Case Study of Organizational Effects in a Distributed Sensor Network. Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04), pp. 1294-1295, 2004.
4. Chapelle, J., Simonin, O., Ferber, J.: How Situated Agents can Learn to Cooperate by Monitoring their neighbors' Satisfaction. ECAI'02, Lyon, pp. 68-72, 2002.
5. Comanici, G., Precup, D.: Optimal Policy Switching Algorithms for Reinforcement Learning. Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS'10), Toronto, Canada, pp. 709-714, 2010.
6. DeLoach, S. A., Valenzuela, J. L.: An Agent-Environment Interaction Model. In L. Padgham and F. Zambonelli (Eds.): AOSE 2006, LNCS 4405. Springer-Verlag, Berlin Heidelberg, pp. 1-18, 2007.
7. Ferber, J.: Multi-Agent Systems - An Introduction to Distributed Artificial Intelligence. Addison-Wesley, Harlow, 1999.
8. Ghavamzadeh, M., Mahadevan, S.: A Multiagent Reinforcement Learning Algorithm by Dynamically Merging Markov Decision Processes". Proc. of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02), Bologna, Italy, pp. 845-846, 2002.
9. Grzes, M., Hoey, J.: Efficient Planning in R-max. Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS'11), Tumer, Yolum, Sonenberg and Stone (eds.), Taipei, Taiwan, 2011.
10. He, M., Rogers, A., Luo, X., Jennings, N. R. Designing a successful trading agent for supply chain management. Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, Hakodate, Japan, pp. 1159-1166, 2006.
11. Kaelbling, L. P., Littman, M. L., Moore, A. W.: Reinforcement learning: A survey. Journal of Artificial Intelligence Research, v. 4, pp. 237-285, 1996.
12. Knox, W. B, Stone, P.: Combining Manual Feedback with Subsequent MDP Reward Signals for Reinforcement Learning. Proc. of 9th Int. Conf. on Autonomous Agents and Mul-

tiagent Systems (AAMAS'10), vander Hoek, Kaminka, Lespérance, Luck and Sen (eds.), Toronto, Canada, pp. 5-12, 2010.

13. Littman, M. L.; Kaelbling, L.P.: Reinforcement Learning: A Survey, Journal of intelligence Research 4, pp. 237-285, 1996.
14. Mataric, M. J.: Using Communication to Reduce Locality in Distributed Multi-Agent Learning, Journal of Experimental and Theoretical Artificial Intelligence, special issue on Learning in DAI Systems, Gerhard Weiss, ed., 10(3), pp. 357-369, 1998.
15. Miyazak, K., Kobayashi, S.: Rationality of Reward Sharing in Multi-agent Reinforcement Learning. New Generation Computing, Ohmsha, Ltd. and Springer-Verlag, v. 19, no. 2, pp. 157-172, 2001.
16. Porta, J. M., Celaya, E.: Reinforcement Learning for Agents with Many Sensors and Actuators Acting in Categorizable Environments. Journal of Artificial Intelligence Research, v. 23, pp. 79-122, 2005.
17. Ribeiro R.; Koerich A. L., Enembreck F.: Noise Tolerance in Reinforcement Learning Algorithms, Int. Conf. on Inteligent Agent Technology, Silicon Valley, California, USA, 2007.
18. Ribeiro, R., Borges, A. P., Enembreck, F.: Interaction Models for Multiagent Reinforcement Learning. International Conference on Computational Intelligence for Modelling, Control and Automation - CIMCA08, Vienna, Austria, pp. 1-6, 2008.
19. Ribeiro, R., Enembreck, F., Koerich, A. L.: A Hybrid Learning Strategy for Discovery of Policies of Action. International Joint Conference X Ibero-American Artificial Intelligence Conference (IBERAMIA 2006) and XVIII Brazilian Artificial Intelligence Symposium (SBIA 2006), Ribeirão Preto, SP, Brazil. LNCS, Vol.4140, pp. 268-277, 2006.
20. Ribeiro, R.: Avaliação e Descoberta de Políticas de Ação para Agentes Autônomos Adaptativos. Dissertação de Mestrado, Programa de Pós-Graduação em Informática Aplicada (PPGIA), Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, 2006.
21. Ribeiro. R., Borges, A. P., Koerich, A., Scalabrin, E. E., Enembreck, F.: A Strategy for Converging Dynamic Action Policies. In: IEEE Symposium Series on Computational Intelligence, 2009, Nashville. Proceedings of IEEE Symposium Series on Computational Intelligence, v. 10, pp. 136-143, 2009.
22. Ribeiro. R.: Análise do Impacto da Teoria das Redes Sociais em Técnicas de Otimização e Aprendizagem Multiagente Baseadas em Recompensas. Tese de Doutorado. Programa de Pós-Graduação em Informática (PPGIa), Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, 2010.

23. Sichman, J. S.: Raciocínio Social e Organizacional em Sistemas Multiagentes: Avanços e Perspectivas. Tese (Escola Politécnica da Universidade de São Paulo, para obtenção do título de Professor Livre Docente) - USP, São Paulo, 2003.

24. Stone, P., Veloso, M.: Towards Collaborative and Adversarial Learning: a Case Study in Robotic Soccer. International Journal of Human-Computer Studies, v. 48, issue 1. Evolution and learning in multiagente systems. Academic Press, Inc. Duluth, MN, USA, p. 83-104, 1996.

25. Sutton, R. S., Barto, A. G.: Reinforcement Learning: An Introduction. A Bradford book, The MIT Press, London, England, 1998.

26. Tesauro, G.: Temporal Difference Learning and TD-Gammon Communications of the ACM, vol. 38, no. 3, p. 58-68, 1995.

27. Vidal, J. M.: The effects of cooperation on multiagent search in task-oriented domains. Journal of Experimental and Theoretical Artificial Intelligence, 16(1): pp. 5-18, 2004.

28. Walsh, T. J., Goschin, S., Littman, M.L: Integrating Sample-Based Planning and Model-Based Reinforcement Learning. In Proceedings AAAI 2010. Twenty-Fourth AAAI Conferences on Artificial Intelligence, Georgia, USA, July, 2010.

29. Watkins, C. J. C. H., Dayan, P.: Q-Learning, M. Learning, vol. 8(3), pp.279-292, 1992.

30. Weiss, G., Sen, S.: Adaptation and Learning in Multiagent Systems, Lecture Notes in Artificial Intelligence. Berlin, Germany: Springer-Verlag, v. 1042, p. 1-21, 1996.

31. Wooldridge, M. J.: An Introduction to MultiAgent Systems. John Wiley ands Sons, 2002.

32. Zhang, C., Lesser, V. R.: Multi-Agent Learning with Policy Prediction. Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 927-934. 2010.