USO DE APRENDIZAGEM POR REFORÇO E CONTROLE EM NÍVEL META PARA MELHORAR A PERFORMANCE DA COMUNICAÇÃO EM GERÊNCIA DE TRÁFEGO AÉREO

Daniela Pereira Alves
Ministério da Ciência e Tecnologia – MCT
Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT

Li Weigang
Bueno Borges de Souza
Universidade de Brasília - UnB
Departamento de Ciência da Computação - CIC

RESUMO
Este artigo descreve um modelo de computação inteligente para auxiliar o processo de tomada de decisão de um sistema distribuído, no caso específico, *Air Traffic Flow Management* - ATFM. Ele é indicado para atuar na comunicação via troca de mensagens entre aeroportos, trabalhando como uma camada adicional que usa os meta-dados em suas decisões. O modelo é considerado inovador porque usa aprendizagem por reforço adequada às características deste ambiente estocástico, preocupando-se com a velocidade e qualidade do processo de tomada de decisão. Na modelagem, três melhorias foram propostas para a própria aprendizagem: heurística inicial, epsilon adaptativo e heurística baseada em performance. Elas são combinadas aos algoritmos de aprendizado por reforço: Q-learning e SARSA. Dois estudos de caso avaliam o desempenho e a qualidade do aprendizado quanto às três melhorias propostas.

ABSTRACT
This article describes a model of intelligent computation to assist the make decision process in a distributed system, in specific case, Air Traffic Flow Management - ATFM. It is indicated to act in the communication saw exchanging of messages among airports and working as an additional layer that uses the goal-given ones in its decisions. The model is considered innovative because it uses reinforcement learning adjusted to the characteristics of this random environment cosidering the speed and quality of decision make process. In the modeling we proposed three learning: initial heuristic, epsilon adaptive and heuristic based in performance. They are combined to the reinforcement learning´s algorithms: Q-learning and SARSA. Two studies of case evaluate the performance and the quality of the learning according of improvement proposals.

1. INTRODUÇÃO

O fluxo em tráfego aéreo tende a crescer nos próximos anos e, como conseqüência, a demanda sob os recursos dos aeroportos aumentará de forma considerável. Por outro lado, a capacidade para atendê-la não irá acompanhar este crescimento. Esta discrepância ocorrerá porque investir em infra-estrutura é caro e nem sempre é possível fazê-lo. Existem aeroportos em que não há como construir uma nova pista porque não existe espaço para isso e, neste caso, teria que construir um novo aeroporto com maior capacidade. Assim, uma solução para este problema é investir em soluções que visem o uso de forma mais eficiente dos recursos atuais dos aeroportos.

Uma proposta de solução para o problema de tráfego aéreo é um modelo que atua em *Air Traffic Flow Management* - ATFM com informações distribuídas que são compartilhadas via troca de mensagens, conforme pode ser visto em (Weigang et. al, 1997, Bonzano, 1996, Prevôt, 2002, Nguyen-Duc, 2003 e Dib, 2004). Estes sistemas empregam os conceitos de sistemas distribuídos e se caracterizam pela intensa troca de mensagens. Estas mensagens ocorrem com a finalidade de coordenar, negociar e trocar informações entre os agentes do sistema.

O sistema de informações distribuídas que se comunica via troca de mensagens proposto para ATFM por Weigang et. al, (1997) e Dib (2004) permitiu melhor eficiência no tratamento do problema de tráfego aéreo atuando sobre o escalonamento dos vôos. Este sistema distribuiu as informações disponíveis entre os aeroportos e se mostrou vantajoso por vários pontos. Primeiro, deixa de existir um ponto único que, na verdade, torna-se o gargalo. Segundo as consultas sobre as informações podem ocorrer de forma mais rápida. Terceiro, não são todas as informações que precisam ser mantidas em todos os pontos, mas somente aquelas que são importantes e que exercem influência no escalonamento local. Contudo, a distribuição da informação pode gerar uma grande quantidade de mensagens que se não forem bem tratadas leva a situações indesejáveis como a paralisação da

comunicação. Desta forma, surge a necessidade de um mecanismo para auxiliar o processo de tomada de decisão.

Baseado nas pesquisas mencionadas em cima, este trabalho propõe a combinação de controle em nível meta e aprendizagem por reforço com intuito de possibilitar melhor eficiência no processo de tomada de decisão de um sistema distribuído ATFM. Nesta proposta, as mensagens passam a receber uma análise especial via uma camada de gerência em nível meta, baseada em Lesser e Raja (2003). Ela atua sob a camada de controle do sistema. O funcionamento da camada de meta-gerência faz uso de aprendizagem por reforço, (Sutton e Barto, 1998), quando precisa decidir qual ação tomar em determinado momento.

O objetivo do modelo aqui proposto é auxiliar a camada de controle do sistema distribuído ATFM quanto à tomada de decisão. Como conseqüência, o modelo contribui na melhoria de performance no processo da comunicação entre os agentes dos sistemas de controle dos aeroportos que se comunicam via troca de mensagens.

2. METODOLOGIA

2.1. Sistemas Multi-agentes

Um agente inteligente é uma entidade computacional autônoma que percebe o ambiente em que se encontra através de sensores e age sobre esse ambiente para produzir efeitos, (Russel e Norvig, 2004).

Segundo Weiss (2000), os agentes são inteligentes quando operam de forma flexível e racional em uma variedade de circunstâncias ambientais, dadas as informações que possuem. Eles têm a capacidade efetiva de realizar uma tarefa e a capacidade perceptiva adequada à obtenção dos dados. O comportamento flexível e racional tem como objetivo a resolução de problemas, o planejamento, a tomada de decisão e o aprendizado.

Um sistema multi-agentes é um sistema computacional em que dois ou mais agentes interagem entre si de forma a realizarem determinadas tarefas ou satisfazerem um conjunto de objetivos, (Weiss, 2000). Um grupo de agentes, num sistema multi-agente, procura coordenar o comportamento inteligente. O raciocínio dos agentes tem como resultado as ações que devem ser tomadas e o processo de coordenação em si. A organização do sistema está sempre sujeita às mudanças visando à adaptação do sistema às variações no ambiente e do problema a ser resolvido.

2.2. Aprendizagem por Reforço (AR)

A aprendizagem por reforço é interessante para os ambientes estocásticos. Estes ambientes sofrem alterações e as ações que são mais indicadas no instante atual podem ser inadequadas em instantes posteriores devido aos fatores randômicos que ocorrem. Não existe a presença de um professor que avalia a atuação do agente, o aprendizado, neste caso, é decorrência da experiência adquirida através de sua atuação com ambiente, (Russel e Norvig, 2004).

O agente interage com o ambiente diretamente para obter informações, e estas são processadas através de um algoritmo apropriado, a fim de executar a ação que maximize a satisfação dos seus objetivos. A aprendizagem por reforço apresenta quatro características principais, (Sutton e Barto, 1998): Primeiro, o *aprendizado pela interação*, o agente de aprendizado por reforço atua no ambiente e aguarda pelo valor de reforço que é retornado como resposta perante a ação tomada. Segundo, o agente vive o dilema entre *usufruir* o conhecimento que se adquiriu com a experiência ou *explorar* correndo riscos no espaço de estados com novos valores. Para a decisão, um valor q é sorteado e comparado com o valor epsilon (ε), se é menor então, o agente deverá explorar e, caso contrário, deverá usufruir o conhecimento adquirido. Terceiro, o *retorno atrasado*, um valor máximo de reforço é enviado para o agente pelo ambiente. As ações tomadas devem levar a maximizar o retorno total, isto é, a qualidade das ações tomadas é vista pelas soluções encontradas em longo prazo. E quarto, a *orientação ao objetivo*, o agente atua no ambiente desconhecido tentando alcançar um objetivo. O objetivo é geralmente otimizar algum comportamento dentro do ambiente. Nas subseções seguintes são apresentados dois algoritmos de aprendizagem por reforço: Q-learning e SARSA.

2.2.1 Algoritmo Q-learning (QL) (Watkins, 1989)

O agente de aprendizagem que emprega o algoritmo QL busca maximizar a recompensa obtida pelo agente. Por isso, como conseqüência contribui para aumentar o custo computacional em raciocinar sobre as possibilidades

que se tem em determinado estado. Assim, o tempo de execução deste algoritmo reflete a procura pela maximização da recompensa.

2.2.2 Algoritmo SARSA (Sutton e Barto, 1998)
O agente de aprendizagem que emprega o algoritmo SARSA sorteia uma ação a ser executada baseando-se na distribuição das ações tomadas pelo agente. Desta forma, existe uma chance maior de que a ação mais tomada até então, seja tomada novamente. A origem do nome SARSA advém do fato de que a regra de aprendizado utiliza todos os elementos da quíntupla $<s_t, a_t, r, s_{t+1}, a_{t+1}>$ que define a transição de um par estado-ação para o próximo na atualização da estimativa.

2.2.3 Aprendizagem por Reforço Acelerada por Heurística
A aprendizagem por reforço acelerada por heurística combina as vantagens de AR com o desempenho alcançado com uso de heurísticas. AR aprende através da experiência o que é bom em ambientes estocásticos. Nesses ambientes, a melhor ação depende de qual estado está sendo observado em determinado momento e não se consegue prever, a priori, quais são as melhores ações a serem tomadas. Fatores randômicos interferem nos resultados alcançados com o tempo. As heurísticas auxiliam a tornar o processo de aprendizagem algo mais rápido. O trabalho tomado como referência, (Bianchi, 2005), combina as duas estratégias: AR e heurística com o intuito de obter melhores resultados.

2.3 Controle em Nível Meta
O controle em nível meta (m*eta-level control*) é uma camada adicional inserida em sistemas de informações com o objetivo de auxiliar a camada de controle já existente. Ele procura compensar, ou justificar sua existência apontando ao controle caminhos alternativos que melhorem a eficiência do sistema como um todo, garantindo que tarefas de maior utilidade e com prazos de conclusão menores sejam priorizadas, além de auxiliar o uso mais adequado dos recursos computacionais limitados.

O modelo de controle em nível meta proposto por Lesser e Raja (2003), abrange a construção de um sistema multi-agentes onde a camada de controle em nível meta monitora, através de uma série de parâmetros, analisando desde as ações que se referem à comunicação entre os agentes até as ações relativas ao conjunto de tarefas escalonadas pelo sistema para execução. Ele utiliza meta-parâmetros que são adicionados às mensagens trocadas entre os agentes e o controle em nível meta, através de uma análise dentro do processo decisório de Markov, (Clarke e Disney, 1970), aprendizagem por reforço, (Sutton e Barto, 1998) e um conjunto de heurísticas aplicadas sobre tal meta-parâmetros.

3. MODELAGEM DO PROBLEMA

3.1 Gerenciamento de Fluxo de Tráfego Aéreo (*ATFM*)
ATFM é uma área de pesquisa cujo objetivo principal é apresentar propostas com intuito de garantir que os vôos sejam efetuados de forma rápida, segura, ordenada e econômica (Targa, 2001).

O excesso de mensagens no ambiente ATFM ocorre por causa da necessidade de negociação entre os aeroportos para que seja estabelecido o escalonamento dos vôos. Vários aspectos interferem na quantidade de mensagens trocadas na negociação. Inicialmente, o mau tempo que interfere no escalonamento dos vôos previamente planejados. Depois, o interesse das companhias aéreas em disponibilizarem vôos nos mesmos horários por motivo de demanda garantida e o exemplo disso são os vôos às 18 horas. Em seguida, efeitos sanzionais como férias e feriados. E por último, eventos menos freqüentes como alerta terrorista que apesar de não acontecerem no Brasil influenciam os vôos internacionais que partem de aeroportos brasileiros.

Sob as condições anteriormente mencionadas, verifica-se que somente o alto desempenho da tecnologia de redes e dos processadores modernos não são suficientes para tratarem a quantidade de mensagens porque a operação ocorre sob condições extremas. Um outro fato que reforça esta afirmação é que as redes não são projetadas para operarem em sua demanda máxima por causa do alto custo do projeto. Embora os recursos das redes estejam bastante evoluídos em relação há vários anos atrás, por causa da complexidade dos problemas atuais, eles se mostram limitados. E assim, boas soluções para o problema envolvem tanto *hardware* como *software*.

Um outro fator que influencia o escalonamento em tráfego aéreo é a presença de incerteza no processo. Tráfego Aéreo é um ambiente estocástico e neste caso, os fatores mencionados no parágrafo anterior: mau tempo,

congestionamento, demanda de recursos maior que disponibilidade ocorrem de forma, nem sempre, previsível. Desta forma, soluções que trabalham com processos estocásticos definem melhor esse ambiente.

3.2 Modelo Meta Gerente de Mensagens (MGM)

O MGM foi desenvolvido neste trabalho como uma camada de controle em nível meta cujo objetivo é gerenciar a troca de mensagens. Ele é composto pelo Módulo de Decisão e Controle – MODEC e pelo Módulo de Aprendizagem por Reforço – MAP. Tanto o MODEC quanto MAP seguem o formalismo matemático conhecido como Processo Decisório Markoviano (Clarke e Disney, 1970).

O MGM conta com três estruturas de dados principais (Entrada Lista, Agenda e Pronto Lista) que auxiliam no processo de gerência e tomada de decisão. A *Entrada Lista* guarda as mensagens que chegam da Rede. A Agenda guarda aquelas armazenadas para processamento futuro e na Pronto Lista estão as mensagens encaminhadas para o Sistema Aeroporto.

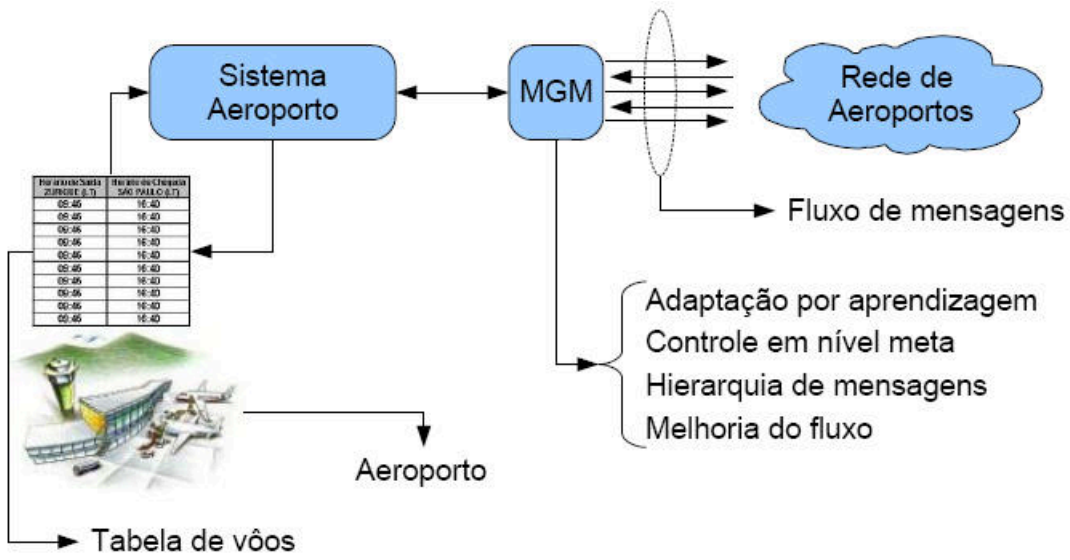


Figura 1: Modelo Meta Gerente de Mensagens aplicado a ATFM.

O meta gerente de mensagens trabalha com um conjunto de sete parâmetros sendo que os dois principais são a utilidade da mensagem e o prazo de resposta. Os outros meta-parâmetros são obtidos internamente pela análise do ambiente e das respostas obtidas pelo sistema hospedeiro. As mensagens descrevem as características do ambiente e permitem refletir a realidade dos aeroportos em determinado momento.

Uma mensagem que chega ao controle em nível meta é colocada na fila de prioridade Entrada Lista. A hierarquia dentro desta fila obedece ao grau de utilidade e o prazo de resposta da mensagem. O MODEC (Módulo de Decisão e Controle) retira a primeira mensagem da fila e decide se a mensagem deve ser agendada, encaminhada diretamente ao sistema hospedeiro ou simplesmente descartada. O processo decisório envolve um conjunto de estados markovianos e a aprendizagem por reforço (Sutton e Barto, 1998). O módulo de aprendizagem garante a capacidade de adaptação em situações estocásticas, ou seja, adaptação às alterações do ambiente que exijam uma mudança no procedimento decisório atual.

Um ciclo completo no modelo meta gerente de mensagens pode ser resumido da seguinte forma. Uma mensagem chega ao MODEC e, a seguir, é gerada uma consulta ao Módulo de Aprendizagem para que este indique uma ação a ser tomada. Depois que uma ação é escolhida pelo MAP ela é informada ao MODEC que verifica a situação do ambiente a partir da ação e retorna um reforço atrasado ao MAP para que este adquira experiência ao longo do tempo.

3.3 Módulo de Aprendizagem

O módulo de aprendizagem conta com um algoritmo de aprendizagem por reforço que pode ser SARSA ou QL. E além do algoritmo, algumas melhorias foram propostas com intuito de se acelerar o aprendizado: heurística inicial, epsilon adaptativo e heurística baseada em performance. Elas são explicadas nos parágrafos seguintes.

A primeira melhoria proposta, Heurística Inicial (HI), combina as sugestões presentes em Lesser e Raja (2003). Para tanto, uma estrutura de dados armazena uma ação mais indicada a ser tomada pelo MODEC de acordo com o estado atual do ambiente que ele está atuando. As regras seguem a lógica de priorizar as mensagens com utilidade alta e prazo curto, em seguida, utilidade alta e prazo médio e assim, sucessivamente, são atendidas. Depois, são tratadas as mensagens com utilidade baixa e prazo longo que podem, às vezes, ser descartadas sem prejuízos ao sistema porque uma análise criteriosa foi realizada anteriormente.

A segunda melhoria proposta, Epsilon Adaptativo (EA), procura adequar o *trade-off* existente em algoritmos de aprendizagem por reforço que é a escolha entre usufruir o conhecimento que se tem até o momento ou explorar novas ações com o intuito de se atingir melhores resultados. A idéia é que se bons resultados estão sendo atingidos deverá ser aumentada à distribuição de probabilidades para usufruto, por outro lado, se resultados ruins estão sendo atingidos deverá ser aumentada a distribuição de probabilidades para exploração.

A terceira melhoria proposta é a inserção de uma Heurística baseada em Performance (HP) do sistema. Nela, o resultado da equação 1, a seguir, representa a avaliação do desempenho atual do sistema:

$$AvD = \frac{(3*SDS + 2*RF + SEL + SAL + SPL)}{8} \quad (1)$$

Onde SDS, significa a situação do sistema hospedeiro; RF, razão de fluxo e representa a diferença entre a quantidade de mensagens encaminhadas para o sistema hospedeiro em relação às mensagens que chegaram; SEL, situação da entrada lista; SAL, situação da agenda e SPL, situação da lista de pronto.

Um valor de desempenho ideal (100%) é tomado como referência e o valor de desempenho atual é avaliado e comparado ao anterior. A nota, obtida com a diferença, interfere no usufruto do agente de aprendizado por reforço. O objetivo desta heurística é minimizar a distância entre o desempenho atual avaliado em determinado instante de análise em relação ao desempenho ideal tomado como referência. Em todas as melhorias propostas (EA e HP), HI é aplicada. Isto ocorre porque nos momentos iniciais de cada agente ela se mostra útil.

4. ESTUDO DE CASO

O estudo de caso apresentado neste artigo avalia a meta gerência de mensagens segundo o interesse de ATFM que é a aprendizagem em um ambiente estocástico, o processo de tomada de decisões rápidas e a qualidade da decisão tomada.

O problema em análise, ATFM, lida com dois aspectos que se contrapõem. De um lado, tem a necessidade de se tomar decisões rápidas e, de outro, a qualidade das decisões que são tomadas. Acontece que para tomar boas decisões é exigido um tempo de raciocínio em cima das opções disponíveis e, nem sempre, existe esse tempo necessário em ATFM.

4.1 Simulação

Cada fonte geradora de mensagem pode ser abstraída como sendo um aeroporto no modelo criado. Desta forma, várias informações podem ser abstraídas. Um aeroporto que em determinado momento esteja com um tráfego leve necessitará trocar poucas mensagens, uma vez que, menos negociações serão necessárias.

Na simulação, três dessas fontes geradoras de mensagens existem e representam três aeroportos enviando mensagens ao quarto aeroporto em análise. Em geral, não existem diferenças entre a lógica de cada fonte geradora. A intenção é que elas representem as adversidades que ocorrem nos aeroportos. As saídas produzidas por elas representam características importantes de cada aeroporto que precisam ser tratadas.

As adversidades enfrentadas pelos aeroportos são simuladas através do intervalo de geração entre as mensagens. O contexto em análise, ATFM, e a escala de horário dos vôos são considerados em relação ao intervalo de 2 horas a 48 horas antes do pouso ou decolagem. Assim, para representar um processo de negociação intensa, as mensagens são geradas com espaçamento curto entre uma e outra mensagem. De outra forma, um processo de negociação leve é representada com espaçamento longo entre as mensagens.

Além do prazo das mensagens, um outro parâmetro importante no processo de tomada de decisão do meta gerente de mensagens é a utilidade das mensagens. Segundo o modelo tomado como referência, (Lesser e Raja, 2003): se uma mensagem com alta utilidade chega no tempo presente, então existe uma alta probabilidade de que

a próxima mensagem a chegar seja com alta utilidade. No modelo MGM este valor foi considerado como 60 % de chance.

No primeiro caso, avaliações do desempenho, três simulações foram realizadas para cada uma das três propostas. As variações ocorreram quanto à quantidade de períodos considerados: 16, 32 e 64. Como os gráficos convergiram por volta do período 16, foi selecionada a primeira simulação porque bem representou a situação desejada.

Para avaliar a qualidade da aprendizagem é estabelecida uma nota ao agente que define em determinado instante o quão boa está sua aprendizagem em relação ao padrão tido como referência. Uma nota é gerada a partir da ação tomada em relação ao que é sugerido pela heurística inicial e também considerando o grau de congestionamento do meio externo. Neste segundo caso, avaliações da qualidade do aprendizado, tanto a quantidade de períodos quanto o intervalo de geração entre as mensagens foram variados através de quatro simulações. Optou-se por 32 períodos e intervalos entre as mensagens: de 1 segundo para QL e 5 segundos para o SARSA.

4.2 Avaliação do Desempenho no Meta Gerente de Mensagens

Cada algoritmo foi testado com as combinações de melhorias propostas, vide seção 3.3 deste artigo. Em resumo, as melhorias incluem a utilização de parâmetros adaptativos e heurísticas baseadas em desempenho que proporcionam maior flexibilidade e rapidez no aprendizado do agente.

A figura 2 mostra o desempenho do meta gerente de mensagens obtido com a implementação tanto do algoritmo QL quanto do SARSA considerando as melhorias propostas. As curvas de desempenho do QL são piores, nas três avaliações, se comparada às curvas do algoritmo SARSA. A literatura justifica os resultados alcançados pela natureza dos dois algoritmos.

Quanto às melhorias propostas tem-se que a abordagem que prevê somente o uso de Heurística Inicial e Epsilon Adaptativo apresentaram as melhores curvas para o algoritmo SARSA. Já para o QL, a Heurística Inicial e Heurística baseada em Performance tiveram as melhores curvas em termos de desempenho, mas inferiores aos resultados obtidos com o SARSA.

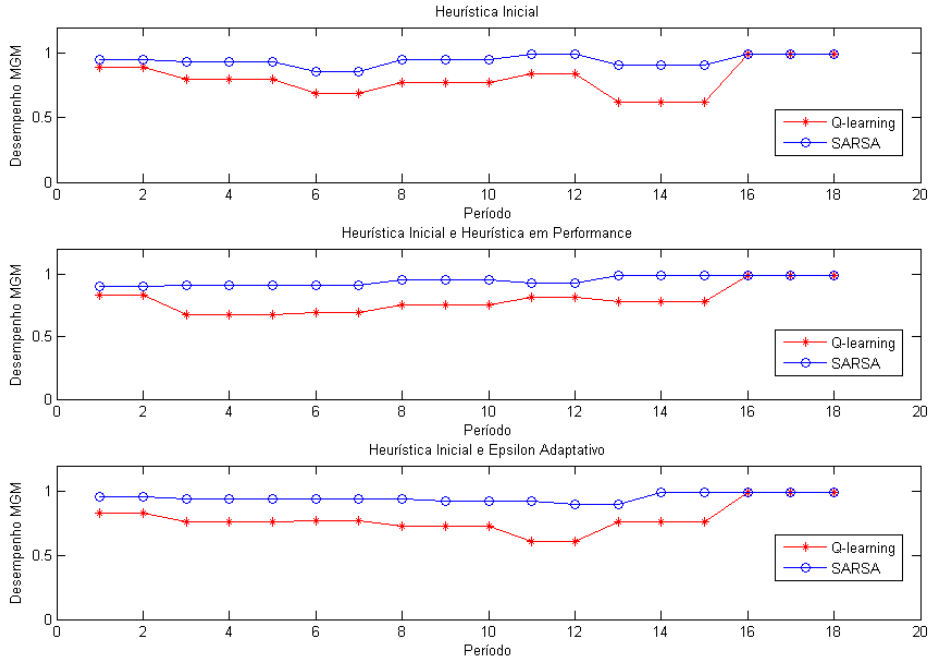


Figura 2: Avaliação do desempenho MGM – Comparação entre o algoritmo SARSA e o QL.

4.3 Avaliação da Qualidade da Aprendizagem dos Agentes

Neste segundo estudo de caso, foi definido um vetor de uplas (estado, ação, recompensa) que considera o valor aceitável dependendo das condições em que o ambiente se apresente. Uma vez que a decisão é tomada ela é comparada com o padrão aceitável. Estando essa decisão dentro do padrão aceitável um parâmetro nota é incrementado e armazenado no vetor de notas. Após o preenchimento completo do vetor de notas que inicialmente tem cem posições retira-se a média aritmética das notas e altera-se o valor da nota. Isso garante que boas decisões conduzirão ao crescimento da nota e más decisões ao decrescimento da nota.

Na figura 3 observa-se que as decisões tendem a seguir o padrão esperado sendo que nos pontos iniciais do gráfico algumas decisões ruins são tomadas. Inicialmente é atribuída nota zero e à medida que o módulo toma boas decisões essa nota é aumentada. Deve-se esperar por algumas decisões ruins em virtude da aprendizagem ser por reforço. O algoritmo SARSA tem tempo de excução mais rápido do que o QL, (Figura 2), no entanto, os experimentos realizados mostram que ele aprende mais lentamente visto que precisa de mais tempo para atingir uma nota próxima de cem. Em contra-partida, o algoritmo QL leva uma quantidade de períodos menor para atingir uma nota próxima de cem (Figura 3).

Acerca dos resultados obtidos com a Figura 3 verifica-se que a melhoria Epsilon Adaptativo combinada a HI com uso de QL obtém a melhor curva de aprendizagem até aproximar a convergência que é por volta do período 16. E daí em diante, a melhor abordagem para o SARSA é a Heurística em Performance.

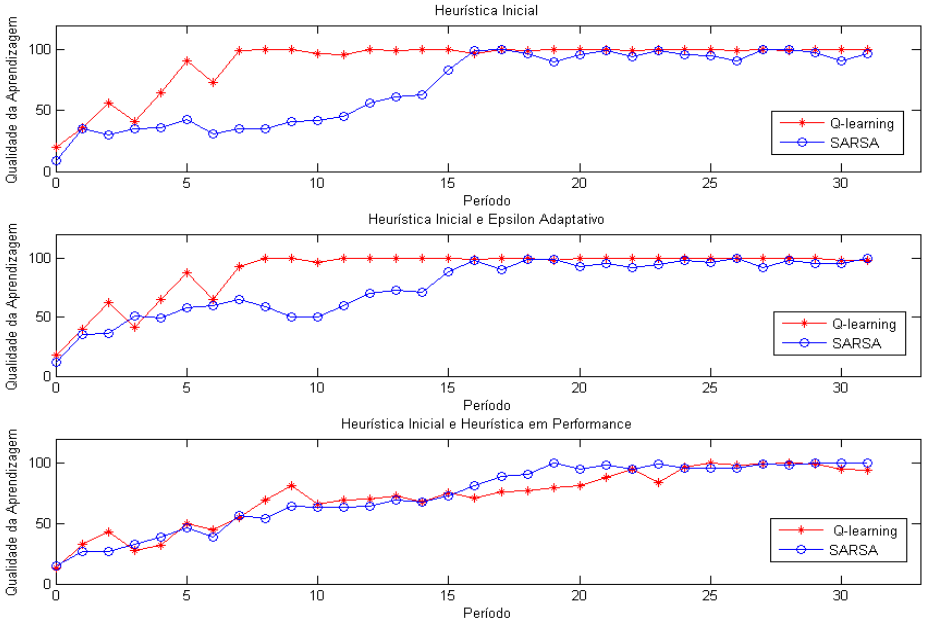


Figura 3: Avaliação da qualidade da aprendizagem – Comparação entre o algoritmo SARSA e o QL.

5. CONCLUSÕES E TRABALHOS FUTUROS

O modelo proposto se mostrou interessante de ser combinado à proposta de distribuição de informações entre os aeroportos por vários aspectos. Inicialmente, porque insere um processo de avaliação com intuito de escolher uma melhor ação a ser tomada sob certos critérios (congestionamento de mensagens, grande tempo para avaliação, entre outros). Em seguida, porque ele insere aprendizagem no processo permitindo adaptação ao longo do tempo. E por último, na aprendizagem é observada a característica estocástica do ambiente. Não são consideradas somente regras estáticas que *a priori* foram previstas, mas uma combinação delas através da heurística inicial com as outras propostas: EA e HP.

Os estudos de caso aqui apresentados mostram que a proposta é interessante para a situação freqüente enfrentada pelos aeroportos do País que é o congestionamento das pistas em determinados horários. E assim sugere a combinação dos dois algoritmos de forma que inicialmente se atinja um nível de aprendizado satisfatório rapidamente, pelo uso do QL com melhorias, e em seguida, passando-se a usar o SARSA sobre a base de conhecimento montada com o QL. O SARSA garantirá um desempenho melhor no processo decisório. Contudo,

vale ressaltar que esta proposta não funcionará bem em todos os casos porque o ambiente é estocástico e, portanto, muda ao longo de tempo.

Como trabalhos futuros, fatores externos podem ser associados ao meta gerente de mensagens no processo de tomada de decisão, como por exemplo: a rede e os sistemas hospedeiros. E também a análise de outros algoritmos de aprendizagem por reforço, como por exemplo: o algoritmo Dyna, (Sutton e Barto, 1998), pode ser interessante.

REFERÊNCIAS BIBLIOGRÁFICAS

Bianchi, R. A. C. (2005). *Uso de Heurísticas para a Aceleração do Aprendizado por Reforço*. XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo, RS.

Bonzano, A.; Cunningham, P. & Meckiff, C. (1996) . *A CBR System for Decision Support in Air Traffic Control*; August (In I. Smyth & B. Faltings (Eds.), Advances in Case-Based Reasoning. Amsterdam: Springer Verlag).

Clarke, A. B. & Disney, R. L. (1970). *Probability and Random Process for Engineers and Scientists*. New York: Wiley.

Dib, M. V. P. (2004). *Sistema Multi-agentes para Sincronização e Gerenciamento de Fluxo de Tráfego Aéreo em Tempo Real*; Dissertação de Mestrado, Universidade de Brasília.

Lesser, V. & RAJA, A. (2003). *Automated meta-level control reasoning in complex agents*, in the Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003), Mexico, August.

Nguyen-Duc, M.; Briot, J. P.; Drogoul, A. & Duog, V. (2003). *An application of Multi-Agent Coordination Techniques in Air Traffic Management*, in the Proceedings of the IEEE/WIC International Conference on Intelligent Agent Technology (IAT'03).

Prevôt, T. (2002). *Exploring the Many Perspectives of Distributed Air Traffic Management: The Multi Aircraft Control System MACS.* In S. Chatty, J. Hansman, and G. Boy (Eds.), Proceedings of the International Conference on Human-Computer Interaction in Aeronautics (HCI-Aero 2002), AAAI Press, Menlo Park, CA, pp. 149-154.

Russel, S. & Norvig, P. (2004). *Inteligência Artificial*. Editora Campus.

Sutton, R. & Barto, A. (1998). *Reinforcement Learning – An Introduction*. A Bradford Book, Cambridge, Massachussetts.

Targa, D. (2001). *Uma ferramenta automatizada no auxílio à alocação de slots para o problema de gerenciamento de fluxo de tráfego aéreo brasileiro*. Instituto Tecnológico de Aeronáutica.

Tidhar, G. (1992). *Distributed Air Traffic Flow Management*, In A. Adams and L. Sterling, editors, Proceedings of the 5th Australian Joint Conference on Artificial Intelligence. Australian Artificial Intelligence Institute, World Scientific, November.

Watkins, C. (1989). *Learning from delayed rewards*, PhD thesis, University of Cambridge. Psychology Department.

Weigang, L.; Alves, C. J. P. & Omar, N.; (1997). *An expert system for Air Traffic Flow Managemen*t. Journal of Advanced Transportation", Vol. 31, No. 3, 1997, pp. 343-361.

Weiss, G. (2000). Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. The MIT Press. England.