Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mecanismo de Negociação por Aprendizado por Reforço no Sequenciamento de Decolagem de Aeroportos

Diego Santos Kieckbusch

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Li Weigang

Brasília
2019

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mecanismo de Negociação por Aprendizado por Reforço no Sequenciamento de Decolagem de Aeroportos

Diego Santos Kieckbusch

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Li Weigang (Orientador)
CIC/UnB

Prof. Dr. Vinicius Ruela Pereira Borges
CIC/UnB

Prof. Dr. Geraldo P. R. Filho
CIC/UnB

Prof. Dr. Edison Ishikawa
Coordenadora do Bacharelado em Ciência da Computação

Brasília, 6 de março de 2019

Dedicatória

Dedico esta obra a meus país, Andrea e Rafael, por durante toda minha vida terem me mostrado o valor da educação, do esforço e da perseverança através do exemplo. Dedico também a minha namorada, Katarine, cujas palavras de incentivo e companheirismo ajudaram ao longo da jornada.

Agradecimentos

Agradeço primeiramente ao meu Orientado, Prof. Li Weigang pela oportunidade e pelos projetos desenvolvidos.

Agradeço ao Prof. Geraldo pelos conselhos e revisões.

Agradeço aos demais professores do curso de Ciência da Computação, cujos ensinamentos me permitiram concluir este trabalho.

Aos colegas dos Translab, Lucas, Helena, Vitor, Igor e Iuri pelas inúmeras discussões que permitiram um melhor entendimento da área.

Aos amigos, pela compreensão das ausências e afastamento temporário.

Resumo

Aprendizado por Reforço busca ensinar um agente como se comportar a partir da experimentação do ambiente e aquisição de recompensas envolvidas no processo. Para isso, o agente mapeia os estados encontrados no ambiente, as ações possíveis em cada um destes estados e as recompensas da execução destas ações nestes estados. Em um ambiente de negociação, as configurações possíveis entre as partes podem ser descritas como estados do ambiente, com as ofertas realizadas entre os oponentes constituindo ações e os acordos encontrados gerando recompensas às partes. Este trabalho busca ensinar uma aeronave em espera pela utilização da pista de decolagem a negociar sua posição com as demais. O trabalho apresenta melhorias em relação aos modelos existentes para o problema de sequenciamento de decolagens e apresenta resultados positivos nos estudos de caso realizados.

Palavras-chave: Aprendizado pro Reforço, Alocação de Slots, Gerenciamento de Partida de Aeroportos

Abstract

Reinforcement Learning aims to teach an agent how to behave using only the rewards received by exploring an environment. To achieve this, the enviroment is modeled in states, the available actions in each of these states and the associated rewards gained by executing these action. In a negotiation environment, the configurations between the stakeholders may be described as states of the enviroment, with offers done by each participant beign described as actions and the outcome of the negotiation, as rewards. This paper presents a model to teach aircrafts to negotiate their position in the queue for airport runway use. This papes presents improvements over the existing methods for aircraft departure scheduling and postive results in the case studies carried out.

Keywords: Reinforcement Learning, Slot Alocation, Departure Scheduling

Sumário

1 Introdução **1**
1.1 Motivação 1
1.2 Objetivos 2
1.3 Organização do Trabalho 3

2 Fundamentação Teórica **4**
2.1 Teoria dos jogos 4
2.1.1 Conceitos Gerais 4
2.1.2 Jogos na forma normal 5
2.1.3 Jogos na forma Extensiva 5
2.1.4 Equilíbrio em Jogos 7
2.1.5 Jogos Evolucionários 8
2.2 Aprendizado por Reforço 8
2.2.1 Conceitos Gerais 8
2.2.2 Processo de Decisão de Markov 9
2.2.3 Q-Learning 10

3 Trabalhos Relacionados **13**
3.1 Alocação de slots 13
3.2 Justificativa metodológica 14

4 Solução Proposta **16**
4.1 Descrição do Problema 16
4.2 Solução Proposta 18
4.2.1 Visão geral 18
4.3 Modelo RELEASE 19
4.3.1 Modelagem do ambiente 19
4.3.2 Modelagem das aeronaves 21
4.3.3 Modelagem do Processo de Negociação como MDP 27

5 Resultados **30**
5.1 Implementação 30
5.1.1 Base de Dados 31
5.2 Cenário de experimento 32
5.3 Casos de Experimento 34
5.3.1 Alocação Homogênea 34
5.3.2 Alocação Heterogênea 36

6 Conclusão **44**
6.1 Considerações Finais 44
6.2 Trabalhos Futuros 45

Referências **46**

Lista de Figuras

2.1 Exemplo de Jogo na forma Extensiva, reproduzido de Hart [1]. 7
2.2 Diferentes Nomenclaturas para conceitos análogos em distintas áreas de conhecimento. 8

4.1 Modelo Geral da Negociação de Slots de Decolagem 18
4.2 Árvore de Negociação entre duas Aeronaves, reproduzido de Ribeiro [2]. . . 23
4.3 Fluxograma de concepção da melhor oferta. 24
4.4 Fluxograma de avaliação da oferta. 26
4.5 Fluxograma de concepção da contra-oferta. 27

5.1 Classes da Solução. 31
5.2 Plano de Voo Regular. 32
5.3 Fluxograma de execução do experimento. 33
5.4 Taxa de exploração por Época. 37
5.5 Atraso dos Agentes. 39
5.6 Custo dos Agentes. 39
5.7 Comparação dos resultados por agente. 40
5.8 Recompensa por Ação nas Propostas aos Antecessores. 40
5.9 Recompensa por Ação na resposta ao Sucessores. 41
5.10 Custo das Aeronaves Adjacentes aos Agentes. 41
5.11 Velocidade das Aeronaves Adjacentes. 42
5.12 Tempo até o Destino das Aeronaves Adjacentes. 43

Lista de Tabelas

3.1 Diferença entre modelos. 15

4.1 Separação Miníma por combinação de aeronaves 18
4.2 Parâmetros do cenário 20
4.3 Parâmetros da aeronave 21
4.4 Processo de Decisão de Markov. 27

5.1 Informações presentes no plano de voo repetitivos. 31
5.2 Agentes Homogêneos Treinados. 34
5.3 Modelo de Referência Caso 1. 35
5.4 Aeronave TAM3769 após 25 épocas. 35
5.5 Aeronave TAM3608 após 25 épocas. 35
5.6 Aeronave TAM3608 após 100 épocas. 36
5.7 Aeronave GLO1720 após 25 épocas. 36
5.8 Aeronave GLO1720 após 100 épocas. 36
5.9 Atributos Aeronaves. 37
5.10 Agentes Heterogêneos Treinados. 38
5.11 Modelo de Referência Caso 2. 38

Capítulo 1

Introdução

1.1 Motivação

Atualmente, a área de tráfego aéreo enfrenta problemas de congestionamento. Nos principais aeroportos do mundo, a demanda por recursos aeroportuários ultrapassa a capacidade instalada. A expansão física destes recursos, além de onerosa, apresenta um horizonte de implementação muito distante[3]. A curto e médio prazo, soluções baseadas na melhor utilização dos recursos disponíveis apresentam melhores resultados. O principal gargalo no sistema é a pista de decolagem, desta forma, mesmo pequenas melhorias na performance de utilização da pista acarretam em ganhos consideráveis ao sistema. Busca-se garantir que as aeronaves decolem nos horários previstos e que os custos decorrentes de eventuais atrasos sejam minimizados[3].

As soluções atuais em alocação da pista apresentam limitações. A solução baseada em restrições busca alocar as aeronaves nos slots, intervalos de tempo correspondentes á permissão de utilização da pista, com base nas restrições operacionais, como separação de segurança entre aeronaves, sequenciamento de diferentes voos no sistema, entre outros[4]. Embora tenha sucesso em diminuir o custo de atraso do ponto de vista da entidade aeroportuária, os interesses das companhias aéreas não são levados em conta. A medida tradicionalmente utilizada para medir o custo dos atrasos é o tempo. Trabalhos realizados nos últimos anos propões que novas medidas devem ser avaliadas, como o valor financeiro destes atrasos. A solução baseada no interesse das companhias aéreas, buscou diminuir o número de voos para liberação da pista, distribuindo os slots restantes entre as companhias segundo o Equilíbrio de Nash do jogo na competição dos slots[5]. No entanto, este trabalho não leva em consideração os interesses do sistema aeroviário.

A partir destas abordagens, Ribeiro propõe uma solução baseada na negociação entre aeronaves na fila de espera pela utilização da pista [2]. As aeronaves realizariam ofertas entre si, buscando novas posições na fila, cada uma com o objetivo individual de diminuir

o custo de seu atraso. As ofertas são constituídas de deslocamentos no horário previsto de decolagem. O comportamento do agente é definido pelo Equilíbrio de Nash encontrado no jogo que representa a negociação. Se as aeronaves não forem capazes de chegar a uma decisão, a entidade aeroportuária intervém, escolhendo a oferta que será efetivada. Este modelo apresenta algumas limitações.

A primeria limitação é a necessidade de trocas de informação perfeita sobre o custo do atraso individual das aeronaves durante a etapa de negociação. Como no mundo real as companhias aéreas competem entre si, estas podem não estar dispostas a compartilharem estas informações. Outra limitação é que função custo utilzada pelo agente e pala entidade aeroportuária é a mesma. Isso é um problema, uma vez que os interesses das partes diverge e diferentes métricas podem ser utilizadas. A última limitação encontrada é que, as ofertas são limitadas a posições na fila, que já se encontra saturada. Assim, o grau de liberdade na escolha de ofertas para os jogadores é muito pequeno.

A solução apresentada neste trabalho busca atacar as limitações encontradas no trabalho de Ribeiro[2]. No modelo proposto, o agente é capaz de compartilhar parte dos ganhos esperados decorridos da aceitação da oferta com o oponente, como uma forma de incentivo para que este aceite-a. Desta forma, é criada uma nova dimensão na negociação, permitindo um grau maior de liberdade na elaboração de ofertas e permitindo a existência de ofertas benéficas a ambos os jogadores que, do contrário não existiriam.

O comportamento dos agentes passa a ser definido por Aprendizado por Reforço. Aprendizado por Reforço busca compreender como um agente pode se tornar proficiente em um ambiente previamente desconhecido valendo apenas de suas percepções do ambiente e de recompensas ocasionais[6]. Busca-se mapear situações à ações, com objetivo de maximizar o sinal de recompensa. O agente não é instruído em quais ações retornam os melhores resultado a-priori, senão aprende experimentando cada ação. Desta forma, não é necessário que o agente seja instruído sobre como os demais jogadores tomam suas decisões na negociação, aprendendo a partir das repostas dadas às ofertas realizadas. Este mesmo mecanismo, permite que o agente utilize sua própria função de custo.

1.2 Objetivos

Este trabalho tem como objetivo verificar se um agente que aprende por reforço é capaz de aprender o comportamento ótimo em um processo de negociação com outros agentes já treinados.

Para alcançar o objetivo geral, foram traçados os seguintes objetivos específicos:

1. Modelar o problema de Negociação de Slots como um Processo de Decisão de Markov

2. Verificar o resultado do aprendizado na performance do agente treinado.

3. Avaliar o desempenho do sistema a partir do treinamento de diversos agentes.

1.3 Organização do Trabalho

O trabalho está organizado da seguinte maneira: o capitulo 2 apresenta os principais conceitos em teoria dos jogos e aprendizado por reforço, necessários ao entendimento da solução proposta. No capitulo 3 são apresentados trabalhos relacionados na área de alocação de slots. No capitulo 4 é apresentado a solução proposta, definindo o problema o qual ela busca resolver e sua modelagem. A Capitulo 5 apresenta a implementação da solução e resultados obtidos nos estudos de caso. Por fim, o Capitulo 6 apresenta as conclusões sobre a realização do trabalho.

Capítulo 2

Fundamentação Teórica

2.1 Teoria dos jogos

Teoria dos jogos é um campo de estudo da matemática que estuda e modela as interações entre agentes racionais na tomada de decisão. Um agente é racional se ele possui um objetivo e emprega a melhor estratégia disponível para alcança-lo. Noções sobre teoria dos jogos nos permitem entender melhor como modelar o problema.

2.1.1 Conceitos Gerais

Jogos podem ser classificados em diferentes níveis de detalhe[7]. Jogos de coalizão envolvem uma descrição de alto nível, especificando apenas as possíveis recompensas que um grupo, ou coalizão pode receber se cooperarem. Os mecanismos pela qual a interação ocorre não são delineados. Um exemplo são os integrantes de um parlamento. Cada partido possui poder proporcional ao número de assentos que ocupa no parlamento. Jogos de coalizão descreveriam as possíveis coalizões que levariam a uma maioria no parlamento, mas desconsideraria o processo de negociação que levaria a votação em bloco [8].

Em contraposição aos Jogos de Coalizão, a Teoria dos Jogos Não-Cooperativos analisa as decisões estratégicas. No paradigma não cooperativo, a ordem e o momento das escolhas dos jogadores é crucial. Em oposição ao modelo de coalizão, no modelo não cooperativo é postulado o processo específico pelo qual é pré-definido quem realiza uma oferta, ou toma uma ação, em um determinado momento. Cooperação entre os jogadores é passível de emergir, se esta melhorar a performance para estes jogadores.

Jogos ainda podem ser separados quanto ao horizonte de ação dos jogadores [7]. Jogos na Forma Normal avaliam as estratégias dos jogadores e as recompensas associadas, também chamada de utilidade, que avalia a estratégia o jogador. Já Jogos na forma Extensiva avaliam como o jogo é jogado através do tempo. Isto inclui a ordem que os jogadores to-

mam as ações, momentos o qual informação é provida aos jogadores e o momento qm que a incerteza associada é resolvida.

2.1.2 Jogos na forma normal

Em Jogos na Forma Normal, ou Jogos Estratégicos, a situação de conflito é modelada entre n jogadores [9]. Neste jogo, cada jogador interage com os outros escolhendo uma ação ou estratégia. A escolha é feita sem informação sobre as estratégias dos demais jogadores. O jogo pode ser descrito como uma tupla (n, A1,..., n, R1,...n) na qual:

- 1,..., n é uma coleção de participantes, chamados de jogadores;
- A_k é o conjunto finito de ações disponíveis ao jogador;
- $R_k : A_1 \times ... \times A_n \rightarrow \mathbb{R}$ é a função de recompensa individual para o jogador K, que especifica a recompensa recebida para uma jogada $\mathbf{a} \in A_1 \times ... \times A_n$

O jogo se desenvolve a partir da escolha independente de cada jogador sobre uma ação individual α pertencente ao seu conjunto particular de ações A_k. A combinação de ações de todos os jogadores constitui um perfil de ações, ou uma ação conjunta $\mathbf{a} \in \mathbb{A}$ [10].

Uma estratégia $\sigma_k : A_k \rightarrow [0, 1]$ é um elemento de $\mu(A_k)$, o conjunto de distribuições de estratégias de A_k do jogador k. Uma estratégia é dita pura se $\sigma_k(a) = 1$ para uma alguma $a \in A_k$ e 0 para todas as outras ações, caso contrário, a estratégia é dita mista. Um perfil estratégico $\sigma = (\sigma_1, ..., \sigma_k)$ é um vetor de estratégias contendo uma estratégia para cada jogador. Se todas as estratégias dos jogadores forem puras, o perfil estratégico corresponde a uma ação conjunta $a \in \mathbb{A}$. A recompensa esperada por um jogador k em um perfil estratégico $\sigma \in \mu(A_k)$ é dada por:

$$R_k(\sigma) = \sum_{a \in \mathbb{A}} \prod_{j=1}^{n} \sigma_j(a_j) R_k(\mathbf{a}) \tag{2.1}$$

em que a_j é a ação do jogador j no perfil de ação $\mathbf{a}$.

2.1.3 Jogos na forma Extensiva

Jogos na forma Extensiva possuem descrições profundas sobre o funcionamento do jogo, descrevendo qual jogador deve atuar, quando, quais são suas escolhas, o resultado desta escolha e a informação disponível em cada estágio. Jogos na Forma Extensiva podem ser utilizados para modelar interações que se prologam e nas quais ações tomadas repercutem ao longo do tempo [1].

Formalmente, um jogo de $n - jogadores$ na forma extensiva pode ser definido por:

- Um conjunto N = 1,2,..., n de jogadores
- Uma árvore T, chamada árvore de jogo
- Uma partição do conjunto de nós não terminais de T em n+1 subconjuntos, denotados por P^0, P^1, ... , P^n. Os membros de P^0 são chamados de nós de chance. Para cada $i \in N$, os membros de P^i são chamados de nós do jogador i.
- Para cada nó em P^0, existe uma distribuição de probabilidades entres os ramos de saída.
- Para cada $i \in N$, exite uma partição de P^i em $k(i)$ conjuntos de informação, U_1^i, ..., $U_{k(i)}^i$, tal que para cada $j = 1, ..., k(i)$:
    1. Todos os nós em U_j^i possuem o mesmo número de nós de saída, existe uma correspondência um-para-um entre os conjuntos de ramos de saída para os diferentes nós em U_j^i.
    2. Cada caminho em uma arvore da raiz até um nó terminal, pode passar por um U_j^i no máximo uma vez.
- Para cada nó terminal $t \in L(T)$, existe um vetor de dimensão n de recompensas associadas a cada jogador. $r(t) = (r^1, ..., r^n)$.
- A descrição completa do jogo é de conhecimento entre os jogadores.

Exemplo: Para N=$\{1,2,3\}$; raiz = a; P^0 = d; P^1 = {a,e,f}; U_1^1 = {a}; U_2^1 = {e,f}; $P^2 = U_1^2 = \{b, c\}$; $P^3 = U_1^3 = \{g\}$. Os ramos, probabilidades associadas a d e recompensa associada a estados terminais é representado em 2.1. Cada nó representa a tomada de decisão de um jogador, as arestas que ligam os nós representam as ações executadas pelos jogadores. Os nós terminais representam as recompensas associas à sequencia de ações realizada.

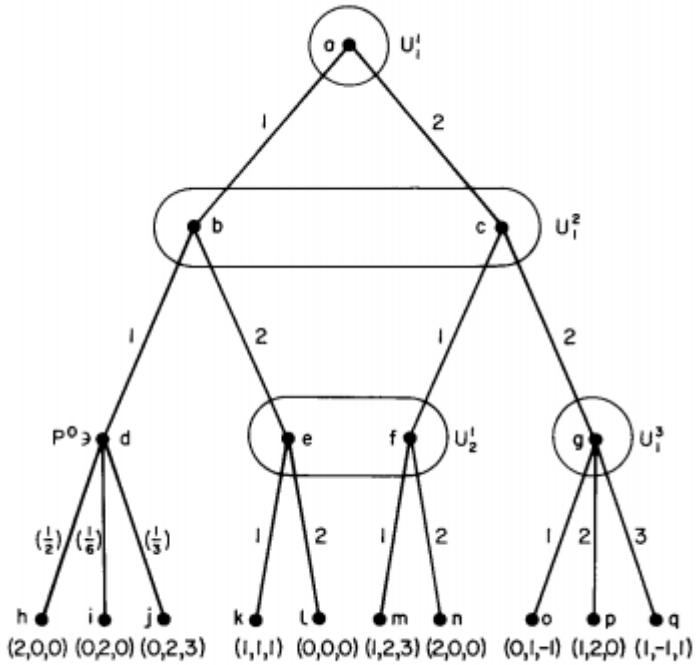

Figure 6. The game tree of Example 1.3.

Figura 2.1: Exemplo de Jogo na forma Extensiva, reproduzido de Hart [1].

2.1.4 Equilíbrio em Jogos

Como as recompensas em jogos são funções relativas a cada jogador, o conceito de solução não é simples. Alguém poderia dizer que a solução de um jogo seria a configuração na qual todos os jogadores maximizam suas recompensas. No entanto, para muitos jogos, é impossível que todos os jogadores atinjam todos os seus objetivos em um mesmo tempo.

Do ponto de vista de um jogador isolado, podemos definir o conceito de melhor resposta. Seja $\sigma = (\sigma_1, ..., \sigma_n)$ o perfil estratégico e σ_{-k}, o mesmo perfil com a ausência da estratégia σ_k do jogador k. A estratégia $\sigma_k \in \mu(A_k)$ é a melhor estratégia se:

$$R_k(\sigma_{-k} \cup_k^*) \geq R_k(\sigma_{-k} \cup_k^{'}) \forall \sigma_k^{'} \in \mu(A_k), \tag{2.2}$$

no qual $R_k(\sigma_{-k} \cup_k^{'})$ denota o perfil estratégico, em que todos os jogadores jogam com o mesmo perfil estratégico que σ exceto pelo jogador k que joga com $\sigma_k^{'}$.

Definida a melhor resposta, podemos definir Equilíbrio de Nash como um perfil estratégico $\sigma = (\sigma_1, ..., \sigma_n)$ no qual, para cada jogador k, a estratégia σ_k é a melhor resposta para o perfil dos outros jogadores σ_{-k}.

2.1.5 Jogos Evolucionários

Teoria dos Jogos Clássica assume que todos os jogadores possuem informação completa sobre o jogo, o que somada a suposição que os agentes atuam de forma racional implica que a modelagem muitas vezes não reflete situações do mundo real. Jogos evolucionários diminuem a ênfase em racionalidade e a substitui por dinâmicas biológicas como mutações e reprodução.

Um conceito chave à teoria de jogos evolucionários, está a dinâmica de replicação, que descreve como uma população de indivíduos evolui ao longo do tempo quando submetida a pressão evolutiva. Seu sucesso reprodutivo é definido pela sua adequação resultante destas interações. A dinâmica de replicação dita que a porção da população com adequação superior a média da adequação de toda população irá aumentar, enquanto as parcelas com adequação inferior irão diminuir.

No entanto o modelo evolucionário também pode ser entendido como a estratégia de um jogador, desta forma a proporção de indivíduos de um determinado tipo na população é análoga a distribuição de probabilidades das ações do jogador[11]. As dinâmicas de replicação agora descrevem como as estratégias do jogador evoluem ao longo do tempo, fazendo a ponte entre Teoria dos jogos e os métodos de aprendizado.

Reinforcement Learning	Game Theory	Evolutionary Game Theory
environment	game	game
agent	player	population
action	action	type
policy	strategy	distribution over types
reward	payoff	fitness

Figura 2.2: Diferentes Nomenclaturas para conceitos análogos em distintas áreas de conhecimento.

2.2 Aprendizado por Reforço

2.2.1 Conceitos Gerais

Aprendizado por reforço é uma abordagem computacional que busca automatizar aprendizado orientado ao objetivo e tomada de decisão. Se distingue de outras abordagens de aprendizado, como aprendizado supervisionado e não supervisionado, por focar em um agente que aprende através de interação direta com o ambiente, sem necessitar supervisão ou modelos completos do ambiente [6].

A tarefa de aprendizado por reforço consiste em utilizar recompensas observadas para aprender uma política ótima ou quase ótima para o ambiente. Imagine disputar um jogo cujas regras não são conhecidas. Depois de aproximadamente uma centena de movimentos, seu oponente anuncia que "você perdeu" [12].

Além do agente e do ambiente, um sistema de aprendizado por reforço pode ser descrito por uma política, um sinal de recompensa, uma função valor e, opcionalmente, um modelo para o ambiente [6].Uma política define como comportamento do agente para um determinado instante. A política mapeia os estados percebidos as ações que devem ser tomadas em tais estados. Em certos casos, a política pode ser definida por uma função simples ou uma tabela de consulta, enquanto em outras, processos computacionais mais complexos podem estar envolvidos. A definição de políticas é a essência de aprendizado por reforço, sendo suficiente para determinar o comportamento do agente.

O sinal de recompensa instancia o objetivo do problema de aprendizado. A cada instante, o agente recebe um valor escalar do ambiente chamado de recompensa. O objetivo do agente é maximizar a recompensa total recebida ao longo de um longo tempo. Desta forma, o sinal de recompensa indica ao agente o que são bons e maus eventos no ambiente. O sinal de recompensa funciona como o principal de indicador para ajustes nas políticas do agente. Se o sinal de recompensa avalia o desempenho do agente no instante, a função valor busca identificar o desempenho do agente a longo prazo. A função valor computa para cada estado, a recompensa esperada futurada partindo deste estado.

O quarto elemento é um modelo do ambiente. Este modelo imita o comportamento do ambiente, permitindo que inferências sobre como o ambiente reage ao agente. Isto permite que o próximo estado seja previsto, bem como a recompensa associada. Modelos são utilizados para planejamento, isto é, antes de escolher uma ação, o agente considera os estados futuros antes destes serem visitados de fato. Podemos separar os métodos para resolução de problemas de aprendizado por reforço entre os baseados em modelo, que utilizam planejamento, e os livres de modelo, que aprendem baseado em tentativa e erro.

2.2.2 Processo de Decisão de Markov

O processo de Decisão de Markov pode ser utilizado como estrutura base ao aprendizado por reforço [6]. Um processo de decisão de Markov pode ser descrito como uma tupla (S, A, T, R, γ), na qual S é o conjunto de estados do ambiente, A é o conjunto finito de ações do agente, T é a função de transição de estados $f : S \times A \times S \rightarrow [0, 1]$ e R é a função de recompensa $R : S \times A \times S- > \mathbb{R}$ e γ é o fator de desconto, utilizado para modelar a diferença de preferência entre recompensas imediatas e futuras.

S_t é o sinal de estado que descreve o ambiente no instante t. O agente pode alterar o estado a cada instante tomando uma ação $a_k \in A$. Como resultado o agente transita

de um estado s_t para um estado $s_{t+1} \in S$, de acordo com a função T. A probabilidade que o agente chegue em S_{t+1} ao agir de forma a_t é dada por $T(s_t, a_t, s_{t+1})$ e a recompensa associada é dada por $T(s_t, a_t, s_{t+1})$. Para cada ação o agente recebe uma recompensa $r_t \in \mathbb{R}$ dada por $R(s_t, a_t, s_{t+1})$. Esta recompensa está associada somente a ação tomada em t, e não relaciona diretamente ao impacto a longo prazo.

Em modelos determinísticos, T é substituída por $T : SxA \to S$ e R pode ser definida por $R : S \times A \to \mathbb{R}$. A política h do agente pode ser estocástica, dada por $h : SxA \to [0.1]$ ou determinística $h : S \to A$.

O objetivo do agente é maximizar a cada instante t o retorno descontado esperado:

$$R_k = E(sum_{i=0}^{\infty} \gamma^j r_{k+j+1}) \tag{2.3}$$

A agente busca maximizar o resultado a longo prazo R_k, enquanto recebe resposta baseada em ações imediatas. A maneira encontrada para isto é calcular uma função *action-value* (Q-function), $Q^h : S \times A \to \mathbb{R}$, que retorna o valor esperado de recompensa para cada par ação estado:

$$Q^h = E(sum_{j=0}^{\infty} \gamma^j r_{k+j+1} | s_k = s, a_k = a, h) \tag{2.4}$$

Podemos entender a tarefa de aprendizagem como o calculo da função Q^*, definida como Q-function ótima:

$$Q^* = arg \max_h Q^h(x, u) \tag{2.5}$$

A partir de Q^*, o agente pode adotar uma estratégia determinística e gananciosa, escolhendo para cada estado a ação que retorna o maior valor:

$$h^* = arg \max_a Q(s, a) \tag{2.6}$$

2.2.3 Q-Learning

Q-learning é um método de aprendizado por reforço livre de modelo que provê a capacidade aos agentes de aprenderem a agir de forma ótima em ambientes markovianos. Os agentes não necessitam de um modelo completo do ambiente para que o aprendizado se inicie, utilizando dos resultados de suas ações como guia [13]. O processo de aprendizagem ocorre através de episódios, no episódio n o agente:

1. Observa seu estado atual x_n
2. Seleciona e performa a ação a_n
3. Observa o estado subsequente y_n

4. Recebe uma recompensa imediata r_n, e

5. Ajusta o valor Q_{n-1} utilizando um fator de aprendizado α na forma:

$$Q_n(x,a) = . \begin{cases} (1\text{-}\alpha_n)Q_{n-1}(x,a) + \alpha_n[r_n + \gamma V_{n-1}(y_n), \text{ se x= } x_n \text{e a= } a_n \\ Q_{n-1}(x,a), \text{ caso contrario} \end{cases} \tag{2.7}$$

no qual:

$$V_{n-1} \equiv \max_b Q_{n-1}(y,b) \tag{2.8}$$

representa o maior valor esperado que o agente entende que pode ser obtido a partir do estado y_n. Na etapa inicial de aprendizado os valores da função podem ainda não refletirem a política que eles definem, ou seja as ações aprendidas ainda não maximizam V_{n-1}. Assume que existem valores para todos os pares de ação estado.

A garantia de optimalidade de Q-learning, ou seja que a função $Q_n(x,a)$ converge para $Q^*(x,a)$ a medida que $n \to \infty$ é dada pelo teorema [13]: Dadas as recompensas limitadas $|r_n| < \mathbb{R}$, taxas de aprendizado $0 \leq \alpha < 1$ e:

$$\sum_{i=1}^{\infty} \alpha_{n^i(x,a)} = \infty, \sum_{i=1}^{\infty} [\alpha_{n^i(x,a)}]^2 < \infty, \forall x, \alpha, \tag{2.9}$$

então $Q_n(x,a) \to Q^*(x,a)$ quando $n \to \infty, \forall x, \alpha$, com probabilidade 1.

Chamamos de política gananciosa a politica que busca maximizas a recompensa recebida a cada iteração. A política gananciosa implícita em 2.8 designa probabilidade 1 para a ação que maximiza o valor de Q(x,a) e pela garantia de optimalidade, aproxima Q de Q^*. No entanto se utilizada durante o processo de aprendizagem, a política gananciosa não garante que todos os estados serão visitados um número suficiente de vezes. Desta forma, a política gananciosa pode ser utilizada para explotar de forma ótima o modelo aprendido, mas não para aprende-lo.

O equilíbrio entre exploração e explotação do modelo aprendido é fundamental a tarefa de aprendizado. Um agente que apenas explora, mas não age em acordo com os valores aprendidos falha em colocar em prática o aprendizado, enquanto um agente que apenas utiliza as melhores ações já aprendidas não evolui sua performance.

Uma maneira de solucionar este problema é escolhendo ações que são gananciosas no limite com exploração infinita(GLIE)[14]. A política utilizada pelo agente irá convergir à política gananciosa eventualmente, no entanto exploração é permitida no meio tempo. Um exemplo é a política ϵ-greedy. Seja $n(s)$ o número de vezes que o estado s foi visitado. Ao visitar o estado s o agente opta pela ação gananciosa com probabilidade $1 - \epsilon(s)$ ou opta por uma ação aleatória com probabilidade $\epsilon(s)$, sendo $\epsilon(s) = c/n(s)$ e $0 < c < 1$.

Um agente converge em comportamento se no limite a distribuição de probabilidades das ações disponíveis se torna estacionária. Agentes adotando uma política GLIE podem não convergir em comportamento uma vez que empates entre ações igualmente valiosas são decididos aleatoriamente.

No entanto, se existir apenas uma política ótima e a função Q-value convergir, um agente utilizando Q-learning irá convergir em comportamento [15]. Um agente converge em comportamento se a partir de um determinado momento, a política gulosa não se altera, ou seja, para cada estado a ação de maior valor permanece a mesma, independete de quantas vezes aquele estado for visitado futuramente e de qual ação é escolhida. A convergência de comportamento é particularmente importante em um sistema multi-agente, uma vez que as ações de um agente dependem das ações dos demais. Desta forma, um agente poderia prever o comportamento de outro agente ao assumir que este irá agir de forma gananciosa. A convergência de comportamento permite que esta previsão da ação dos demais seja consistente ao longo do tempo.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são apresentados trabalhos relacionados à alocação de slots em aeroportos. O objetivo é apresentar o contexto no qual o trabalho se insere e que contribuições ele traz a área.

3.1 Alocação de slots

O sistema de pistas foi identificado como sendo o principal gargalo na capacidade aeroportuária devido às exigências operacionais nas operações de pista. Desta forma, mesmo melhorias singelas no desempenho da pista tem impactos significativos no atraso sistemático [16]. As áreas terminais, nas quais os voos se iniciam e encerram, constituem ambientes dinâmicos e incertos, com constante atualização dos estados das aeronaves através de sistemas de monitoramento. A natureza dinâmica do problema torna necessário o desenvolvimento de algoritmos eficientes que possam replanejar a utilização dos recursos quando novos eventos ocorrem, como quando aeronaves adentram o perímetro ou novos dados se tornam disponíveis.

Balakrishnan et al [4] propõem um modelo baseado em restrições e programação dinâmica no paradigma CPS (Constraint Position Shifiting). CPS estipula que uma aeronave não pode ter sua posição na fila de espera pelo uso da pista alterada mais do que um determinado número de posições. As restrições operacionais do modelo são divididas em: justiça, requerimentos de separação mínima, restrições quanto a janela de tempo e Restrições de precedência. No entanto, o modelo não leva em consideração as preferências das companhias aéreas.

No mundo real cada companhia área tem interesses próprios e busca maximizar seus ganhos. Surge, então, a necessidade de modelar o comportamento competitivo das aeronaves no processo de competição dos *slots*. Vaze e Barnhart [**?**] apresentam uma solução utilizando Teoria dos Jogos. A solução é baseada na divisão da capacidade total entre as

companhias baseada na maximização dos lucros, na qual estas disputam fatias de mercado. Nos experimentos realizados, o número de *slots* alocados no aeroporto de LaGuardia foi diminuído e os *slots* remanescentes distribuídos de acordo com o equilíbrio de Nash encontrado entre as companhias. A diminuição em 12% do número de *slots* resultou em um aumento de lucro operacional de 19,2% e uma diminuição do atraso por voo de 40,97% e de atraso por passageiro 59.03% com a redução do número de passageiros de apenas 2,27%. No entanto, o modelo tem como limitação o fato que a redução da capacidade de um aeroporto, embora lucrativa as companhias, pode não atender aos interesses do sistema aeroportuário.

O modelo baseado em restrições trata dos conflitos na competição por *slots* das aeronaves, porém não garante que a sequencia final montada está livre de atrasos desnecessários. Com o intuito de resolver este problema, Ribeiro [2] propõe uma abordagem baseada em teoria dos jogos que leva não apenas em consideração os interesses das aeronaves, mas permite que a entidade aeroportuária arbitre o resultado das negociações. Nesta abordagem, as aeronaves são alocadas em uma fila de acordo com sua ordem de chegada. A fila é percorrida, dando a cada aeronave a chance de negociar sua posição com a anterior. A negociação se dá numa série de pares de propostas, nos quais, as aeronaves envolvidas alternam nos papeis de proponente e oponente, buscando dividir de forma justa o atraso entre si. Se nenhuma oferta for aceita, ou a oferta acordada entre as partes violar restrições operacionais, a entidade aeroportuária interfere na negociação. Nos experimentos realizados no aeroporto de Brasília, o sistema teve bons resultados. Na alocação estática, realizada a partir dos horários programados de decolagem das aeronaves, sem considerar as operações de pouso, o sistema foi capaz de diminuir o atraso total do sistema em 17%. Considerando a chegada de novas aeronaves o sistema diminui o atraso total em 19%.

3.2 Justificativa metodológica

A solução proposta por este trabalho, denominada RELEASE, parte da solução apresentada por Ribeiro[17] para elaborar um processo de negociação entre aeronaves. A solução apresentada possui duas distinções fundamentais: A utilização de aprendizado por reforço para estabelecimento da política dos agentes e diferenciação dos critérios de decisão das aeronaves e da entidade aeroportuária.

A utilização de aprendizado por reforço em detrimento de uma solução baseada em teoria dos jogos permite que a política resultante do aprendizado seja dinâmica. Na solução baseada no equilíbrio de Nash, se uma das premissas nas quais o jogo se baseia for violada, o equilíbrio se perde e a política designada se torna obsoleta devido modificação das recompensas dos jogadores. Em uma solução por aprendizado por reforço é possível

que o agente se adapte com o tempo ao novo paradigma, desde que seja capaz de modelar os atributos relevantes.

Além disso, a solução por teoria dos jogos necessita que todas os jogadores tenham informação perfeita sobre os custos dos outros jogadores. Assumisse que a equação da função custo é a mesma para todas as aeronaves e que estas estão dispostas a compartilhar o valor dos atributos utilizados no cálculo. Esta situação pode não refletir o mundo real, uma vez que as companhias aéreas são concorrentes entre si, podem não estar dispostas a compartilhar esta informação. No aprendizado por reforço, o agente não necessita conhecer o processo de decisão do outro agente para tomar uma decisão. Desde que os atributos envolvidos sejam mapeados, o agente pode mapear o comportamento dos outros agentes apenas observando suas ações.

Outra diferença, é que no modelo base, o critério de decisão da entidade aeroportuária é o mesmo das aeronaves. A solução proposta permite que as aeronaves negociem segundo seus próprios critérios de avaliação de custos, enquanto a entidade aeroportuária arbitre segundo sua própria lógica. Assim como no custo das outras aeronaves, o agente não necessita saber o mecanismo interno de decisão, apenas os resultados da arbitragem.

Na solução proposta é criada uma nova dimensão na negociação, na qual é permitido ao agente negociar parte de seus ganhos com o oponente, além das posições na fila. Além disso, a abordagem proposta permite que os interesses da entidade aeroportuária divirjam dos interesses das aeronaves e não é necessário a cada entidade ter conhecimento perfeito sobre os atributos dos demais. As diferenças entre os modelos são ilustradas na Tabela 3.1.

Trabalho	Melhoria do Atraso	Interesses das Companhias	Interesse Aeroportuário	Diferenciação dos Interesses
Balakrishnan	✓	×	✓	×
Vaze e Barnhart	✓	✓	×	×
Ribeiro	✓	✓	✓	×
RELEASE	✓	✓	✓	✓

Tabela 3.1: Diferença entre modelos.

Capítulo 4

Solução Proposta

Este capitulo tem como objetivo apresentar a solução desenvolvida neste trabalho. Para isso, primeiramente é exposto o problema de sequenciamento de partidas em aeroportos. Em sequência, é apresentada a modelagem da solução, denominada RELEASE *(Reinforcement Learning Airport Scheduler)*, na qual são descritos o ambiente, o agente e a processo de negociação.

4.1 Descrição do Problema

Todo voo parte de uma origem e transita em vias de navegação até chegar a seu destino. Os pontos de origem e destino são denominados aeródromos, lugares onde é possível pouso, decolagem e a movimentação de aeronaves. Aeroportos são aeródromos públicos, dotados de infraestrutura para facilitar as operações das aeronaves no embarque e desembarque de passageiros e carga. A execução do voo se divide nas seguintes fases [18]:

1. Fase pré-voo: Compreende as fases de planejamento pré-tático e tático.
2. Taxiamento: O piloto é autorizado a deslocar a aeronave até a margem da pista e acionar os motores.
3. Decolagem: A torre de controle (TWR) do aeródromos autoriza a decolagem da aeronave. Ao atingir a altitude de 35 pés, o controle passa da torre para o controle de aproximação (APP), que será responsável por guiar a aeronave até que ela saia do setor terminal (TMA).
4. Ascendência: A aeronave continua ascendendo até a altura de cruzeiro prevista no plano de voo, saindo do setor terminal e entrando na aerovia prevista.
5. Cruzeiro: A aeronave transita entre as aerovias conforme seu plano de voo e o controle da área responsável (ACC)

6. Descida Inicial: A aeronave reduz gradualmente a sua altitude ao se aproximar da TMA. Caso a área terminal esteja congestionada, são aplicadas medidas restritivas para que a aeronave adie sua entrada na área terminal. A aeronave passa a voar em círculos aguardando que a autorização seja dada.

7. Aproximação Final: Quando chega a área terminal de destino, o controle volta a ser da ACC. A aeronave diminui sua altitude e velocidade afim de chegar a pista de pouso nas condições adequadas. Para que possa aterrissar, a aeronave depende da aprovação da torre.

8. Taxiamento: Após a aterrissagem, a aeronave é manobrada pela equipe em terra de forma a não atrapalhar as demais operações de pouso e decolagem.

A fase mais crítica do voo se dá nas movimentações próximas as áreas terminais devido a intensa movimentação de outras aeronaves. Dependendo da capacidade da área terminal, APP pode requisitar que as aeronaves em solo atrasem sua decolagem até que o número de aeronaves no ar diminua. De forma análoga, se a capacidade do aeródromo estiver lotada, a APP coloca as aeronaves no ar em espera até que haja vagas no aeródromo ou redireciona-a a outros aeródromos. As alterações dos horários de decolagem e aterrissagem se dão na etapas de planejamento que podem ser divididas em [18]:

1. Planejamento estratégico: Nesta fase busca-se balancear a demanda e a capacidade. O planejamento acontece entre seis meses antes do voo e encerra-se na antevéspera e é realizado pelo Centro de Gerenciamento de Navegação Aérea brasileiro em conjunto com os CINDACTAS;

2. Planejamento pré-tático: é iniciado na véspera e se encerra até duas horas antes do inicio do voo. São consideradas as projeções de demanda junto aos órgãos de controle e possíveis alterações na capacidade devido a mudanças na infraestrutura aeroportuária ou condições meteorológicas.

3. Planejamento tático: se inicia duas horas antes da execução do voo e se encerra somente na sua aterrissagem. Nesse planejamento, estão inseridas as medidas de regulação para resolução de eventos não previstos.

Para que as manobras feitas pelas aeronaves não ponham em risco a si próprias, os tripulantes, bem como a própria infraestrutura do aeródromo, é estipulado uma separação segura entre as aeronaves baseada no porte das aeronaves envolvidas como mostrado na Tabela 4.1.

Aeronave antecedente	Aeronave em Sequencia	Separação mínima
pesada	pesada	2 min
	média	2 min
	leve	3 min
média	leve	3 min

Tabela 4.1: Separação Miníma por combinação de aeronaves

4.2 Solução Proposta

Nesta seção é apresentado a solução RELEASE *(Reinforcement Learning Airpot Scheduler)*, um mecanismo por aprendizado por reforço para ensinar aeronaves a negociarem seus atrasos na fila de decolagem.

4.2.1 Visão geral

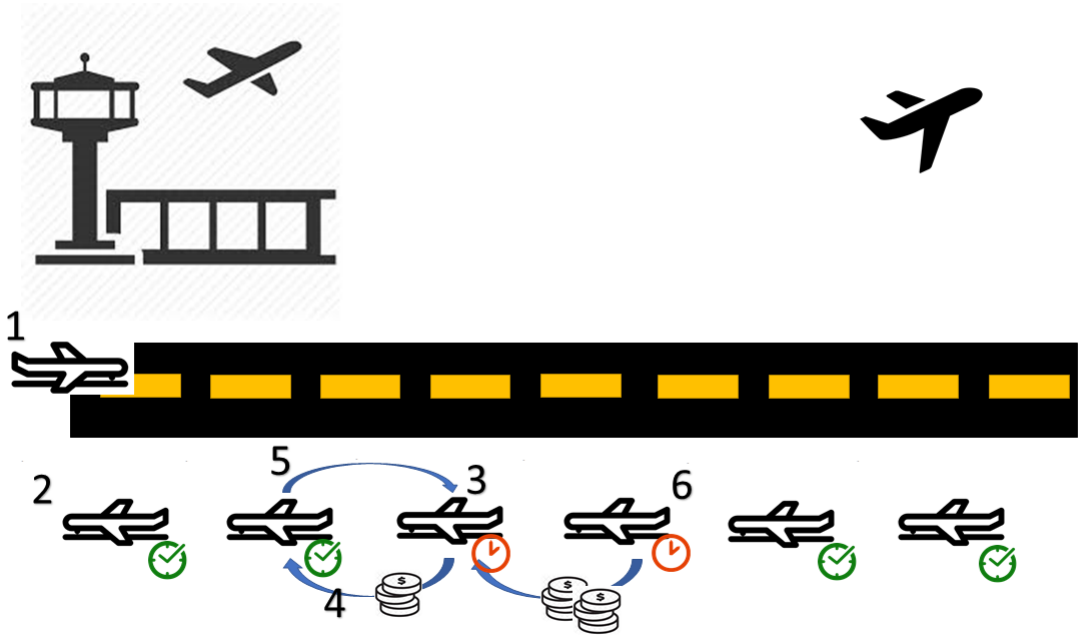

Figura 4.1: Modelo Geral da Negociação de Slots de Decolagem

A Figura 4.1 representa o cenário de negociação de *slots* de decolagem em um aeroporto. O problema pode ser definido como a necessidade de realizar o agendamento do uso da pista da melhor maneira possível atendendo às garantias operacionais de segurança (Figura 4.1,

rótulo 1). Isso inclui como distribuir as medidas restritivas de forma que o custo total do sistema seja diminuído, mantendo um critério de justiça, no qual uma aeronave não é penalizada de forma assimétrica.

Os voos são ordenados em uma fila quanto os seus horários de decolagem pretendido e é aplicado a separação segura entre eles (Figura 4.1, rótulo 2). Esta separação acarreta em possíveis atrasos as aeronaves (Figura, rótulo 3). É dada então, a chance para que elas negociem a distribuição dos custos associados entre si.

A cada atualização da fila, é iniciada uma nova etapa de negociações. Em cada negociação, as aeronaves alternam nos papéis de proponente e oponente. Em cada turno da negociação, o proponente envia uma oferta ao oponente, que a avalia (Figura 4.1, rótulo 4). A oferta consiste em deslocamentos das duas aeronaves na fila de espera, bem como uma recompensa associada.

Se o opoente aceitar ou rejeitar, a negociação se encerra. Caso o oponente escolha contra-atacar (Figura 4.1, rótulo 5), a negociação segue para um novo turno com as aeronaves trocando de papéis. Este arranjo se prolonga até que um acordo seja atingido entre as aeronaves ou a entidade aeroportuária interfira. Para que uma aeronave consiga convencer outra, é necessário que seja feita uma proposta que diminua o custo de atraso da aeronave oponente(Figura 4.1, rótulo 6).

Para resolução do problema de agendamento de decolagens e distribuição justa dos atrasos, propõe-se um modelo baseado em aprendizado por reforço onde as aeronaves aprendem a dividir os custos relacionados às medidas restritivas entre si, baseadas em seus interesses pessoais com arbitragem final da entidade aeroportuária.

Para que a aeronave possa negociar, é necessário que ela saiba qual a ação que lhe traz maior ganho. Para isso as aeronaves são treinadas em uma série de simulações onde negociam entre si.

4.3 Modelo RELEASE

Nesta seção é apresentada a modelagem da solução, formada pelo ambiente no qual o agente atua, as demais aeronaves que agem como oponentes e o agente em si. São apresentados duas modelagens de negociação, uma referente ao comportamento padrão das aeronaves oponentes e outro referente ao aprendizado do agente, modelado como um Processo de Decisão de Markov.

4.3.1 Modelagem do ambiente

Para modelagem do ambiente é utilizada o modelo apresentado por Ribeiro [2]. O cenário aeroportuário é modelado com base nos parâmetros definidos na Tabela 4.2:

F	Conjunto dos voos ativos. Formado por F_g e F_a
F_g	Aeronaves em solo
F_a	Aeronaves no ar
I_p	Capacidade aeroportuária instalada, condicionada pelo de estacionamento e pistas
$C_a(t)$	Capacidade de aterrissagem no aeroporto no instante t
$C_d(t)$	Capacidade de decolagem no aeroporto no instante t
Qa	Conjunto de aeronaves com autorização para pouso.
Qd	Conjunto de aeronaves no patio esperando para decolagem
K	Custo total de atraso

Tabela 4.2: Parâmetros do cenário

O conjunto Q_d constitui as aeronaves que necessitam de *slots* para decolagem e constituem o conjunto de agentes que participarão das negociações. Quando uma aeronave decola ela é removida do conjunto Q_d e inserida no conjunto F_a^-. De forma análoga, quando uma aeronave pousa ela sai do conjunto F_a e é inserida no conjunto Q_d^-. O modelo assume que embora influenciem na capacidade do aeroporto, os conjuntos F_a^- e Fg^- não possuem interesse iminente de utilização da pista e são removidos da disputa por *Slots*.

A capacidade de decolagem é igual ao número de aeronaves que o aeroporto pode despachar em um instante t. Este número é igual ao número de pistas disponíveis, observando a utilização prioritária das pistas para pousos. A capacidade de decolagem é dada pela formula:

$$C_d(t) = I_p^r - (I_a(t) + Id(t)) \tag{4.1}$$

A capacidade de aterrissagem é defina pela Eq.(4.2):

$$C_a(t) = minI_p^f - F_g^-, I_p^r - (I_a(t) + Id(t)) \tag{4.2}$$

O custo total dos atrasos é igual a soma do custo individual dos n voos ativos.

$$K = \sum_{i=1}^{n} k_i \tag{4.3}$$

4.3.2 Modelagem das aeronaves

t_d	Horário de decolagem previamente alocado para a aeronave.
t_a	Horário de chegada previsto para a aeronave no aeroporto de destino
v	Velocidade de cruzeiro
t_f	Duração estimada de voo
γ	Fator de carga da companhia aérea
D	Máximo período de atraso aceitável para a aeronave
p	Número de passageiros presentes na aeronave
B	Porte da aeronave
σ	Significância da aeronave
l	Atraso imposto a aeronave
k	Custo ponderado individual de atraso da aeronave
S	*Slot* alocado para a aeronave

Tabela 4.3: Parâmetros da aeronave

As aeronaves são os agentes do sistema. Os atributos das aeronaves são representados na Tabela 4.3. O fator B se relaciona à exigência de separação miníma entre as decolagens devido a existência de uma esteira de turbulência, decorrente a movimentação das aeronaves. Pressupõe-se ainda, que o consumo de combustível de aeronaves de maior porte é maior, aumentando o custo de situações de espera em ar.

As aeronaves utilizadas no transporte de passageiros são utilizadas continuamente, assim, o aeroporto de destino se torna o aeroporto de origem para o voo seguinte. Logo o atraso de uma aeronave é propagado para os voos executados por aquela mesma aeronave. A janela máxima de atraso é modelada pela variável D.

Considera-se que o porte, o número de passageiros e a distância percorrida pela aeronave podem ser utilizados como indicadores para o custo operacional associado a aeronave. Desta forma, a significância de uma aeronave é dada por:

$$\sigma = vt_f B + p \tag{4.4}$$

A significância da aeronave é utilizada para calcular o custo ponderado do atraso individual das aeronaves. O custo é tido como unidade genérica. O fator de carga mensura a influência da companhia aérea proprietária da aeronave no cenário aeroportuário em questão.

$$k = \frac{l\gamma}{D}\sigma \tag{4.5}$$

Para a negociação de slots, primeiramente as aeronaves são alocados em uma fila na ordem *first in, first out*. Cada aeronave F_i é inserida no sequenciamento com o objetivo de anular k_i, o que ocorre na situação ideal. Para ocupar um novo espaço na lista, a aeronave gera deslocamento da posição das outras aeronaves. O atraso real é igual a diferença do horário previamente alocado no plano de voo e o horário de decolagem real.

A negociação entre aeronaves é efetuada com o objetivo de diminuir o custo global, com menor atraso individual possível. Na negociaçã,o a aeronave pode adotar uma de três ações:

1. Tentar capturar o *slot* seguinte, ganhando uma posição na fila;
2. Ceder seu *slot* para outra aeronave, perdendo uma posição;
3. Permanecer em seu *slot*.

Sendo detectado o conflito de interesses, surge uma situação de negociação, na qual está associado um *payoff*. Este *payoff*, para as aeronaves que não estão em aprendizado, é apenas uma medida da qualidade da negociação não sendo utilizado no sequenciamento.

O fluxo básico do processo de negociação segue as seguintes etapas:

1. Busca-se na fila de decolagens a aeronave cujo horário de decolagem alocado pelo plano de voo é o mais próximo do horário corrente;
2. Se esta aeronave tiver interesse em negociar, ela irá executar a tarefa de calcular uma oferta e enviá-la à aeronave que está imediatamente a sua frente na fila;
3. Se esta outra aeronave aceitar a oferta, a troca é efetuada e a próxima aeronave na fila é avaliada;
4. Caso contrário, o fator de desconto é incrementado e uma nova oferta é elaborada, desta vez pela aeronave oponente;
5. Se nenhuma oferta for aceita, o aeroporto decidirá qual delas é a mais interessante a nível global e tal oferta será aplicada.

O processo de negociação pode ser modelado na forma de um jogo extensivo na forma: Sejam f1 e f2 as duas aeronaves, a aeronave f1 sugere a aeronave f2 uma das seguintes propostas:

- avançar: é fornecido a f2 um horário anterior ao que está assinalado. Pode ser interessante a f2 aceitar a oferta.
- atrasar: é fornecido a f2 um horário posterior, com a finalidade de liberar seu *slot* para f1.

- permutar: f2 recebe a proposta de trocar seu *slot* com f1. O caráter parcialmente colaborativo do jogo favorece esta decisão.

Dada estas três alternativas, a aeronave f2 pode decidir por uma de três opções:

- aceitar: f2 aceita a proposta e recebe a recompensa cabível;
- recusar: f2 recusa a oferta e mantém seu custo de atraso, enquanto f1 arca com a intervenção do árbitro de negociação.
- contra-atacar: f2 elabora uma contra-oferta, reiniciando o processo de negociação e tornando-se o agente ativo da barganha.

Este processo pode ser modelado por uma árvore na qual o agente f1 sugere uma opção possível e obtém uma ação reposta do agente f2. A recompensa só é recebida quando a oferta é aceita ou rejeitada. Nos casos no qual o agente f2 opta pela contra-oferta, a arvore cresce. Cada nível da árvore é uma etapa temporal, como mostrado na Figura 4.2, sendo *acc*, *rej* e *ctr* abreviaturas para aceitar, rejeitar e contra-atacar respectivamente.

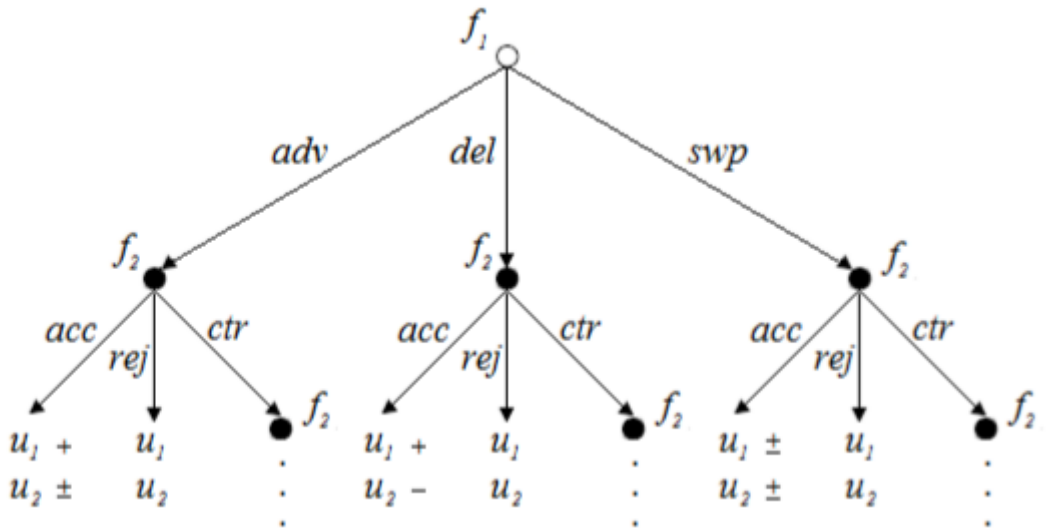

Figura 4.2: Árvore de Negociação entre duas Aeronaves, reproduzido de Ribeiro [2].

Ao iniciar uma negociação, o proponente procura fazer a melhor oferta possível. Primeiramente, tenta-se diminuir o máximo possível o atraso individual, sem incutir um atraso ao oponente. Se a separação entre o proponente e o oponente for insuficiente, propõe-se ao oponente adiantar-se o máximo possível para que o proponente possa ocupar um espaço melhor. Caso não haja espaço suficiente para que o oponente se adiante, o proponente pede que o oponente se atrase. Como o comportamento das aeronaves tende

a cooperação, o máximo de atraso proposto é que o oponente ocupe o *slot* seguinte na fila, que equivale a posição do proponente, situação onde os dois trocam de posição na fila. Este processo é descrito na Figura 4.3

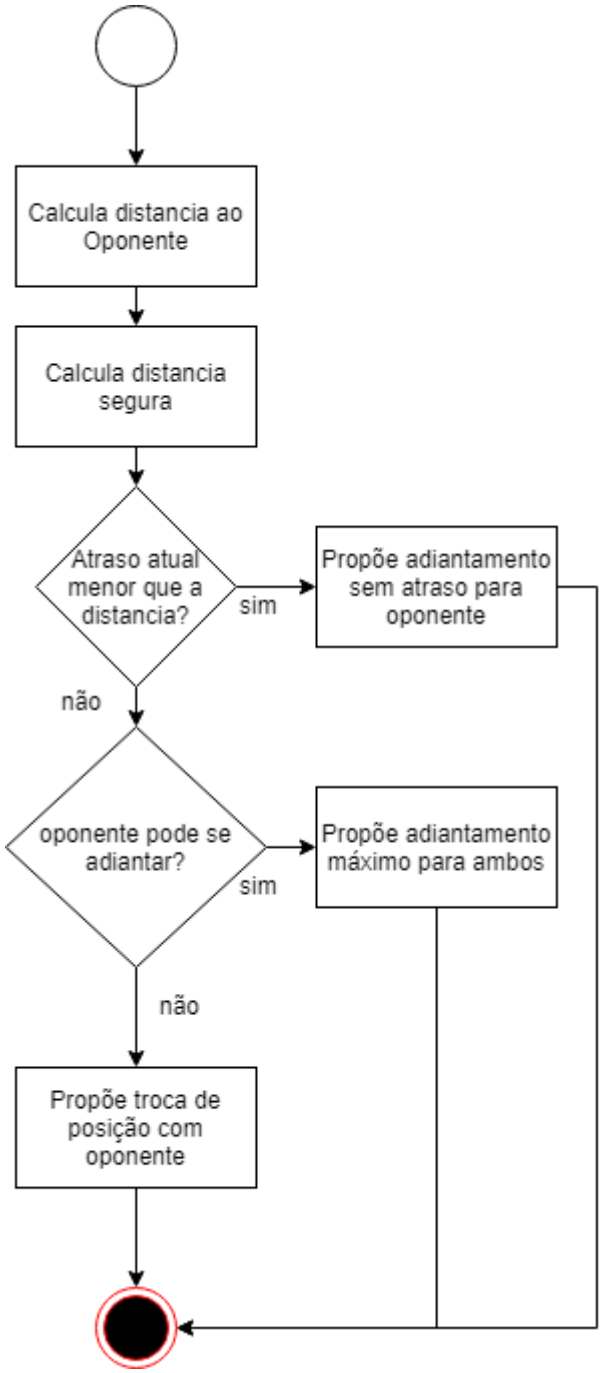

Figura 4.3: Fluxograma de concepção da melhor oferta.

O oponente então avalia a oferta. A primeira condição avaliada é se a oferta proposta diminui o custo individual de atraso. Se a oferta diminuir, o agente aceita e a negociação é finalizada. Caso a oferta aumente o custo de atraso da aeronave, a aeronave avalia se a

oferta faz com que ela se adiante ao seu horário de decolagem programado. Se a aeronave estiver se adiantando e o adiantamento for menor que o máximo, a aeronave calcula se o *payoff* de aceitar a oferta atual é maior que o *payoff* descontado do oponente aceitar uma possível contra-oferta. A aeronave opta pela ação de maior *payoff*. Caso a aeronave estiver se atrasando, a aeronave verifica se havia espaço para outra ação. Se houver, a aeronave recusa a oferta, se não, o agente verifica se o *payoff* global melhora ou piora com a ação, se melhorar ele aceita, caso contrário, ele rejeita. Este processo é ilustrado na Figura 4.4:

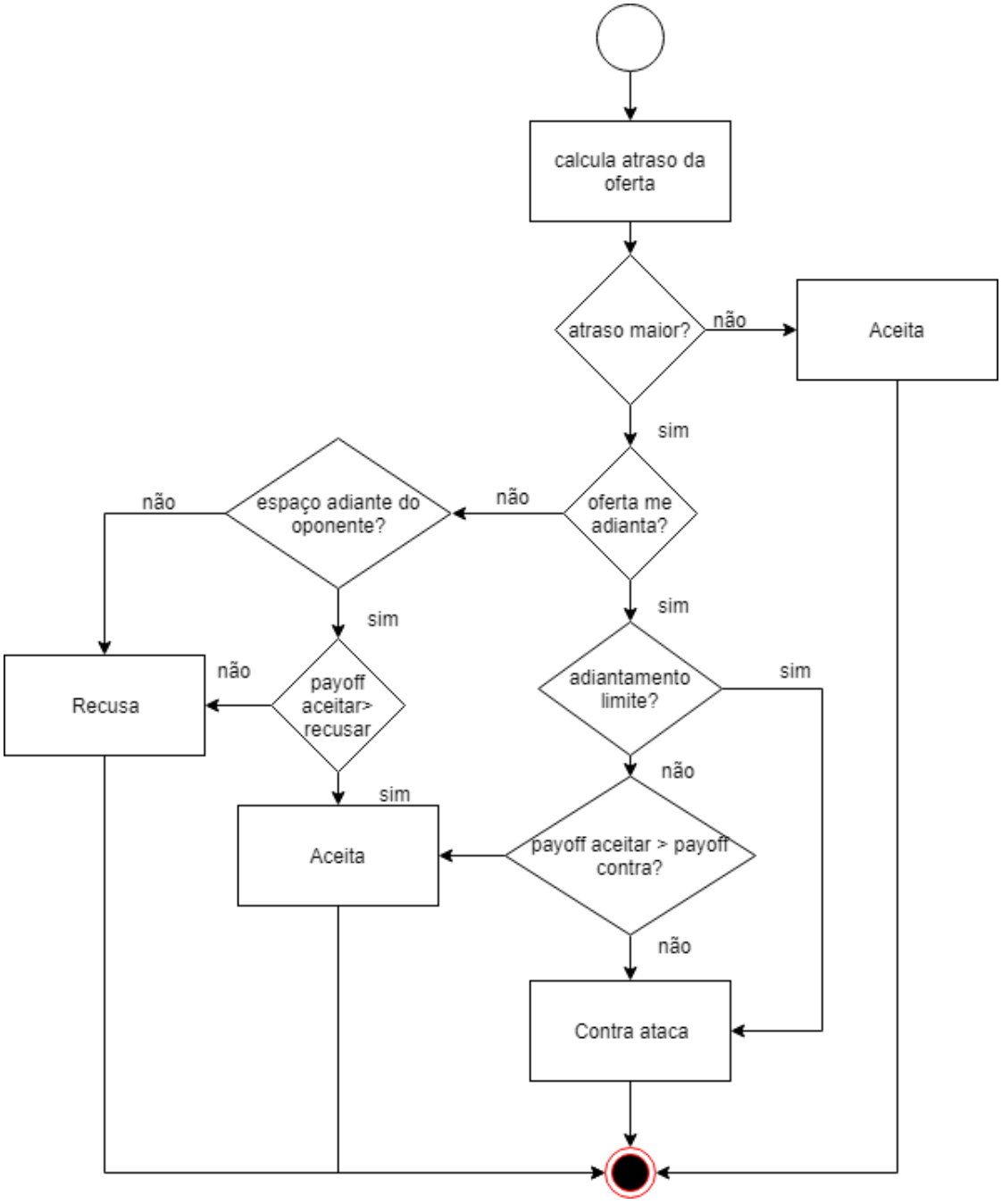

Figura 4.4: Fluxograma de avaliação da oferta.

Existem três possibilidades de contra oferta para o oponente. Se o agente estiver adiantado, ele pode realizar um atraso programado para liberar sua posição, oferecendo ela à outra aeronave. Se a aeronave já está contente com sua posição, ela verifica se exite um espaço livre adiante na fila e o oferece. Se a aeronave não está adiantada e não há espaço adiante que possa ser ofertado, o agente busca dividir igualmente o atraso imposto entre ambas as aeronaves.

S	Conjunto de estados do ambiente
A	Conjunto de ações disponíveis
T	Função de transição
R	Função de recompensa
y	Fator de desconto

Tabela 4.4: Processo de Decisão de Markov.

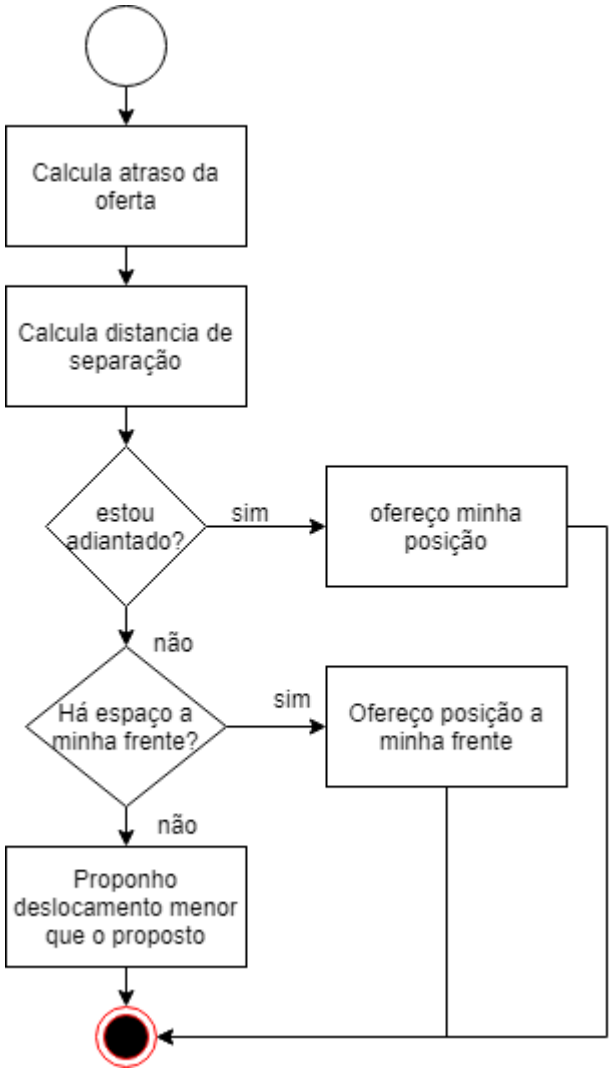

Figura 4.5: Fluxograma de concepção da contra-oferta.

4.3.3 Modelagem do Processo de Negociação como MDP

O processo de negociação entre o agente e as outras aeronave pode ser definido como um Processo de Decisão de Markov como mostrado na Tabela 4.4.

Um estados é definido pela combinação de atraso do agente, atributos da aeronave oponente, rodada de negociação e papel do agente. O atraso do agente é crucial para que

ele possa determinar se as ofertas elaboradas o aproximam de seu horário alvo ou não. Os atributos da aeronave oponente são utilizados para modelar a possível resposta do oponente. A etapa da negociação é utilizada para modelar quão perto o agente está da intervenção do árbitro. O papel do agente serve para definir quais ações são condizentes com a rodada em questão.

Um ponto a ser ressaltado é que não é necessário ao agente de aprendizagem conhecer a lógica utilizada pelas outras aeronaves na elaboração de ofertas e respostas. Uma vez que as outras aeronaves não aprendem durante a interação, o agente é capaz de modelar o comportamento destas apenas observado a resposta à suas ações.

O esforço de exploração do agente é proporcional ao número de estados possíveis para o agente. Desta forma, a maneira pela qual os atributos que constituem o estado são representados afeta o esforço computacional. Para representar o atraso do agente poderíamos armazenar o valor em minutos do atraso, no entanto isso incorreria em uma quantidade muito grande de estados. A diferença de apenas um minuto incorreria em estados completamente diferentes. A solução adotada para a representação do atraso foi definir o atraso pela distância relativa do agente ao horário alvo, reduzindo o número de valores possíveis para apenas três: atrasado, adiantado ou pontual. O mesmo se aplica ao horário da aeronave oponente.

Quanto a representação da aeronave, podemos adotar duas abordagens: representar apenas a identidade da aeronave, escondendo seus atributos, ou passar todos os atributos da aeronave. A utilização da identidade permite que o agente modele o comportamento das aeronaves sem compreender seu processo de tomada de decisão. No entanto, aeronaves com mesmo comportamento acabam por serem representadas por estados distintos.

Por outro lado, utilizando os atributos da aeronave, como tamanho, tempo de voo, velocidade , entre outros, permite ao agente modelar aeronaves cujo processo decisório seja similar em um mesmo estado. No entanto, se a variação destes atributos for muito grande, teremos a criação de muitos estados. Além disso, se os atributos não contemplarem todas as variáveis envolvidas no processo decisório do oponente, os estados gerados podem não ser suficientes.

O conjunto de ações disponíveis ao agente depende do papel desempenhado por ele na etapa de negociação em questão. Neste trabalho, utilizam-se as ações descritas por Ribeiro[2]: como proponente, o agente pode optar por ocupar o melhor espaço livre entre ele e o oponente, solicitar um adiantamento ao oponente ou solicitar que o oponente troque de posição com ele. Como oponente, o agente escolhe por aceitar, recusar ou contra-atacar a oferta recebida. A ação de contra-atacar é expandida para as 4 contra-ofertas possíveis: realizar um atraso programado e ceder sua posição atual ao proponente, oferecer o espaço adiante na fila se este estiver livre, oferecer um espaço posterior, se este estiver livre ou

adiantar-se. Cada ação de oferta tem uma recompensa associada relativa ao ganho do agente, com ofertas entre 0%, 25% e 50% do ganho. Pressupõe-se que o agente sabe como executar a ação escolhida. Isto é, não é necessário ensinar ao agente a calcular quais posições são oferecidas ao oponente ao escolher determinada ação.

A função de transição depende apenas das ações realizadas pelas aeronaves e do estado no qual aeronave se encontra. Embora as ações de aceitar e rejeitar encerrem a negociação com uma aeronave, elas não levam necessariamente o agente a um estado terminal. O resultado de uma ação, ao influenciar a posição do agente na fila, leva a possibilidade de novas negociações e novos estados.

A função de recompensa depende do resultado da negociação. Caso a oferta seja aceita, a recompensa dada ao agente é igual a diferença no custo do atraso entre a nova posição e a última. Caso o resultado seja a recusa, é dado uma recompensa negativa com valor constante. O mesmo ocorre no caso de contra-oferta. Desta forma, o agente é incentivado a agir de maneira que uma oferta vantajosa seja alcançada o mais rápido possível. O fator de desconto é utilizado para que as recompensas futuras sejam levadas em conta no processo decisório.

A política utilizada para exploração do ambiente deve ser gananciosa no limite, isto significa que com o passar das iterações do agente ele deixa de explorar e passa a escolher a ação com maior recompensa esperada. Utiliza-se uma função linear de forma que o agente explore e aja de forma gananciosa durante períodos semelhantes.

Capítulo 5

Resultados

Neste capítulo é mostrado os resultados encontrados a partir da realização do estudo de caso. Primeiramente é apresentado como foi feita a implementação do protótipo e a base de dados utilizada, para então apresentar o cenário de experimento e casos de estudo.

5.1 Implementação

A solução foi desenvolvida em linguagem Python, versão 2.7, que permite orientação a objetos e tem como um de seus tipos básico, o tipo dicionário, o que facilitou a representação dos estados de aprendizado e da função de transição. O Ambiente de Desenvolvimento utilizado foi Spyder, versão 3.1.4, presente na Distribuição da plataforma Anaconda, versão 1.6.2.

As classes do modelo se relacionam de acordo com a Figura 5.1. A classe cenário tem a função de ler a base de dados e instanciar os agentes e demais aeronaves, bem como o simulador. Para cada rodada de treinamento, uma nova instância do simulador é criada e chamada. O Simulador modela o cenário do aeroporto, alocando as aeronaves às pistas existentes e executando a negociação entre as aeronaves. A arbitragem da negociação feita pela entidade aeroportuária é modelada como um método do simulador.

As aeronaves oponentes são modeladas pela classe Aeronave e possuem comportamento predefinido. A classe Agente representa o agente em aprendizado, que aprende um comportamento a partir da negociação com as outras aeronaves.

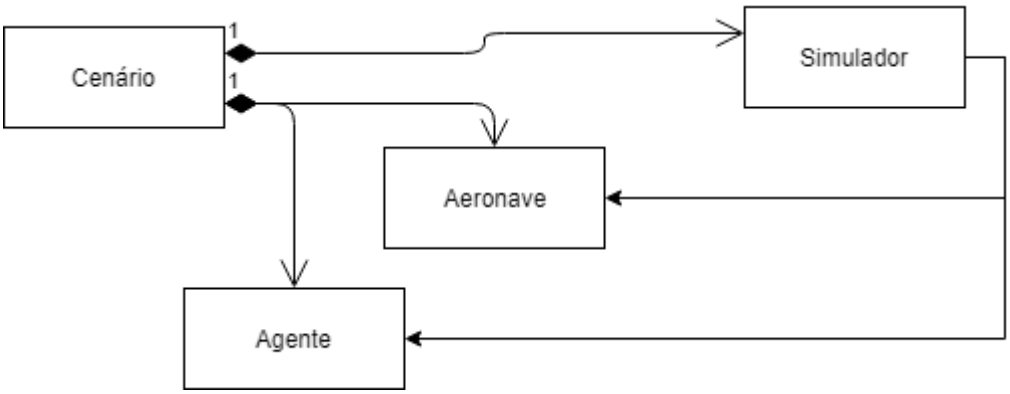

Figura 5.1: Classes da Solução.

5.1.1 Base de Dados

Para realização deste trabalho foram coletados dados referentes a planos de voos regulares disponíveis no portal do CGNA[1]. A Tabela 5.1 lista as informações presentes no documento. Um plano de voo é um documento que contém as informações sobre um determinado voo, elaborado pelo piloto da aeronave e apresentado ao órgão de gestão de tráfego aéreo. Os arquivos são disponibilizados no formato *txt* como mostrado na Figura 5.2.

Aeródromo de partida
Aeródromo de destino
Rota
Horário previsto de decolagem
Tempo previsto de viagem
Velocidade da aeronave
Modelo da aeronave
Tipo de turbina
Altura de voo
Dias de operação

Tabela 5.1: Informações presentes no plano de voo repetitivos.

[1]http://portal.cgna.gov.br/?l=pt$_br$, dados do mês de dezembro de 2018

Planos de Voo Repetitivos
Classificação: EOBT

LOC: SBBR INÍCIO DE VALIDADE: 12Jan2019

VALIDO DESDE	VALIDO ATE	DIAS OP STQQSSD	IDENT ANV	TIPO TURB	ADEP EOBT	VEL	FL	ROTA	DEST EET
181218	160219	0000060	TAM3457	A319/M	SBBR0015	N0450	340	DCT KOGDI UZ76 RCO DCT	SBRB0300
251118	160219	1030000	TAM3457	A320/M	SBBR0015	N0450	340	DCT KOGDI UZ76 RCO DCT	SBRB0300
120119	200119	1234567	GLO2070	B738/M	SBBR0045	N0462	360	DOLVI UZ24 ILVIS UM423	SBBV0311
130119	150119	1200007	GLO1750	B737/M	SBBR0050	N0450	360	UZ26 PMS UZ26 ESNER DCT	SBMA0140
180119	200119	0000567	GLO1750	B737/M	SBBR0050	N0450	360	UZ26 PMS UZ26 ESNER DCT	SBMA0140
120119	120119	0000060	GLO1750	B738/M	SBBR0050	N0450	360	UZ26 PMS UZ26 ESNER DCT	SBMA0140
160119	170119	0034000	GLO1750	B738/M	SBBR0050	N0450	360	UZ26 PMS UZ26 ESNER DCT	SBMA0140
140119	150119	1200000	GLO1706	B738/M	SBBR0055	N0440	370	MUGIS UZ35 MEBLU	SBCF0119
120119	190119	0000060	GLO1728	B738/M	SBBR0055	N0460	370	UZ2 BETAR UZ2 ENRUS UZ2	SBSL0215
180119	180119	0000500	GLO1732	B738/M	SBBR0055	N0460	370	UZ2 BETAR UZ2 ENRUS UZ2	SBSL0215

Figura 5.2: Plano de Voo Regular.

5.2 Cenário de experimento

Para realização do experimento, os dados coletados dos planos de voo são utilizados para instanciar a fila de aeronaves existentes que disputam as posições nas pistas. Criada a fila, uma aeronave é selecionada para ser o agente em treinamento para o experimento. Atrasos são aplicados aos agentes decorrentes dos conflitos por utilização das pistas e é garantida a separação segura entre as aeronaves. O processo de simulação se inicia com o agente tendo uma determinada taxa de exploração. Cada simulação envolve diversos processos de negociação do agente, que busca melhorar sua posição na fila. Em cada simulação o agente realiza ofertas à aeronave antecessora na fila, enquanto responde às ofertas de aeronaves posteriores.

Ao final de cada simulação, a fila é ordenada nas posições originais, no entanto o agente mantém as informações sobre os estados visitados. Ao fim de cada época, a taxa de exploração do agente decresce. Quando a taxa chega a zero, o agente passa a ter um comportamento ganancioso baseado nos estados aprendidos. A Figura 5.3 mostra o fluxograma do experimento.

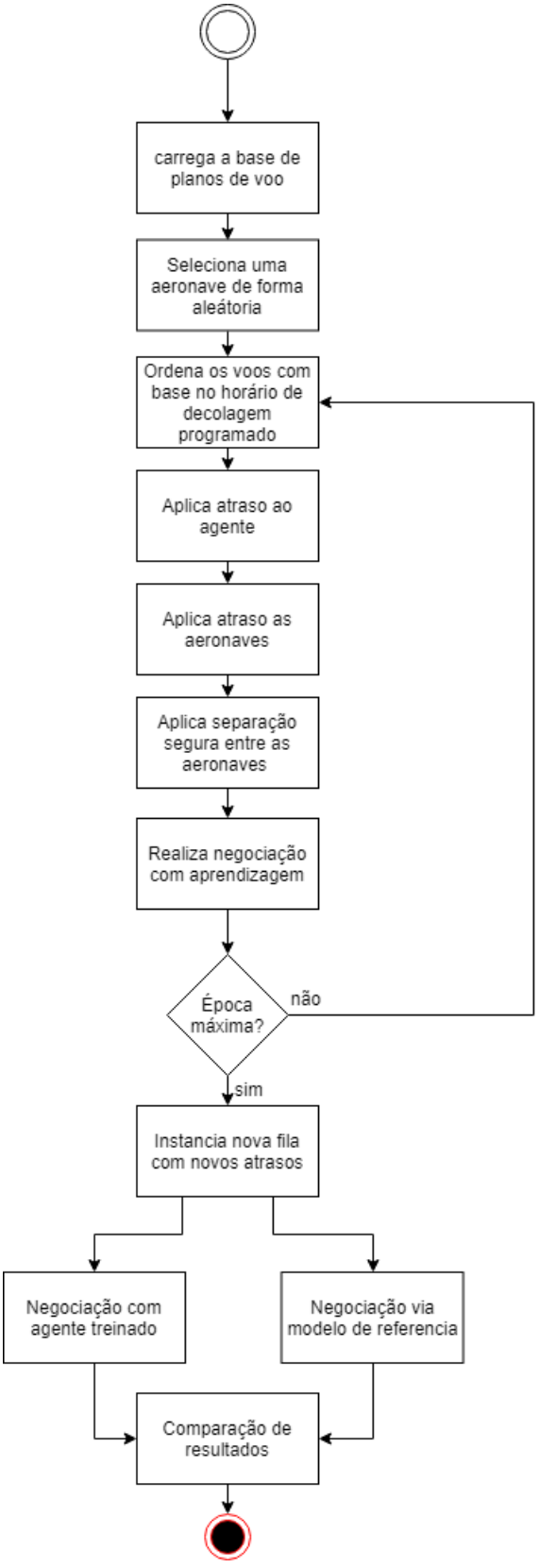

Figura 5.3: Fluxograma de execução do experimento.

Para validação do modelo foram feitos dois casos de experimento:

1. Caso 1: Alocação com Agente Homogêneos

   Para o primeiro experimento são considerados agentes com atributos semelhantes, desta forma o custo operacional depende apenas do atraso do agente. Isso limita o poder de negociação dos agentes, uma vez que a proposta associada ao deslocamento é baseada na diminuição do custo do proponente. Espera-se que o agente aprenda a se comportar como o modelo de referência.

2. Caso 2: Alocação com Agentes Heterogêneos

   Para o segundo experimento são consideradas as características individuais dos voos. O Custo operacional dos voos varia não apenas com o atraso, mas entre aeronaves. Desta forma um agente com maior custo tem maior poder de negociação, uma vez que ele pode negociar parte deste valor com a aeronave vizinha. Espera-se que o agente melhore sua posição em comparação com o modelo de referência.

5.3 Casos de Experimento

5.3.1 Alocação Homogênea

Na Tabela 5.2 são mostrados os agentes treinados no experimento, bem como seu atraso inicial e final. Os custos dependem apenas do atraso da aeronave, sendo que os outros atributos necessários ao cálculo são mantidos constantes entre todas as aeronaves. Os estados visitados representam o número de estados mapeados pelo agente durante o treinamento.

Aeronaves	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final	Estados Visitados
TAM3769	0	0	0	0	0
TAM3608	5	5	135000	337500	21
TAM3074	5	5	135000	135000	14
GLO1720	10	15	270000	405000	15
ONE6224	10	10	270000	270000	13
ONE6233	5	5	135000	135000	20

Tabela 5.2: Agentes Homogêneos Treinados.

Na Tabela 5.3 estão representados os resultados dos agentes segundo o comportamento do modelo de referência. O comportamento esperado é que os agentes não consigam melhorar a sua posição. Isso se deve ao fato de não haver oferta de horário possível que

melhore o custo do oponente. De forma análoga, espera-se que o atraso do agente não aumente e que o agente seja capaz de identificar e recusar ofertas desvantajosas.

Aeronaves	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final
TAM3769	0	0	135000	135000
TAM3608	5	5	135000	135000
TAM3074	5	5	135000	135000
GLO1720	10	10	270000	270000
ONE6224	10	10	270000	270000
ONE6233	5	5	270000	270000

Tabela 5.3: Modelo de Referência Caso 1.

Cada um destes agentes apresentam comportamentos diferentes baseados nas ações aprendidas e na posição no qual estavam inseridos. A Aeronave TAM3769 ao início do experimento já se encontrava na posição desejada, com aeronaves subsequentes na fila sem intenção de negociação, como mostrado na tabela 5.4. Desta forma, o agente não é estimulado a negociar, o que explica o número de estados visitados. Este resultado é importante por demonstrar que o agente não age de forma desnecessária.

	Aeronave	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final
Antecessor	AZU2855	0	0	0	0
Agente	TAM3769	0	0	0	0
Sucessor	ONE6300	0	0	0	0

Tabela 5.4: Aeronave TAM3769 após 25 épocas.

A aeronave TAM3608 embora não tenha aumentado seu atraso, apresentou aumento do custo ao final do experimento. Isso pode ser explicado pelas ações tomadas pelo agente durante o aprendizado. A aeronave realizou ações inúteis associadas a recompensas à outra aeronave. Desta forma, houve gastos sem ganhos nestas ações. A diferença entre os custos é de 1.5 vezes o custo inicial, o que equivale a uma combinação de ações com recompensa associada de 25% e 50%, como mostrado na Tabela 5.5. Ao aumentar o número de épocas de aprendizado de 25 para 100, o agente apresentou o comportamento esperado como mostrado na Tabela 5.6.

	Aeronave	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final
Antecessor	ONE6308	0	0	0	-202500
Agente	TAM3608	5	5	135000	337500
Sucessor	ONE6374	5	5	135000	135000

Tabela 5.5: Aeronave TAM3608 após 25 épocas.

	Aeronave	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final
Antecessor	ONE6308	0	0	0	0
Agente	TAM3608	5	5	135000	135000
Sucessor	ONE6374	5	5	135000	135000

Tabela 5.6: Aeronave TAM3608 após 100 épocas.

A aeronave GLO1720 piorou sua posição na fila ao contrário do esperado. Isso ocorre por ela ter cedido sua posição a aeronave posterior, como mostrado na Tabela 5.7. Esta ação difere da ação esperada, isso significa que o agente não explorou suficientemente o ambiente e acabou por eleger a ação errada. Novamente, ao aumentarmos o número de épocas de aprendizado, o Agente passa a adotar o comportamento esperado, como mostrado na Tabela 5.8 .

	Aeronave	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final
Antecessor	GLO1829	10	10	270000	270000
Agente	GLO1720	10	15	270000	405000
Sucessor	GLO1754	15	10	405000	270000

Tabela 5.7: Aeronave GLO1720 após 25 épocas.

	Aeronave	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final
Antecessor	GLO1829	10	10	270000	270000
Agente	GLO1720	10	10	270000	270000
Sucessor	GLO1754	15	15	405000	405000

Tabela 5.8: Aeronave GLO1720 após 100 épocas.

Os demais agentes se comportam da maneira esperada. Para os agentes treinados, não havia horários livres entre estes e as aeronaves antecessoras na fila. Além disso, o custo do atraso é igual para todos os agentes. Deste modo, o melhor comportamento possível é não realizar ofertas ao antecessor que mantenham o atraso do agente e negar as propostas feitas pelas aeronaves posteriores. Dadas as épocas suficientes de aprendizado, os agentes foram capazes de aprender este comportamento.

5.3.2 Alocação Heterogênea

Na Alocação Heterogênea, os atributos individuais das aeronaves são levados em conta no cálculo do custo da aeronave. O custo da aeronave é proporcional ao tempo de viagem, a

velocidade, a relevância da companhia aérea para a operação do aeroporto e do porte da aeronave. A Tabela 5.9 mostra os atributos das aeronaves utilizadas no experimento.

Aeronaves	Tempo de viagem (Minutos)	Velocidade(Nós)	Relevância	Porte
TAM3769	130	450	0.925	M
TAM3608	240	450	0.925	M
TAM3074	115	450	0.925	M
GLO1720	200	440	0.835	M
ONE6224	200	450	0.634	M
ONE6233	150	450	0.634	M

Tabela 5.9: Atributos Aeronaves.

Foram utilizadas 200 épocas de aprendizado com exploração decrescente de forma linear e mais 10 épocas de treinamento segundo a política gananciosa, como na Figura 5.4. Os resultados dos agentes treinados são mostrados na Tabela 5.10. A tabela 5.11 apresenta os resultados do modelo de referência.

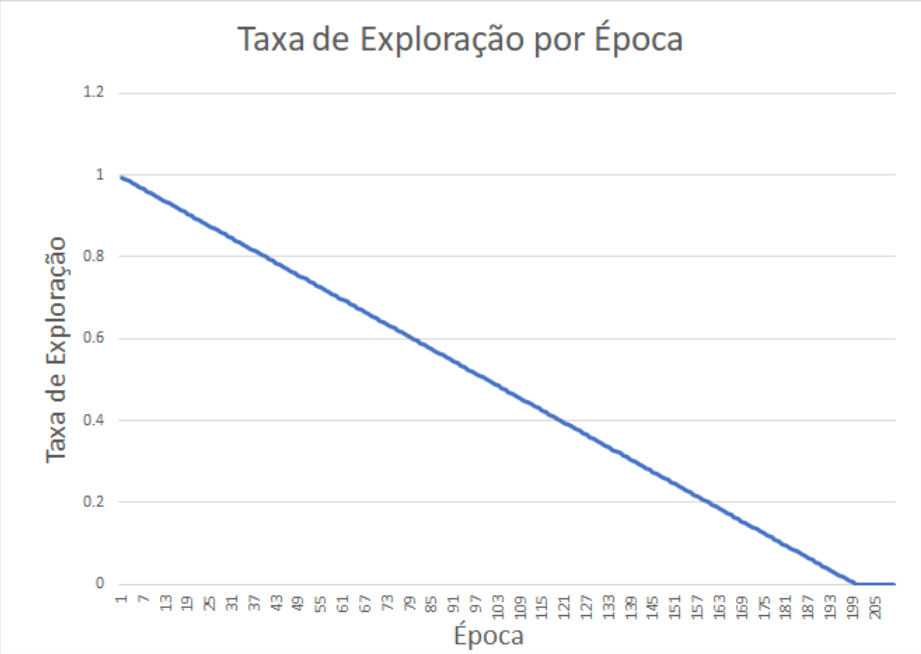

Figura 5.4: Taxa de exploração por Época.

Aeronaves	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final	Estados Visitados
TAM3769	0	0	0	0	0
TAM3608	5	5	135000	135000	20
TAM3074	5	5	216083	216083	14
GLO1720	10	5	264000	198005	12
ONE6224	10	5	579255	579255	13
ONE6233	5	5	281848	281848	20

Tabela 5.10: Agentes Heterogêneos Treinados.

Aeronaves	Atraso Inicial	Atraso Final	Custo Inicial	Custo Final
TAM3769	0	0	0	0
TAM3608	5	5	135000	135000
TAM3074	5	5	216083	216083
GLO1720	10	10	264000	264000
ONE6224	10	10	579255	579255
ONE6233	5	5	281848	281848

Tabela 5.11: Modelo de Referência Caso 2.

Dos agentes treinados, somente a Aeronave GLO1720 foi capaz de reduzir seu atraso, em 50%, como mostrado na Figura 5.5. Com a diminuição do atraso houve também diminuição do custo para a aeronave, como mostrado na Figura 5.6. No entanto, o custo não decaiu a mesma proporção, se não, reduziu-se a apenas 75% do valor original, como mostrado na Figura 5.7.

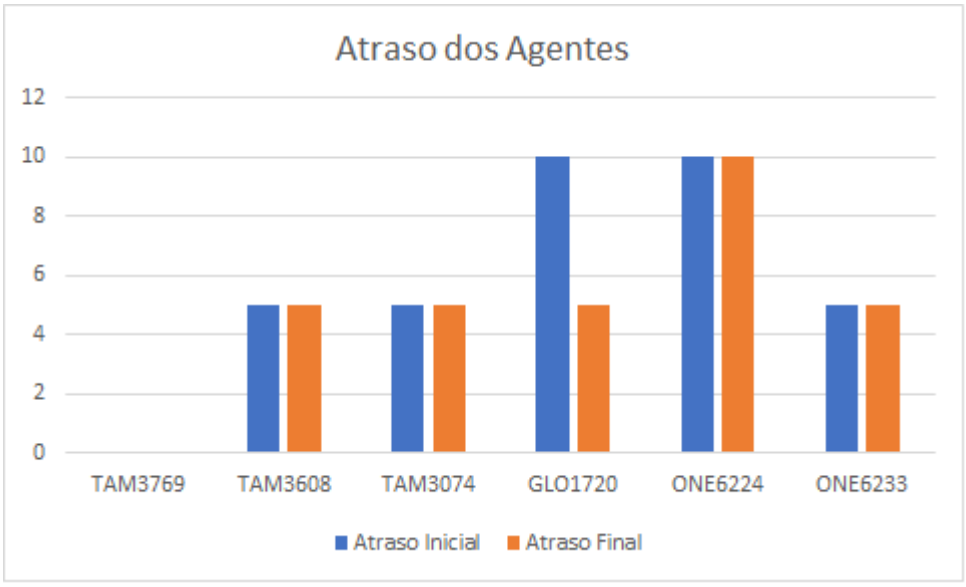

Figura 5.5: Atraso dos Agentes.

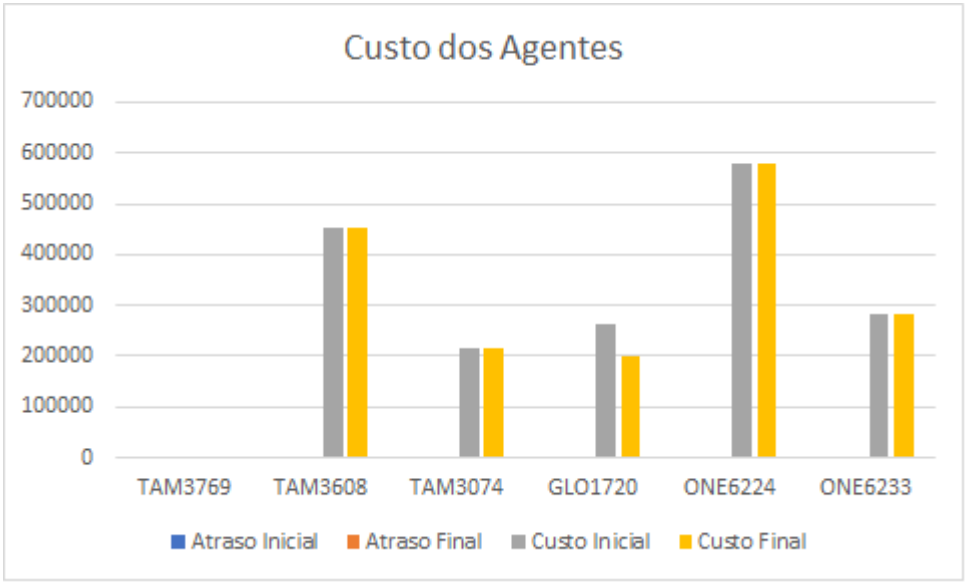

Figura 5.6: Custo dos Agentes.

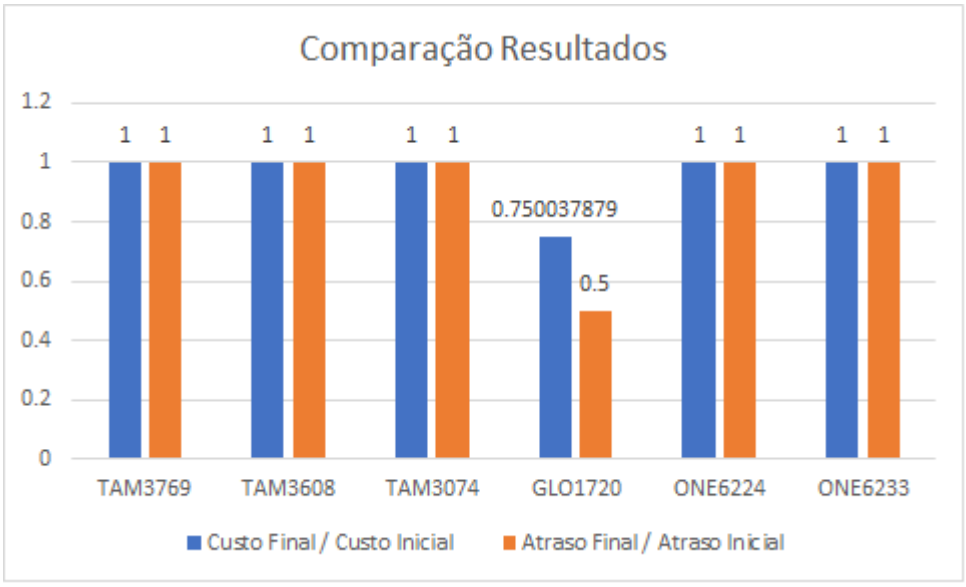

Figura 5.7: Comparação dos resultados por agente.

Podemos explicar esta discrepância através das negociações do agente com as aeronaves anteriores na fila. Ao elaborar uma oferta, o agente pode incluir uma recompensa para tornar a oferta mais atraente ao oponente. As ofertas equivalem a 25% ou 50% do valor economizado pelo agente com a adoção da oferta.

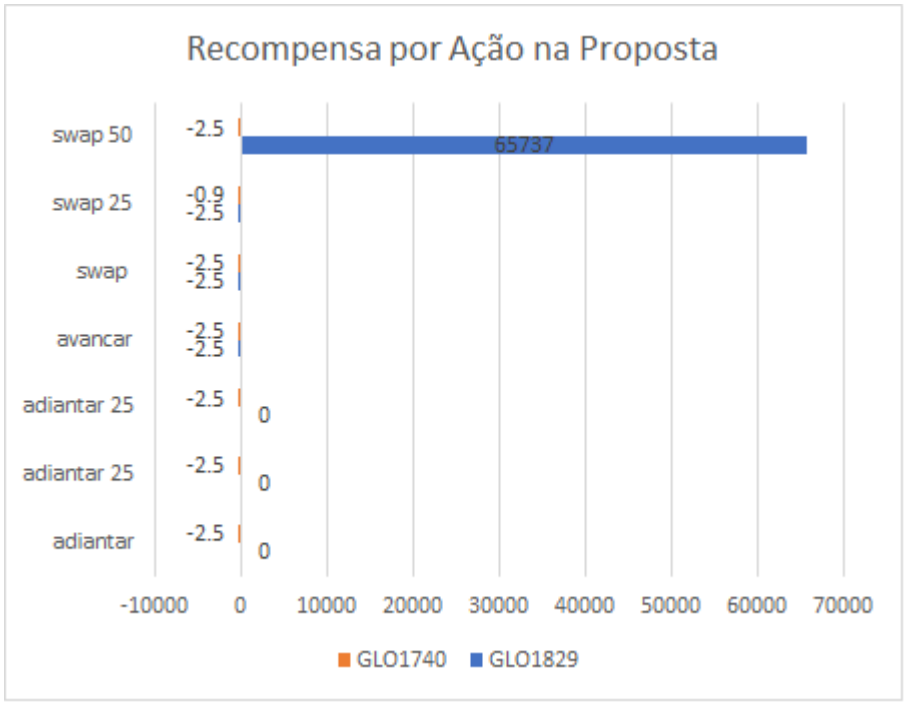

Figura 5.8: Recompensa por Ação nas Propostas aos Antecessores.

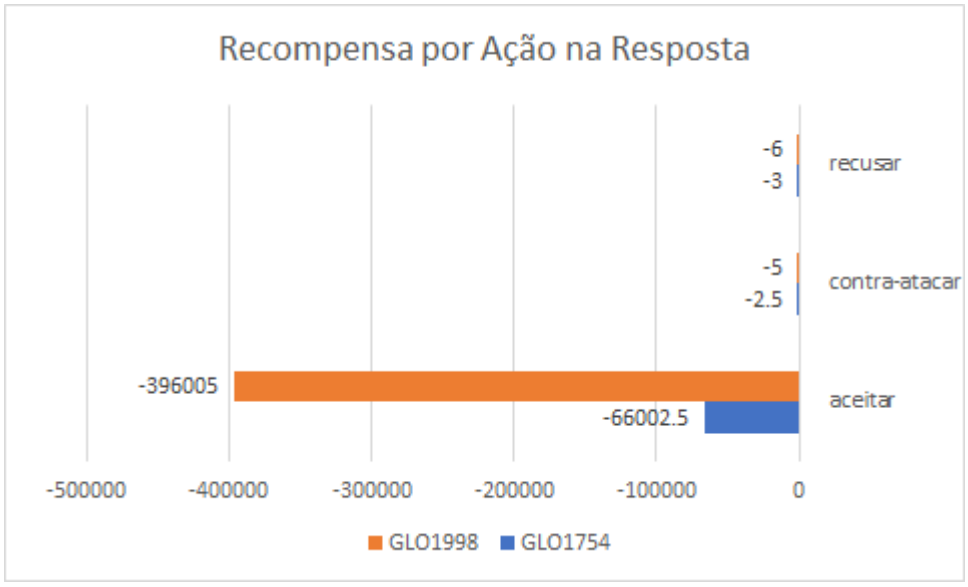

Figura 5.9: Recompensa por Ação na resposta ao Sucessores.

A partir da Figura 5.8, podemos perceber que a única oferta aceita pela aeronave anterior foi a de permuta de posições com uma recompensa associada de 50% dos ganhos. As demais ações apresentam recompensas associadas à recusa da oferta pela outra aeronave. Na Figura 5.9 podemos perceber que ao aceitar as propostas das aeronaves sucessoras na fila, a recompensa sempre é menor, sendo preferível recusar ou contra-atacar.

A partir da recompensa associada à ação de permuta com divisão dos ganhos, surge o questionamento: *"por que esta ação não teve o mesmo retorno nas ofertas realizadas pelos outros agentes?"*

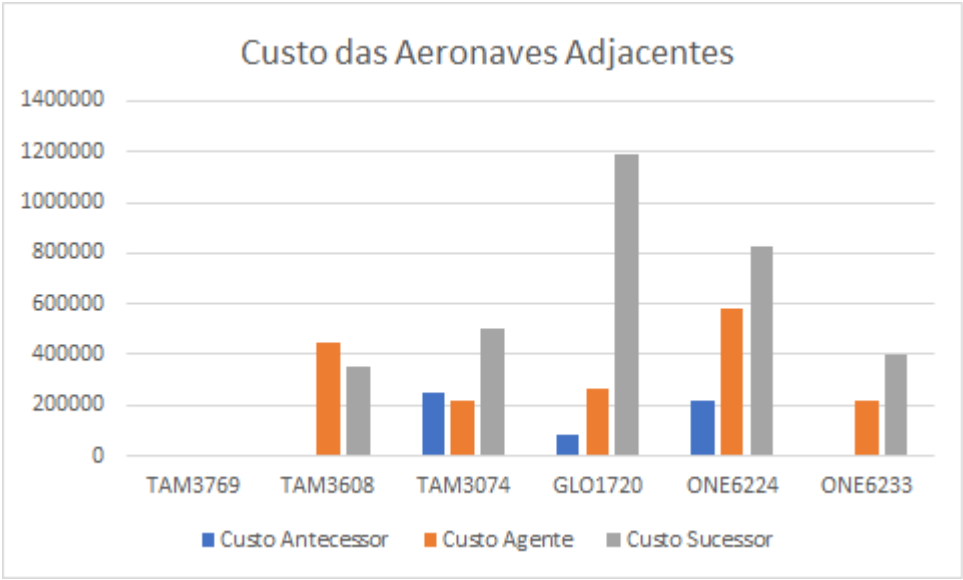

Figura 5.10: Custo das Aeronaves Adjacentes aos Agentes.

A partir dos custos dos agentes mostrados na Figura 5.10, podemos perceber que GLO1720, possui um custo maior que seu antecessor, no entanto o mesmo é verdade para as aeronaves ONE6244 e ONE6233. Para a aeronave ONE6233, seu antecessor já se encontra no horário pretendido, tendo custo zero. O fator determinante não é o custo inicial dos agentes, mas a diferença entre o custo inicial dos agentes e o custo da nova posição alcançada a partir da oferta.

A partir das Figuras 5.11 e 5.12 podemos visualizar a diferença entre a velocidade e o tempo até o destino das aeronaves e seus adjacentes. Para o agente GLO1720, tanto a velocidade do antecessor, como a distância até o destino são significativamente menores que do agente. Tal fato não é verdade para os outros agentes, com velocidades muito similares e tempos relativamente parecidos.

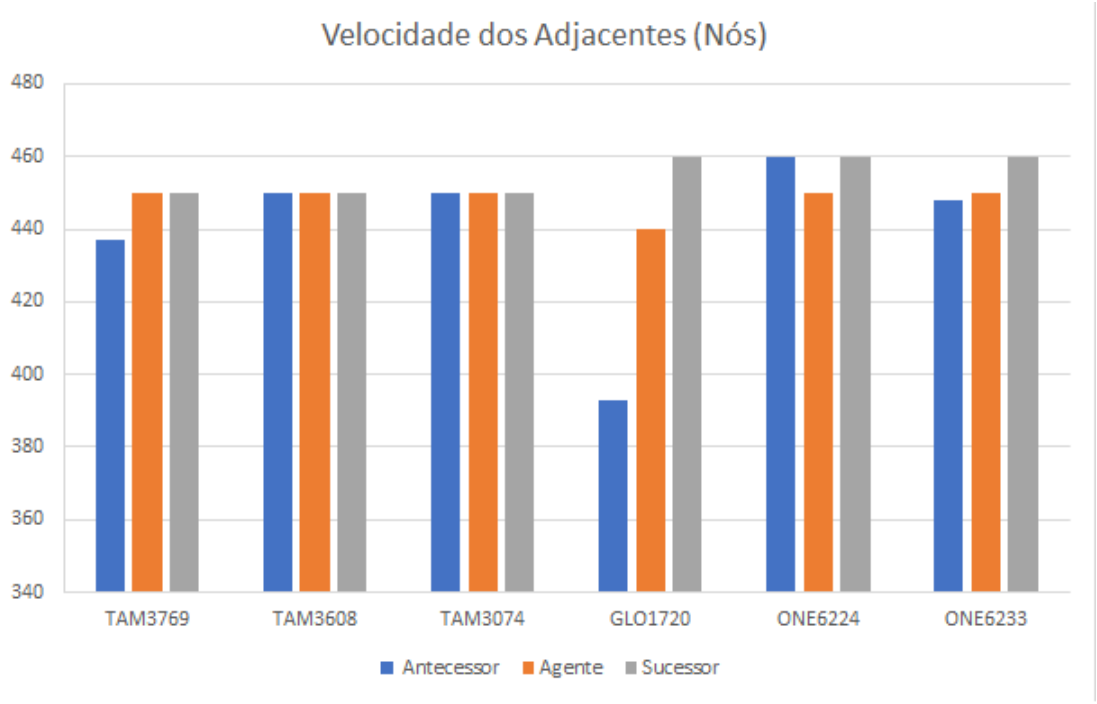

Figura 5.11: Velocidade das Aeronaves Adjacentes.

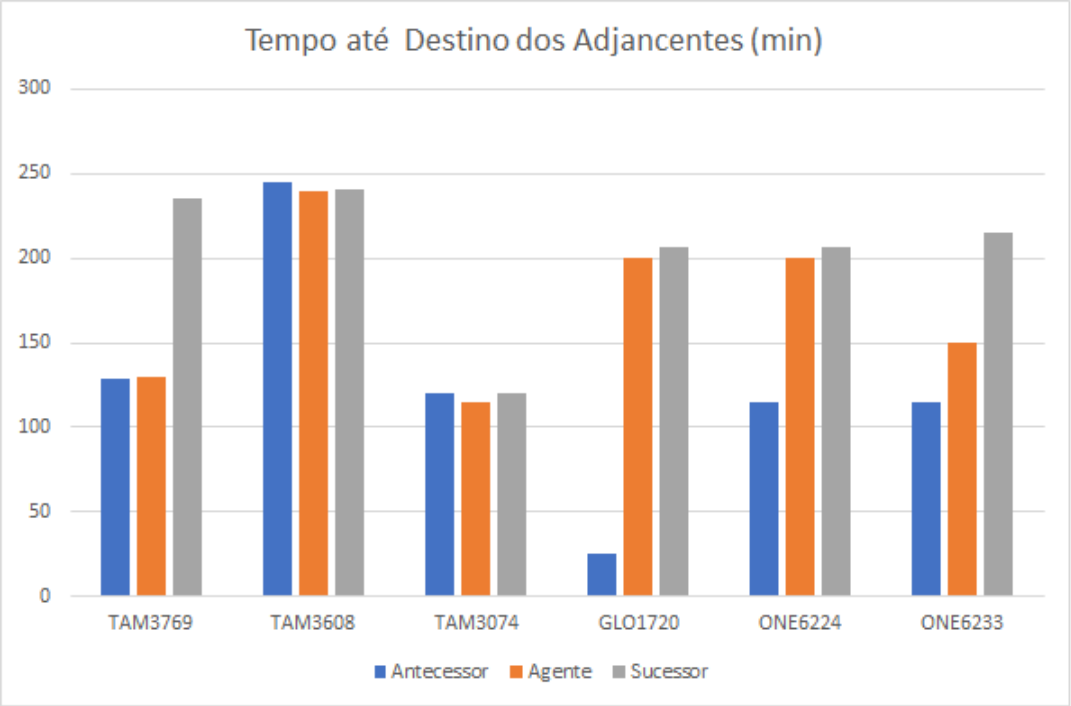

Figura 5.12: Tempo até o Destino das Aeronaves Adjacentes.

A partir dos resultados apresentados podemos concluir que agentes treinados cujos atributos são semelhantes as aeronaves adjacentes na fila, acabam por reproduzir o comportamento do modelo de referência. No entanto, para o agente que possua atributos significativamente maiores que seu antecessor, o comportamento aprendido permite que o agente melhore seu desempenho ao compartilhar seus ganhos com a outra aeronave.

Capítulo 6

Conclusão

Neste capítulo são apresentados as conclusões obtidas a partir da execução do trabalho e as possibilidades para trabalhos futuros.

6.1 Considerações Finais

Este trabalho apresentou uma abordagem por aprendizado por reforço para treinar agentes a melhorarem seu desempenho na negociação por *slots* de decolagem em aeroportos.

Os agentes foram treinados utilizando outras aeronaves oponentes cujo comportamento foi definido por um modelo de referencia. O processo de negociação do agente foi modelado como um processo de decisão de Markov, e o agente foi treinado utilizando o Algoritmo Q-learning. Como estudo de caso, foi utilizado os planos de voo regulares para o Aeroporto de Brasília e os agentes modelados de forma homogênea e heterogênea.

Para agentes homogêneos os agentes treinados tiveram comportamento similar ao modelo de referência. Isso se deu pelo fato de não haver oferta possível que melhorasse o desempenho do agente. No caso Heterogêneo, quando o custo por unidade de atraso do agente era significantemente maior que o custo da aeronave anterior na fila, o agente foi capaz de melhorar seu desempenho ao dividir o lucro da movimentação com o oponente.

Ao analisarmos as escolhas dos agentes nos estados visitados, percebemos que para cada estado, existiam ações claramente melhores. Desta forma, quando os agentes foram submetidos a um número suficiente de épocas, o comportamento convergia para escolha destas estas ações.

Quanto aos estados visitados pelos agentes, embora teoricamente o número de estados possíveis seja muito grande, o número de estados visitados foi pequeno. Isso se dá pelo fato de que embora existam muitas aeronaves no sistema, o agente acaba por negociar apenas com um subconjunto muito reduzido de aeronaves. Assim, com um número menor de oponentes, é possível que o processo de negociação seja aprofundado futuramente com

o intuito de melhorar a oferta realizada. Isso pode ser feito a partir do mapeamento de mais atributos, criação de mais ações ou mudança na forma como os atributos são representados.

A contribuição mais importante do trabalho é permitir que o agente aprenda a melhorar seu desempenho baseado na estratégia de outros jogadores. Além disso, a solução proposta flexibiliza a modelagem do problema, permitindo que os interesses individuais das entidades sejam modelados de forma independente. O modelo pode ser expandido para incorporar novas situações e novas formas de representar os interesses através de recompensas e ações possíveis para o agente.

6.2 Trabalhos Futuros

A solução realizada foi capaz de melhorar o desempenho do agente quando este tinha um custo elevado, no entanto existe possibilidade de melhoria em pontos que ficam aberto para trabalho futuro:

1. Custo Operacional das Aeronaves: A recompensa dos agentes é calculada a partir da diferença no custo operacional dos diferentes horários. Melhoria da modelagem do custo permite que o agente se adéque melhor ao ambiente.

2. Planejamento Tático: O trabalho atua sobre a etapa estratégica dos voos, que ocorre antes dos voos serem iniciados. A negociação durante a fase tática permitiria aos agentes negociar os atrasos imprevistos em tempo real.

3. Aprendizado Multiagente: Os agentes foram treinados com oponentes seguindo um comportamento predefinido. Permitir aos oponentes aprender eleva o nível de complexidade do sistema uma vez que o agente tem que se adaptar a estratégias em evolução.

4. Flexibilização das Ações: As recompensas associadas as ofertas foram definidas em relação á uma porção do ganho do agente. Flexibilizar estes valores permitiria a criação de mais ofertas possíveis, o que auxiliaria na negociação em troca de maior custo computacional devido ao aumento do número de estados.

Referências

[1] HART, SERGIU: *Games in extensive and strategic form.* Em *Handbook of Game Theory*, capítulo 2. Elsevier Science, 1992. ix, 5, 7

[2] Vitor Filincowsky Ribeiro, Li Weigang, Viorel Milea: *Collaborative decision making in departure sequencing with an adapted rubinstein protocol.* IEEE Transactions On Systems, Man, and Cybernetics: Systems, 2016. ix, 1, 2, 14, 19, 23, 28

[3] Konstantinos G. Zografos, Michael A. Madas, Konstantinos N. Androutsopoulos: *Increasing airport capacity utilisation through optimum slot scheduling: review of current developments and identification of future needs.* Journal of Scheduling, 20, 2017. 1

[4] Hamsa Balakrishnan, Bala G. Chandran: *Algorithms for scheduling runway operations under constrained position shifting.* Operations Research, 56, 2010. 1, 13

[5] Vikrant Vaze, Cynthia Barnhart: *A multi-stakeholder evaluation of strategic slot allocation schemes under airline frequency competition.* Ninth USA/Europe Air Traffic Management Research and Development Seminar, 2011. 1

[6] Richard S. Sutton, Andrew G. Barto: *Reinforcement Learning: an Introduction*, volume 2. The MIT Press, 2018. 2, 8, 9

[7] Busoniu, Lucian, Robert Babuska e Bart De Schutter: *Multi-agent reinforcement learning : An overview* . 2010. 4

[8] Theodore L. Turocy, Bernhard von Stengel: *Game theory.* CDAM Research Report, 2001. 4

[9] Karl Tuyls, Ann Nowe: *Evolutionary game theory and multi-agent reinforcement learning.* The Knowledge Engineering Review, 20:63,90, 2005. 5

[10] Ann Nowe, Peter Vrancx, Yann Michael De Hauwere: *Game theory and multi-agent reinforcement learning.* Em *Reinforcement Learning*, capítulo 14. Springer-Verlag, 2012. 5

[11] Daan Bloembergen, Karl Tuyls, Daniel Hennes Michael Kaisers: *Evolutionary dynamics of multi-agent learning: A survey.* Journal of Artificial Intelligence Research, 53, 2015. 8

[12] Stuart Russel, Peter Novig: *Inteligencia Artificial*, volume 3. Elsevier, 2013. 9

[13] Dayan, Christopher J.C.H. WatkinsPeter: *Technical note: Q-learning.* Machine Learning, 8(3-4), 1992. 10, 11

[14] Singh, Satinder, Tommi Jaakkola, Michael L. Littman e Csaba Szepesvári: *Convergence results for single-step on-policy reinforcement-learning algorithms.* Machine Learning, 38(3):287–308, Mar 2000, ISSN 1573-0565. https://doi.org/10.1023/A:1007678930559. 11

[15] Bowling, Michael H. e Manuela M. Veloso: *An analysis of stochastic game theory for multiagent reinforcement learning.* 2000. 12

[16] Husni R. Idris, Bertrand Delcaire, Ioannis Anagnostakis William D. Hall John Paul Clarke R. John Hansman Eric Feron Amedeo R. Odoni: *Observations of departure processes at logan airport to support the development of departure planning tools.* 2 nd USA/Europe Air Traffic Management RD Seminar, 1998. 13

[17] Ribeiro, Vitor Filincowsky: *Decisão colaborativa com utilização de teoria dos jogos para o sequenciamento de partidas em aeroportos*, 2013. 14

[18] Defesa, Ministério da: *Serviço de gerenciamento de fluxo de tráfego aéreo.* https://publicacoes.decea.gov.br/download.cfm?d=4838, acesso em 2018-2-11. 16, 17