UNIVERSIDADE FEDERAL DO PARÁ

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Rennan José Maia da Silva

CONTROLE DE ADMISSÃO DE CHAMADAS CONJUNTO PARA REDES HETEROGÊNEAS BASEADO EM APRENDIZAGEM POR REFORÇO

Belém, Pará
2012

Rennan José Maia da Silva

CONTROLE DE ADMISSÃO DE CHAMADAS CONJUNTO PARA REDES HETEROGÊNEAS BASEADO EM APRENDIZAGEM POR REFORÇO

Dissertação de Mestrado apresentada para a obtenção do grau de Mestre em Ciência da Computação. Programa de Pós Graduação em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.
Área de Concentração Redes de Computadores.
Orientador Prof. Dr. João Crisóstomo Weyl Albuquerque Costa.
Co-orientador Prof. Dr. Gláucio Haroldo Silva de Carvalho.

Belém, Pará
2012

Dados Internacionais de Catalogação-na-Publicação (CIP)

Sistema de Bibliotecas da UFPA

Silva, Rennan José Maia da, 1984 -

Controle de Admissão de Chamadas Conjunto para Redes Heterogêneas baseado em Aprendizagem por Reforço / Rennan José Maia da Silva. - 2012.

Orientador: João Crisóstomo Weyl Albuquerque Costa;

Coorientador: Gláucio Haroldo Silva de Carvalho

Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Ciência da Computação, Belém, 2012.

1. Redes de Computadores. 2. Redes de Computadores - Gerência. 3. Aprendizado do Computador. I. Título.

CDD 22. ed. 004.6

Rennan José Maia da Silva

CONTROLE DE ADMISSÃO DE CHAMADAS CONJUNTO PARA REDES HETEROGÊNEAS BASEADO EM APRENDIZAGEM POR REFORÇO

Dissertação de Mestrado apresentada para a obtenção do grau de Mestre em Ciência da Computação. Programa de Pós Graduação em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.

Data da aprovação: Belém-Pa. 22-08-2012.

Banca Examinadora

Prof. Dr. João Crisóstomo Weyl Albuquerque Costa
Programa de Pós Graduação em Ciência da Computação - UFPA – Orientador

Prof. Dr. Gláucio Haroldo Silva de Carvalho
Programa de Pós-Graduação em Ciência da Computação – UFPA – Co-orientador

Prof. Dr. Dionne Cavalcante Monteiro
Programa de Pós-Graduação em Ciência da Computação - UFPA – Membro Interno

Prof. Dr. Diego Lisboa Cardoso
Instituto de Tecnologia/Faculdade de Engenharia da Computação - UFPA – Membro Externo

A meus pais, que nunca mediram esforços para me dar instrução e, principalmente, educação. E me ensinaram, acima de tudo as leis de Deus e do Amor.

AGRADECIMENTOS

A Deus, que com seu amor infinito me deu o dom da vida, inteligência, saúde e tudo mais, me dando sempre o suficiente, me amparando e me resgatando sempre em todos os momentos.

A meus pais, que sempre cuidaram de mim, desde o nascimento, que tiveram paciência e que me educaram com os valores mais importantes.

A meus irmãos e irmãs, cunhados(as) e sobrinhos(as), que sempre me apoiaram e me ajudaram, me dando carinho, amor e tudo mais, sempre que precisei, e que, mesmo distantes, fazem de tudo para eu estar sempre bem.

Aos meus orientadores João Crisóstomo e Gláucio Carvalho, que mesmo sem me conhecer, acreditaram em mim e me mostraram o caminho da ciência, que tiveram paciência, compreensão e competência, sendo hoje referências profissionais para mim.

Aos demais professores do PPGCC e PPGEE, que me ajudaram no processo de amadurecimento e que me ensinaram além dos conteúdos técnicos.

Aos colegas do laboratório (LPRAD), com quem convivi desde 2007, que além de me proporcionarem boas farras, me ajudaram até onde eu não merecia.

À minha esposa Larissa, que junto comigo soube entender a distância e os desígnios de Deus para nossas vidas, que soube me compreender nos momentos difíceis, e me ajudar em todos os momentos me dando sempre amor, carinho e a sua solidariedade.

A minha família maior, tios, primos, e aos amigos Lira Maia, Alexandre Von e Adilson Pinto, que foram imprescindíveis, me ajudaram e me deram condições para concluir o curso longe de minha cidade natal.

Aos demais amigos, principalmente os de Santarém e de Belém, que cada um a sua forma, ajudaram a construir quem eu sou hoje e a realizar este grande sonho.

A todos aqueles que contribuíram para a conclusão deste trabalho.

Nunca se pode planejar o futuro pelo passado.

(Edmund Burke)

RESUMO

Atualmente, existem muitos tipos de redes sem fio baseados em diferentes tecnologias de acesso a rádio. Além dos existentes e para complementá-los, novos tipos de rede ainda serão desenvolvidos. Entretanto, nenhuma dessas tecnologias será capaz de dar aos usuários atendimento a todos os requisitos de qualidade de serviço (QoS) com cobertura universal e, por isso, a próxima geração de redes sem fio irá integrar múltiplas tecnologias, trabalhando conjuntamente de forma heterogênea. Redes heterogêneas necessitam de mecanismo de gerenciamento conjunto para garantir melhor utilização dos recursos disponíveis e dar aos usuários maior qualidade de serviço. O Controle de Admissão de Chamadas Conjunto (CACC) é um tipo de mecanismo que gerencia conjuntamente recursos em redes sem fio heterogêneas. Assim, neste trabalho é apresentada uma proposta de CACC para o gerenciamento de redes sem fio heterogêneas baseado em aprendizagem por reforço a fim de tratar as tarefas de gerenciamento de aceitação ou rejeição de chamadas e seleção inicial de tecnologia, melhorando o desempenho da rede como um todo. O algoritmo é baseado nas características da própria rede como taxa média de chegada de chamadas, tempo médio de duração das chamadas e um preço atribuído a cada classe de chamadas e os parâmetros usados para medir o desempenho foram probabilidade de bloqueio para novas chamadas e taxa de utilização da rede.

PALAVRAS-CHAVE: Controle de Admissão de Chamadas Conjunto, CAC, CACC, Alocação de Recursos, Aprendizagem por Reforço, Redes Heterogêneas.

ABSTRACT

Currently, there are many wireless networks based on different radio access technologies (RATs). Despite this, new kind of networks will be developed to complement those already existing. As there will be no RAT able to give users full service requirements with universal coverage, the next generation wireless networks will integrate multiple technologies, working jointly on a heterogeneous way. Heterogeneous networks necessitate joint radio resource management (JRRM) mechanism to enhance better resource utilization and give users better quality of service. Joint call admission controls (JCAC) are a kind of JRRM mechanisms. In this paper, we present a JCAC approach to heterogeneous wireless network management based on reinforcement learning to treat call admission and initial technology selection, enhancing the network's performance. The algorithm is based on network parameters like call arrive rate, duration time of a call and a price of class of calls. The effectiveness of this approach is assessed in terms of blocking rate and utilization rate results obtained by two simulation scenarios.

KEYWORDS: Joint Call Admission Control, JCAC, Resource Allocation, Reinforcement Learning, Heterogeneous Networks.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

3G	Terceira Geração de Telefonia Móvel
3GPP	*The Third Generation Partnership Project* - Projeto de parceria para terceira geração de tecnologia de comunicação móvel
3GPP2	*The Third Generation Partnership Project 2* - Projeto de parceria para terceira geração de tecnologia de comunicação móvel 2
AES	Padrão de criptografia avançada do protocolo 802.11
AP	Ponto de acesso
AR	Aprendizagem por Reforço
CAC	Controle de Admissão de Chamadas
CACC	Controle de Admissão de Chamadas Conjunto
CDMA	Acesso Múltiplo por Divisão de Código
DFS	Seleção dinâmica de frequência
DT	Método da Diferença Temporal
FDMA	Acesso Múltiplo por Divisão de Frequência
Gbps	Gigabits por segundo
GHz	Giga Hetz
GSM	Sistema Global para Comunicações Móveis
IA	Inteligência Artificial
IEEE	Instituto de Engenheiros Eletricistas e Eletrônicos
IMT-Advanced	*International Mobile Telecommunications-Advanced* (Telecomunicação Móvel Internacional Avançada)
IrDA	*Infrared Data Association* - Associação de dados por infravermelho
ISM	Larguras de banda de rádio para uso industrial, científico e médico
ITS	Simpósio Internacional de Telecomunicação
ITU-R	*International Telecommunication Union Radiocommunications Sector* - Setor de comunicação por rádio da União Internacional de Telecomunicações
LAN	Rede local
LTE	*Long Term Evolution* - 4ª Geração de Redes Móveis Sem Fio baseado em 3GPP
MAC	Camada de Controle de Acesso ao Meio
Mbps	Megabits por segundo
MC	Método de Monte Carlo
Mesh	Redes em malha sem fio
MIH	*Handover* independente de tecnologia

MIMO	*Multiple-input multiple-output* (múltiplas entradas múltiplas saídas)
MIPv6	Protocolo de Internet Móvel versão 6
OFDM	Multiplexação por divisão de frequência ortogonal
PD	Método de Programação Dinâmica
PHY	Camada Física
PMD	Processo Markoviano de Decisão
PSMD	Processo Semi Markoviano de Decisão
QoS	Qualidade de Serviço
RAT	Tecnologia de Acesso a Rádio
RH	Rede Heterogênea
RHSF	Rede Heterogêneas Sem Fio
RRM	Gerenciamento de Recurso de Rádio
SA	Método de CAC baseado em Seleção Aleatória de tecnologia
SBrT	Simpósio Brasileiro de Telecomunicações
TDMA	Acesso Múltiplo por Divisão de Tempo
TPC	Controle de Potência de Transmissão
ulb	Unidade de Largura de Banda
UMTS	Sistema de Telecomunicação Móvel Universal
VoIP	Voz sobre Protocolo de Internet
WAVE	Acesso sem fio em ambiente veicular
WiFi	Rede local sem fio baseada no padrão IEEE 802.11
WiMax	*Worldwide Interoperability for Microwave Access* - Interoperabilidade Mundial para Acesso de Micro-ondas)
WLAN	Rede local sem fio
WMAN	Rede metropolitana sem fio
WPAN	Rede de área pessoal sem fio
WRAN	Rede sem fio de área regional
WWAN	Rede geograficamente distribuída sem fio

SUMÁRIO

1. Introdução

1.1 Motivação

O sucesso da próxima geração de redes consistirá de uma combinação de heterogeneidade de terminal e heterogeneidade de rede. Heterogeneidade de terminal refere-se a diferentes tipos de terminais em termos de número de redes suportadas (modalidade), tamanho de tela, consumo de energia, portabilidade/peso, complexidade, etc. Por outro lado, a heterogeneidade da rede refere-se à combinação de múltiplas redes sem fio baseadas em diferentes tecnologias de acesso e coexistentes na mesma área geográfica. Esta coexistência, requer gerenciamento conjunto de recursos para atingir melhor qualidade de serviço e maior eficiência na utilização de recursos de rádio (Falowo & Chan, 2010).

Falowo & Chan (2008) afirmam também em seu trabalho que gerenciar conjuntamente os recursos de rede traz benefícios como, por exemplo, dar maior estabilidade à rede, aumento da utilização de recursos bem como maior satisfação dos usuários. Um dos mecanismos de gerenciamento em rede sem fio de suma importância é o Controle de Admissão de Chamadas (CAC). O CAC é uma política que define as regras para aceitação ou rejeição de chamadas. O objetivo é sempre encontrar uma política de CAC que maximize o rendimento/utilidade e, ao mesmo tempo, garanta parâmetros de QoS (*Quality of Service* – Qualidade de Serviço). Em termos simples, a rede quer uma política que aceita tantas chamadas quanto possível, mas nem tantas a ponto da rede ficar congestionada e os requisitos de QoS sejam violados.

Mecanismos de CAC projetados para redes homogêneas não apresentam bom desempenho em redes heterogêneas, fazendo que seja desenvolvida uma nova área de pesquisa que estuda mecanismos de CAC Conjunto (CACC), que gerenciam conjuntamente redes heterogêneas. Diversos mecanismos de CACC têm sido propostos na literatura, mas nenhum deles apresenta uma solução definitiva para o assunto. Um dos tipos de CACC

apresentado na literatura é o baseado em Inteligência Computacional, que se utiliza de algoritmos de Inteligência Artificial para melhorar o desempenho e a utilização de recursos de rede sem fio, dar maior estabilidade e garantir melhor rendimento por parte do operador de rede.

Nesse contexto, maximizar rendimento enquanto garante requisitos de QoS sugere o uso de Processos Semi-Markovianos de Decisão (PSMD). Entretanto, o rápido crescimento do número de estados, de acordo com a complexidade do problema, tem levado à utilização de algoritmo de aprendizagem por reforço (AR) (Mignanti *et. al.*, 2009). AR tem sido aplicada a uma variedade de esquemas como roteamento, gerenciamento de recursos e alocação dinâmica de canal em redes sem fio, especialmente em redes móveis *ad hoc*, redes de sensores sem fio, redes celulares e, recentemente, na próxima geração de redes sem fio, como redes de rádio cognitivo (Yau *et. al.*, 2012), isso porque com AR, têm-se a possibilidade de trabalhar com espaço de estados maior que em Processo Markoviano ou Semi-Markoviano de Decisão, sem inviabilizar o desempenho do sistema. As métricas de melhorias podem ser medidas em termos de vazão, perda de pacotes, consumo de energia, atraso fim-a-fim, probabilidade bloqueio de chamadas, descarte de chamadas de *handoff*, utilização dos recursos de rede e, consequentemente, rendimento do operador de rede, troca de canais de rede, número de reassociações de canal, etc. Por esse motivo, resolveu-se estudar o problema de CACC para redes sem fio heterogêneas associando o método de solução à técnica de aprendizagem por reforço, que já vem sendo utilizada em diversas soluções em redes sem fio.

1.2 Objetivos

Neste trabalho, objetiva-se desenvolver a modelagem de um algoritmo de CACC baseado em aprendizagem por reforço, utilizando-se de parâmetros da própria rede como: taxa média de chegada de chamadas; tempo médio de duração das chamadas; preço de cada classe de chamadas; visando melhorar o desempenho da rede em termos de probabilidade de bloqueio para novas chamadas, bem como aumentar a utilização dos recursos de rede. Esse algoritmo é apresentado nos Capítulos 4 e os resultados de simulação obtidos são explicados no Capítulo 5.

1.3 Contribuição

A principal contribuição deste trabalho foi o desenvolvimento de um algoritmo para

redes heterogêneas que se utiliza da técnica de aprendizagem por reforço para resolução do problema de CACC. Esta abordagem é nova no sentido de se usar AR associado a parâmetros da própria rede. A solução de CACC baseado em IA utilizando-se de AR já foi abordada em outras propostas, mas nenhuma com as funções e parâmetros definidos neste trabalho, que permitem que o algoritmo utilize as características da rede e, a partir daí, gere uma política de controle de admissão. O fato de a solução aqui proposta ser baseada em modelo, não atendo-se a uma tecnologia de transmissão de dados específica, dá margem também para que o algoritmo seja aplicado em quaisquer tecnologias, aumentando a área de abrangência e permitindo sua aplicação universal. Além disso, uma vantagem de se usar AR é que o sistema pode continuar aprendendo mesmo estando em modo de operação.

1.4 Publicação

Os resultados deste trabalho foram publicados nos anais do 7th IEEE/SBrT *International Symposium of Telecommunications* (ITS'10) realizado no ano de 2010 na cidade de Manaus, Brasil sob o título de "*A Reinforcement Learning Based Joint Call Admission Control for Heterogeneous Wireless Networks*".

1.5 Organização da dissertação

Este trabalho está organizado da seguinte forma: no Capítulo 2 é mostrada uma visão geral de redes sem fio, seus principais padrões e tecnologias homogêneas, uma introdução sobre sua operação integrada de modo heterogêneo bem como uma breve descrição da área de gerenciamento de recursos e CACC; no Capítulo 3 é apresentado o algoritmo de aprendizagem por reforço e suas características; no Capítulo 4 é feita a descrição da modelagem do algoritmo proposto neste trabalho; Os cenários de simulação para validar a proposta, bem como seus resultados, serão explicados no Capítulo 5 e, por fim, o Capítulo 6 traz as considerações finais e trabalhos futuros a serem desenvolvidos para aperfeiçoamento do mesmo.

Capítulo 2 - Redes sem fio homogêneas, redes heterogêneas e gerenciamento de recursos

2.1 Introdução

É comum, nos dias atuais, a existência de aparelhos como portões eletrônicos com controle remoto, aparelhos de rádio e televisão, dentre outros tipos de aparelhos que se utilizam de ondas eletromagnéticas para transmissão e recepção de dados. Cada um desses utiliza algum tipo de tecnologia de comunicação sem fio, que implementa um padrão específico de rede, e que possui características distintas. Dentro dessa diversidade, cada uma dessas tecnologias evolui, apresentando melhorias em termos de velocidade de transmissão, cobertura, confiabilidade, novas funcionalidades, etc. a ponto de serem, atualmente, quase que onipresentes e de ser comum ambientes onde mais de uma tecnologia de comunicação sem fio esteja disponível, bem como aparelhos que suportem mais de um protocolo de comunicação sem fio. Um exemplo disso é a existência de telefones celulares que suportam também redes WiFi e Bluetooth, além da tecnologia de telefonia móvel. Nas próximas subseções serão apresentados os principais padrões de redes sem fio, as tecnologias mais utilizadas e suas características. É feito também um resumo de trabalhos de CACC, explicando seus benefícios e características.

2.2 Tipos de Redes sem fio

Existem diversas formas de classificar redes de computadores, dependendo do parâmetro a ser levado em conta. Uma delas é a classificação baseada na região geográfica de abrangência da rede. Tanenbaum (2003) utiliza este critério para classificar as redes de computadores de modo geral em redes de área pessoal, redes locais, metropolitanas, geograficamente distribuídas e a Internet.

Na Tabela 2.1 será apresentado um resumo dos principais tipos de tecnologia de comunicação sem fio que serão detalhados nas próximas subseções. Esta tabela é uma extensão da classificação feita por Tanenbaum (2003), especificamente tratando de redes sem fio e acrescentando-se apenas as redes de área regional, que se utilizam do padrão IEEE 802.22 (IEEE 802.22 Standard, 2011).

Tabela 2.1: Tipos de redes sem fio (Rackley, 2007)

Tipo de Rede	Cobertura	Função	Padrões mais populares
WPAN	Espaço operacional pessoal; normalmente até 10 metros	Tecnologia de substituição de cabeamento; redes pessoais	IrDA, Bluetooth, ZigBee, IEEE 802.15
WLAN	Prédios ou campus; normalmente até 100 metros	Extensão ou alternativa para redes cabeadas	WiFi, IEEE 802.11a, b, g
WMAN	(Alguns quilômetros) Cobertura de um bairro ou uma cidade	Extensão de redes locais, interconexão de rede, redes de acesso a banda larga	WiMax IEEE 802.16
WWAN	Nacional através de vários fornecedores, ou até internacional	Extensão de rede local, interconexão de redes, redes de acesso a banda larga, redes de telefonia móvel	GSM, TDMA, CDMA, EVDO, GPRS, EDGE, WCDMA, UMTS, LTE IEEE 802.20
WRAN	De 30 a 100 km de extensão	Oferta de conexão de rede a áreas remotas onde possuam canais de televisão disponíveis como em áreas rurais	IEEE 802.22

Em seguida, nas subseções seguintes, será mostrado um detalhamento de alguns dos principais tipos e tecnologias de comunicação sem fio, conforme classificação apresentada anteriormente.

2.2.1 WLAN (*Wireless Local Area Networks*) – Redes locais sem fio

Presente no local de trabalho, em casa, em instituições educacionais, em cafés, aeroportos e esquinas, as LANs agora são uma das mais importantes tecnologias de rede de acesso na Internet de hoje (Kurose & Ross, 2006). Embora muitas tecnologias e padrões para LANs tenham sido desenvolvidos, uma classe particular surgiu claramente como vencedora: a LAN sem fio padrão IEEE 802.11, também conhecida como WiFi (Kurose & Ross, 2006).

As redes locais sem fio começaram a ser desenvolvidas no final da década de 1980 seguindo a abertura de bandas de freqüência ISM (*industrial, scientific and medical* – banda para utilização em aplicações industriais, científicas e médicas) para uso não licenciado em 1985 e atingiu seu marco em 1997 com a aprovação e publicação do padrão IEEE 802.11. Este padrão, que inicialmente, especificou modesta taxa de dados (de 1 a 2 Mbps), tem sido

melhorado com o tempo, e as principais revisões são conhecidas pela adição de uma letra no sufixo original IEEE 802.11 como, por exemplo, a, b e g. Um resumo geral pode ser visto na Tabela 2.2, como segue.

Tabela 2.2 – Grupo de padrões IEEE 802.11 (Rackley, 2007)

Padrão	Principais características
802.11a	Padrão WLAN de alta velocidade, suportando até 54 Mbps de taxa de dados usando modulação OFDM em banda de freqüência ISM de 5 GHz
802.11b	O padrão WiFi original, provendo 11 Mbps em espectro de frequência ISM de 2.4 GHz
802.11d	Habilita configuração de freqüências permitidas, nível de potência e largura de banda de sinal em nível de camada MAC para obedecer a regras de rádio frequência locais, mas facilitando operação internacional.
802.11e	Adiciona requisitos de QoS a todas as interfaces de radio 802.11, provendo TDMA, permitindo priorizar e corrigir de erros, melhorando o desempenho de aplicações sensíveis a atraso.
802.11f	Define práticas recomendadas de equipamentos de WLAN de tal forma que os Pontos de Acesso (APs) possam interoperar. Define o protocolo IAPP (*Inter-Access Point Protocol*).
802.11g	Provê melhoria na largura de banda para 54 Mbps usando modulação OFDM na banda de freqüência de 2.4 GHz. Interoperável no mesmo equipamento de rede que o equipamento 802.11b.
802.11h	Versão do protocolo 802.11a que vai de encontro a algumas regulamentações para utilização da banda de freqüência de 5 GHz na Europa. O padrão conta com dois mecanismos que otimizam a transmissão via rádio: a tecnologia DFS (seleção dinâmica de frequência) e a tecnologia TPC (controle de potência de transmissão).
802.11i	Visa aperfeiçoar as funções de segurança como autenticação de usuário e protocolos de criptografia. O padrão emprega padrão de criptografia avançado (AES) e autenticação 802.1x.
802.11j	Diz respeito a bandas que operam as faixas de freqüência de 4.9 GHz a 5 GHz, disponíveis no Japão.
802.11k	Possibilita um meio de acesso para os APs transmitirem dados de gerenciamento. Especifica otimização de desempenho de rede através da seleção de canal, *roaming* e TPC.
802.11n	Provê maior largura de banda de 150, 350 e até 600 Mbps usando a tecnologia de rádio MIMO, canais de radiofreqüência maior e melhorias na pilha de protocolos, enquanto mantém compatibilidade com as versões 802.11 a, b e g.
802.11p	Implementação de redes sem fio para ambientes veiculares (WAVE – *Wireless Access in Vehicular Environments*)
802.11r	Padroniza o *handoff* rápido quando um cliente wireless se reassocia quando estiver se locomovendo de um ponto de acesso a outro na mesma rede, para suportar serviços sensíveis a atraso como VoIP.
802.11s	Padroniza "*self-healing/self-configuring*" para redes em malha (*Mesh*).
802.11t	Provêem práticas recomendadas em métodos de medidas, métricas de desempenho e procedimentos de testes para atingir melhor desempenho em equipamentos e redes.
802.11u	Alterações para as camadas PHY (física) e MAC (Controle de Acesso ao

	Meio) para prover uma abordagem padronizada e genérica para interoperação com redes não 802.11, como Bluetooth, ZigBee e WiMax.
802.11v	Provê melhorias na vazão, reduz a interferência e aumenta a confiabilidade através do gerenciamento de rede.
802.11w	Aumenta a segurança, estendendo segurança da transmissão dos pacotes de camada física.

Mais detalhes sobre o padrão IEEE 802.11 e as redes locais sem fio podem ser encontrados em (Kurose & Ross, 2006) e (Rackley, 2007).

2.2.2 WPAN (*Wireless Personal Area Networks)* – Redes de área pessoal sem fio

As redes de área pessoal sem fio são redes que fazem interconexão de dispositivos de uso pessoal em curto espaço de operação – geralmente na faixa de 1 a 10 metros. Seu objetivo é dar maior flexibilidade, mobilidade e evitar a inconveniência do uso de cabos (Rackley, 2007). O padrão IEEE 802.15, como também é conhecido, é essencialmente uma tecnologia de 'substituição de cabo', de baixa potencia, curto alcance e baixa velocidade para interconectar notebooks, equipamentos periféricos, telefones celulares e PDAs (Kurose & Ross, 2006).

Se livrar da inconveniência e da limitação dos cabos foi a maior motivação por trás de numerosos grupos de trabalho e outras organizações envolvidas, desde a década de 1990 em desenvolver um padrão para redes de área pessoal sem fio (Rackley, 2007). Inicialmente o Bluetooth foi lançado por meio de uma cooperação de empresas e, em 1999 o IEEE estabeleceu o grupo de trabalho 802.15 para prover um padrão para suportar interoperabilidade de baixa complexidade, baixo consumo de energia, mantendo comunicação em espaço pessoal para equipamentos estacionários ou móveis. Uma variedade de padrões WPAN tem sido desenvolvido desde o final da década de 1990, dentre eles os mais notáveis foram Bluetooth, IrDA, e mais recentemente ZigBee e Wireless USB (Rackley, 2007), cada uma dessas tecnologias tem seus pontos fortes e fracos e na Tabela 2.3 será apresentado um resumo dos padrões lançados.

Tabela 2.3 - Resumo do padrão IEEE 802.15 Padrão WPAN e seus Grupos de Trabalho (Rackley, 2007)

Padrão	Descrição	Aplicação
802.15.1	Especificação original FHSS que opera na banda de frequência 2.4 GHz. Publicada em 2002	Bluetooth
802.15.2	Práticas recomendadas para facilitar a coexistência entre as redes sem fio WPAN 802.15 e redes WLAN	

	802.11. Publicado em 2003.	
802.15.3a	WPAN a altas taxas. UWB PHY com DS-UWB x OFDM sob discussão. Rascunho publicado em 2003. Ultrapassado pelo MBOA e Wireless USB. Grupo de trabalho dissolvido em janeiro de 2006	
802.15.3b	Melhorias no Grupo de Trabalho da camada MAC, na parte de implementação e interoperabilidade.	
802.15.3c	Alternativa baseada em milímetro de onda nas bandas de frequência 57-64 GHz. Taxa de dados de 1 Gbps e opcionalmente até 2 Gbps. Formado em março de 2005.	
802.15.4	Padrão que especifica a camada física e MAC para WPAN de baixa taxa. Opera em 2.4 GHz, 915 e 868 MHz. Publicado em 2003	ZigBee
802.15.4a	Destinado a desenvolver uma camada física alternativa.	
802.15.4b	Grupo caracterizado para prover melhorias e clarividência do padrão 802.15.4	
802.15.5	Mecanismo para desenvolvimento das camadas MAC e PHY requerido para habilitar redes PAN em malha.	

Para mais detalhes a respeito da tecnologia Bluetooth e do padrão IEEE 802.15, consultar (Held, 2001) e (Bisdikian, 2001) *apud*. Kurose & Ross (2006).

2.2.3 WMAN (*Wireless Metropolitan Area Networks)* – Redes Metropolitanas Sem Fio

As redes metropolitanas sem fio são redes que abrangem uma grande área como um bairro ou até uma cidade inteira, com extensão de alguns quilômetros. São utilizadas, na prática, entre provedores de acesso, seus pontos de distribuição ou para interligação de redes locais. A tecnologia mais difundida nesse tipo de rede é a WiMax, representada pelo padrão IEEE 802.16, que foi lançado, inicialmente, em 2001 e desde então apresentou uma série de versões de evolução, conforme apresentado na Tabela 2.4.

Tabela 2.4 – Resumo do padrão IEEE 802.16

Padrão	Descrição
80216-2001	Acesso a banda larga fixa sem fio (10-66 GHz)
802.16.2-2001	Práticas recomendadas para coexistência
802.16c-2002	Melhorias incluindo perfis para 10-66 GHz
802.16a-2003	Definições de camada física e camada MAC para 2-11 GHz
802.16-2004	Revisão, incorporação do padrão obsoleto IEEE 802.16-2001 e suas duas melhorias.
802.16.2-2004	Revisão, incluindo expansão para 2-66 GHz
802.16f-2005	Base de informações e gerenciamento para sistemas fixos

802.16-2004/Cor 1-2005	Correções para sistemas fixos (co-publicado com 802.16e-2005)
802.16e-2005	Sistema de acesso a banda larga móvel sem fio
802.16k-2007	Melhorias para pontes em 802.16 (uma correção para IEEE 802.1D)
802.16g-2007	Melhorias na base de informações e gerenciamento móvel. Procedimentos de gestão de planos e serviços
802.16j-2009	Melhorias em *Mobile Multihop Relay*
802.16h-2010	Melhorias no mecanismo de coexistência para operação livre de licença
802.16m-2011	Interface aérea avançada com taxa de dados de 100 Mbps móvel e 1 Gbps fixo. Também conhecido como WiMax móvel versão 2 ou WirelessMAN-Advanced. Objetivando preencher os requisitos do ITU-R IMT-Advanced para sistemas 4G.

Mais informações sobre as redes WMAN e o padrão IEEE 802.16, suas especificações e características podem ser encontrados em IEEE 802.16.2 Standard (2004) ou em Andrews (2007).

2.2.4 WWAN – (*Wireless Wide Area Networks*) Redes geograficamente distribuídas sem fio

As redes geograficamente distribuídas sem fio são redes com grande abrangência geográfica, distribuídas regionalmente, nacionalmente ou até globalmente, voltadas para aplicações móveis que utilizem, por exemplo, telefones celulares, *pagers*, PDAs, etc. Historicamente, as WWANs seguem tecnologias especificadas por dois projetos: 3GPP (*The Third Generation Partnership Project* - Projeto de parceria para terceira geração de tecnologia de comunicação móvel) e 3GPP2 (*The Third Generation Partnership Project* 2 – Projeto de parceria para terceira geração de tecnologia de comunicação móvel 2). Além desses, o IEEE também criou um grupo de trabalho para definir o padrão IEEE 802.20 (IEEE 802.20 Standard, 2008), que trata de redes de acesso a banda larga móvel sem fio. Inicialmente projetado para ser uma rede exclusivamente de telefonia, para transmissão de voz. Entretanto essa necessidade foi mudando e fazendo com que aplicações de transmissão de dados tenham crescido muito nos últimos anos.

O 3GPP é o projeto que une padrões de telecomunicações, conhecido como parceria organizacional, e provê seus membros com um ambiente estável para produzir relatórios e especificações que definem tecnologias 3GPP. Essas tecnologias estão constantemente evoluindo – conhecidas como gerações de sistemas móveis celulares comerciais, conforme mostrado na Figura 2.1.

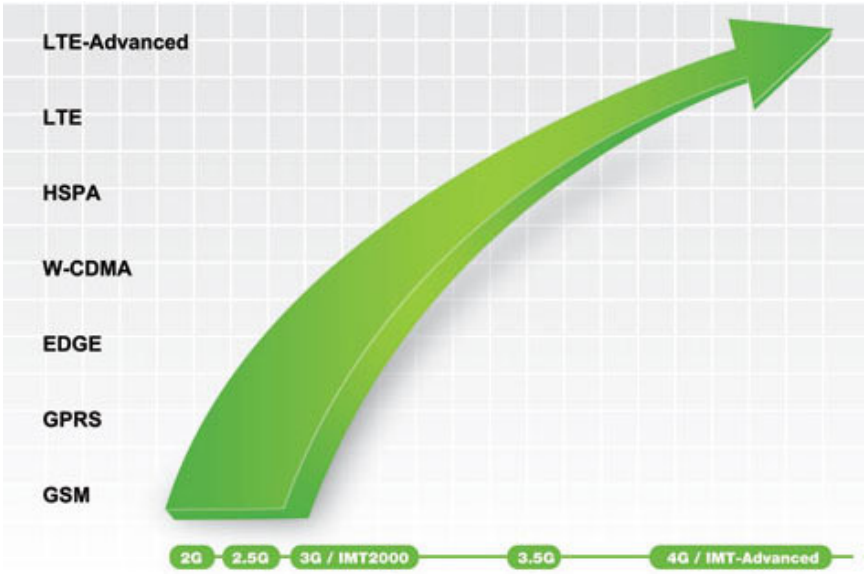

Figura 2.1. - Evolução dos padrões baseados em tecnologias 3GPP (About 3GPP, 2012)

Inicialmente, o 3GPP foi uma parceria para padronizar sistemas GSM (*Global System for Mobile Communications* – Sistema Global para Comunicações Móveis) para sua terceira geração de rede de telefonia móvel (3G), entretanto, desde a conclusão da primeira especificação de LTE (*Long Term Evolution*), o 3GPP tem se tornado um dos focos principais para sistemas móveis pós 3G. Com a mesma finalidade, o 3GPP2 é um projeto colaborativo para especificação de telecomunicação de terceira geração compreendendo interesses de desenvolvimento da América do Norte e da Ásia na evolução das tecnologias CDMA. A diferença das duas linhas de projetos é a origem geográfica e a família de tecnologias que são padronizadas. Um resumo da evolução das tecnologias de redes móveis sem fio foi apresentada por Dias (2010), Figura 2.2, diferenciando a família de tecnologias, a contemporaneidade, a distribuição geográfica e sua respectiva "geração" comercial.

Como a tecnologia WiMax, a partir do padrão IEEE 802.16m, está no cenário de redes móveis sem fio, foi publicada uma análise comparativa entre os padrões WiMax, HSPA+ e LTE (Gray, 2009). Já o padrão IEEE 802.20 foi o grupo de trabalho que especificou a rede sem fio de acesso móvel celular no IEEE e especificou a camada física (PHY), a camada de acesso ao meio (MAC) e a camada de controle de link lógico (LLC). Em março de 2011 o padrão foi colocado em hibernação por falta de atividade.

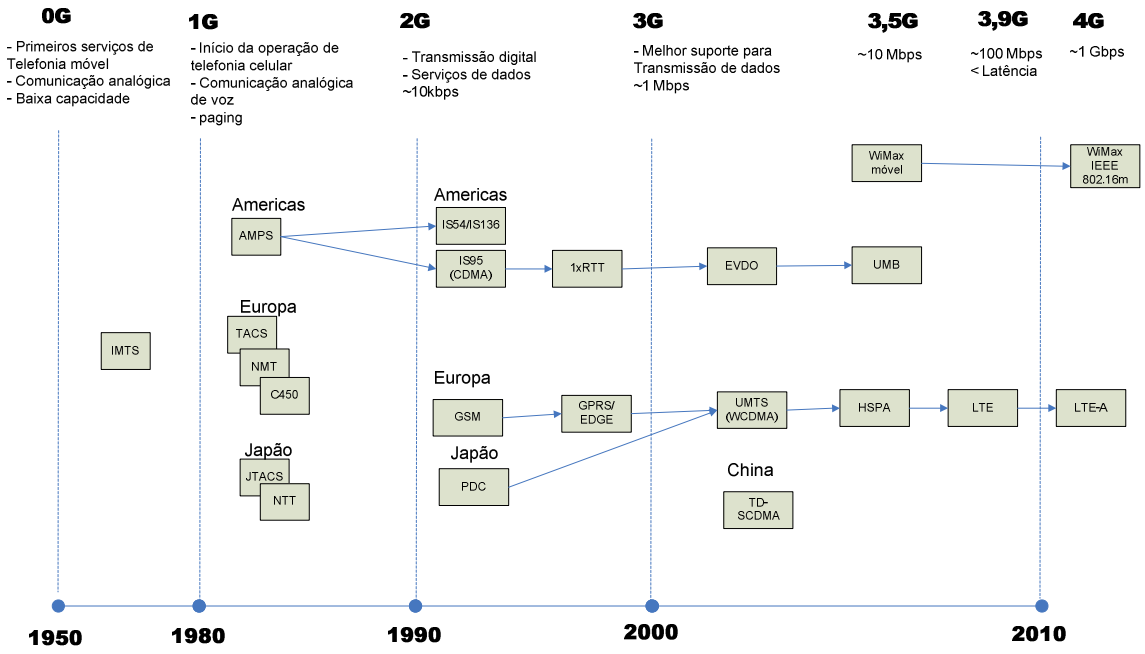

Figura 2.2 – Evolução das Redes Móveis (Dias, 2010)

2.2.5 WRAN (*Wireless Regional Area Networks*) – Redes sem fio de área regional

O grupo de trabalho para redes sem fio de área regional, proposto pelo padrão IEEE 802.22, também conhecido como WRAN (Redes sem fio de área regional), lançado em julho de 2011, objetiva desenvolver um padrão para acesso a banda larga sem fio fixa em áreas rurais e remotas usando freqüências de TV não utilizadas nas faixas de 54 a 862 MHz. Este padrão será baseado em tecnologia da rádio cognitivo. O sistema terá capacidade de "ouvir" o espectro, identificar canais de TV não utilizados, e utilizar estes canais para acesso à rede sem fio sem interferir nos usuários existentes, incluindo receptores de TV e microfones sem fio.

O objetivo desse tipo de rede é prover acesso sem fio com células de raio típico de 30 km e máximo de 100 km. As principais características são: sensibilidade de perceber os canais de TV não utilizados; capacidade do canal de 22 Mbps.

2.2.6 Considerações sobre redes sem fio homogêneas

Conforme se pode observar, existe uma variedade muito grande de tecnologias de redes sem fio e padrões de transmissão de dados. Essas tecnologias possuem características diversas e as aplicações mais distintas possíveis. Entretanto, essas redes são tratadas separadamente e trabalham de modo homogêneo, ficando presas a suas limitações. Uma forma de estender as funcionalidades de cada um desses tipos de rede, que funcionam

separadamente, e aproveitar benefícios de um padrão diferente é integrar padrões e tecnologias distintas em uma mesma rede heterogênea. Já existem grupos de trabalhos específicos destinados a pesquisar a interoperação entre padrões diferentes, como é o caso da versão do padrão IEEE 802.11u e IEEE 802.15.2. Essa interoperação entre redes será detalhada na próxima sessão.

2.3 Redes heterogêneas sem fio e gerenciamento de recursos

Conforme mostrado na sessão 2.2, existem vários padrões baseados em diferentes tecnologias de acesso a rádio (RAT – *Radio Access Technologies*). Apesar disso, ainda serão desenvolvidos novos tipos de redes para complementar aqueles que já existem. Como nenhuma RAT é capaz de atender todos os requisitos de usuários com cobertura universal, vislumbra-se que a próxima geração de redes sem fio seja integrada por múltiplas tecnologias, trabalhando conjuntamente de modo heterogêneo (Falowo & Chan, 2010). Esses padrões de rede sem fio são vistos como complementares e não como tecnologias concorrentes, o que leva ao desenvolvimento de redes heterogêneas que irão suprir a necessidades dependendo do espaço em que são aplicados.

Uma rede heterogênea sem fio (RHSF), conforme mostrado na Figura 2.3, é composta por mais de uma RAT, como LTE, UMTS, WLAN, WiMax, Bluetooth, etc. que coexistem na mesma área geográfica. Neste tipo de rede, cada tecnologia é limitada e têm suas próprias características como cobertura, largura de banda, nível de segurança, custo de serviço, nível de qualidade de serviço oferecido pelo operador de rede, etc. Além disso, é comum nos dias de atuais, aparelhos que suportam mais de uma tecnologia de comunicação sem fio no mesmo equipamento.

A motivação para a existência de redes sem fio heterogêneas vem do fato de que nenhuma tecnologia homogênea provê cobertura ubíqua e altos níveis de QoS em múltiplos espaços como residências, escritórios, espaços públicos e etc. (Falowo & Chan, 2008). Do ponto de vista econômico, em uma rede heterogênea, o usuário móvel pode também optar qual rede atenderá sua chamada usando o critério do menor custo monetário. Alem disso, o 3GPP adotou a tecnologia de redes heterogêneas a afim de atingir melhor desempenho, e introduziu sua utilização na padronização do LTE-Advanced (Cao *et al.*, 2012).

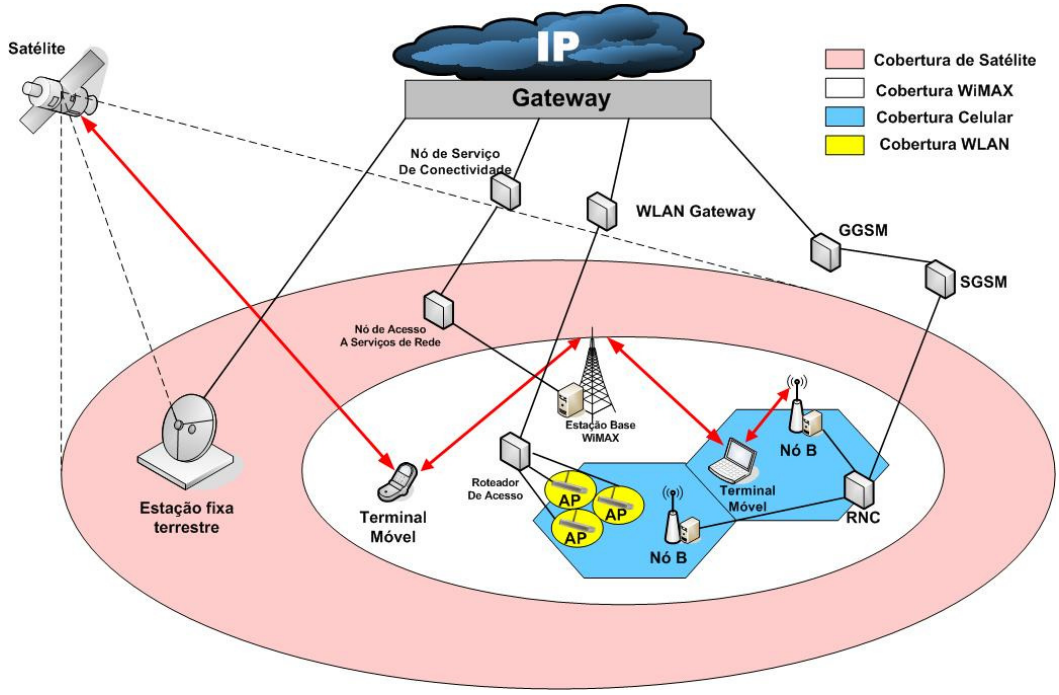

Figura 2.3 – Um exemplo de arquitetura de rede sem fio heterogênea (Falowo & Chan, 2008)

Segundo Falowo & Chan. (2008), existem vários benefícios para utilização de RHSF, dentre eles pode-se citar utilização eficiente de recursos de rádio, provisionamento de QoS consistente, estabilidade na rede como um todo, melhoria na satisfação do usuário e aumento no rendimento do operador de rede. Para isso, é desejável que algoritmos de CACC tenham certos requisitos para garantir os benefícios mencionados como: suporte a multi-serviços, eficiência, simplicidade, alta velocidade de execução, escalabilidade e estabilidade.

Além dos benefícios citados, gerenciar redes conjuntamente pode ajudar no momento da expansão de redes homogêneas atuando de forma complementar com outras tecnologias. No caso das redes de telefonia móvel, por exemplo, que atualmente são redes homogêneas, alterações da rede planejada e implantada, por conta de necessidades de melhoria de capacidade, se torna um processo complexo e iterativo. Além disso, aquisições de estações base para macro células, como torres, se tornam mais difíceis em áreas urbanas densas. Dessa forma, um modelo de desenvolvimento mais flexível é necessário para que os operadores de rede possam melhorar a experiência de usuário de forma ubíqua e rentável (Qualcomm Incorporated, 2011). Assim, ganhos futuros de redes sem fio serão obtidos através de topologia avançada de rede, no caso, redes celulares heterogêneas, onde a rede regular existirá sobreposta de várias "*pico base stations*", "*femto base stations*" e "*relay base stations*", conforme mostrado na Figura 2.4, como uma forma de aumentar a capacidade em alguns pontos da área de cobertura. Nesses casos, diferentemente do que acontece em redes

homogêneas, onde o terminal móvel é servido pela estação base com sinal mais forte, em redes celulares heterogêneas haverá estratégia de seleção e técnicas avançadas para gerenciamento de chamadas (Qualcomm Incorporated, 2011).

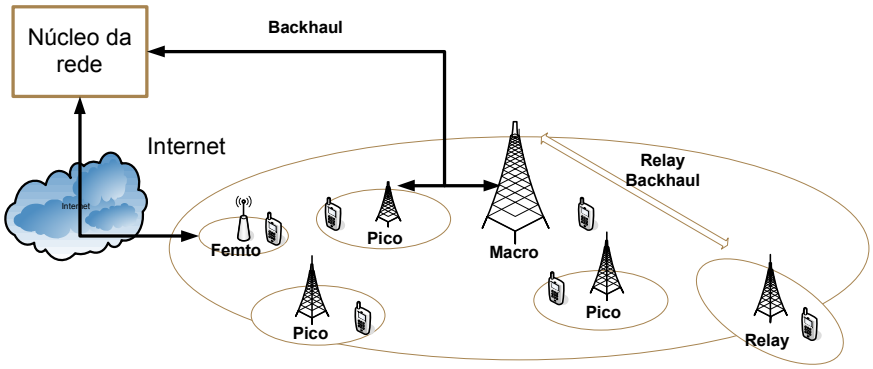

Figura 2.4 – Rede heterogênea utilizando estações base macro, pico, femto e relay (Qualcomm, 2011)

Vários aspectos devem ser considerados no projeto de redes heterogêneas. De uma perspectiva de demanda, volume de tráfego, localização de tráfego e taxa de dados são fatores importantes. De uma perspectiva de oferta aspectos importantes incluem ambiente de rádio, cobertura de macro células, transmissão de *backhaul*, espectro e integração com macro células existentes. Aspectos comerciais como competição de tecnologia, modelos de negócio, estratégias de mercado e preço também devem ser considerados (Landstrom *et. al.*, 2011).

Vislumbrando esta tendência, existem no mercado e na academia muitos esforços e realidades que integram tecnologias diferentes em uma única rede heterogênea gerenciada conjuntamente. Exemplo disso é a criação do grupo de trabalho IEEE 802.21, que trata de um padrão habilitando *handover* transparente e interoperabilidade tanto entre redes do mesmo tipo como em redes de tecnologias diferentes, conhecido como MIH (*Media Independent Handover* – Handover independente de tecnologia).

A integração de tecnologias em redes heterogêneas tem sido estudada por vários pesquisadores. A maior parte dos estudos nessa área foca na integração de redes UMTS e tecnologias de dados sem fio como WiFi e WiMax (Kassab *et. al.*, 2010). Kassab (2010) propõe um framework para integração de redes heterogêneas sem fio baseada no esqueleto e num mecanismo de gerenciamento de handover. Sun & Wang (2010), em seu trabalho, propõem duas arquiteturas avançadas de funcionalidade de intercomunicação entre WiMax e redes baseadas nas tecnologias 3GPP, baseado em arquiteturas centralizadas, concentrando

sua pesquisa na camada de link genérica (GLL) e em mecanismos de gerenciamento de recursos de rede. Em Lee *et. al.* (2009) é apresentado um estudo que gerencia *handover* vertical de uma rede 3G para uma rede WLAN. A política de decisão foi probabilisticamente derivada para evitar *handover* vertical desnecessário de 3G para WLAN.

Em Nogueira *et. al.* (2007) são apresentadas duas abordagens para gerenciamento de mobilidade em redes integradas 3G UMTS e IEEE 802.11, uma baseada em SIP (*Session Initiation Protocol*) e outro baseado em MIPv6 (*Mobile IP version 6*). Em Hasib *et. al.* (2010) é proposto um esquema de seleção de mobilidade de rede adaptativo, para redes sem fio heterogêneas integradas por WWAN e WLAN, baseado em informações da camada física, onde a validação é feita através de cadeias de Markov.

Além desses, trabalhos de integração de entre redes diferentes como WiFi e WiMax (Kassab *et. al.*, 2010) e integração de redes da mesma família como LTE-UMTS (Vucevic *et. al.*, 2011) são desenvolvidos, mostrando que este é um assunto em pauta, independente da tecnologia e do tipo de rede em questão. Por esse motivo, o algoritmo aqui proposto é um modelo genérico, independente de tecnologia, e que pode se aplicar a qualquer tipo de rede, visto que a necessidade de se integrar é comum a todas as redes.

Nesse contexto, gerenciamento de recursos de rádio (RRM – Radio Resource Management) é um assunto chave, visto que recursos de rádio são frequentemente escassos e caros, o que faz o seu estudo uma área de pesquisa constante. Além disso, a coexistência de diferentes RATs requer gerenciamento de recursos conjuntos para atingir níveis de QoS e utilização eficiente de recursos de radio (Falowo & Chan, 2008).

Quando uma chamada solicita recursos em uma rede, esta pode ser aceita ou rejeitada dependendo das condições da rede e da política de gerenciamento de recursos. O mecanismo que gerencia a aceitação e o bloqueio de chamadas é conhecido como CAC (Controle de Admissão de Chamadas). O propósito principal de algoritmos de CAC em redes sem fio é o melhor uso dos recursos disponíveis, garantindo que requisitos de QoS sejam satisfeitos para todas as chamadas aceitas (Falowo & Chan, 2008).

Diversos algoritmos de CAC têm sido desenvolvidos para redes homogêneas. Entretanto, algoritmos de CAC para redes homogêneas não provêem uma solução simples e boa para arquitetura heterogênea. Esta limitação tem levado ao desenvolvimento novos

algoritmos específicos para redes heterogêneas, chamados CACC (Controle de Admissão de Chamadas Conjunto) (Falowo & Chan, 2008).

Em algoritmos de CACC, além da tarefa de decidir se aceita ou não uma chamada que é iniciada, eles devem também decidir qual RAT é mais adequada para receber esta chamada, conforme é mostrado na Figura 2.5. Uma atividade é tão importante quanto a outra, pois contribuem da mesma forma para o desempenho, a estabilidade e utilização de recursos da melhor forma possível. Alguns algoritmos foram propostos para gerenciar recursos conjuntamente em redes heterogêneas.

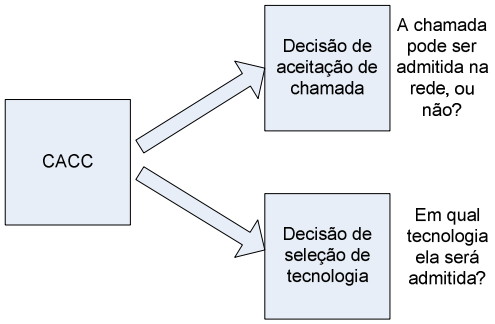

Figura 2.5 – Funções Básicas de um algoritmo de CACC

Em sua concepção, segundo Falowo & Chan (2008) os algoritmos de CACC podem ser classificados da seguinte forma: baseado em seleção aleatória; baseado em carga; baseado em classes de serviço; baseada em perda de percurso; baseado em custo de serviço; baseado em camada de rede; baseado em função custo/utilidade; baseado em inteligência computacional, além de CACC baseado em múltiplos critérios (que integram dois ou mais dos tipos citados anteriormente). Essa especificação de projeto do CACC é quem vai direcionar seu desenvolvimento e, principalmente, sua base de operação em busca da política de otimização da utilização de recursos de rede.

2.4 Considerações finais

Conforme mostrado, é grande o volume de pesquisa na área de redes heterogêneas e no gerenciamento conjunto dos recursos de redes. Essa necessidade tende a ficar mais evidente à medida que as redes começam a interagir e cooperam visando diversos benefícios. Portanto, o trabalho aqui abordado é um tema relevante, atual, e que pode gerar inúmeras discussões benéficas no desenvolvimento de tecnologias novas bem como no melhoramento das já existentes.

Capítulo 3 - Inteligência Artificial e Aprendizagem por Reforço

3.1 Introdução

Em geral, os dicionários da língua portuguesa, como em Michaelis (1998), definem inteligência como sendo a "faculdade de aprender, compreender e adaptar-se" e apresentam, pelo menos, mais duas acepções distintas para esse termo:

- Filosófica: "princípio espiritual e abstrato considerado como a fonte de toda a intelectualidade";
- Psicológica: "capacidade de resolver situações novas com rapidez e êxito".

Derivado do termo inteligência, a Inteligência Artificial (IA) é uma área de pesquisa dedicada a buscar métodos ou dispositivos computacionais que possuam ou simulem a capacidade racional de resolver problemas, pensar ou, de forma ampla, ser inteligente. Segundo Russel & Norvig (2004) as definições de IA encontradas na literatura podem ser agrupadas em quatro categorias principais, conforme Tabela 3.1. As que estão na parte superior da tabela se relacionam a processos de pensamento e raciocínio, enquanto as que estão na parte inferior se referem ao comportamento. As definições do lado esquerdo medem o sucesso em termos de fidelidade ao desempenho humano, enquanto as definições do lado direito medem o sucesso comparando-o a um conceito ideal de inteligência, chamado racionalidade. Historicamente, as quatro estratégias para o estudo de IA têm sido seguidas. Assim, existe uma tensão entre abordagens centradas em torno de seres humanos (que constitui uma ciência empírica) e abordagens centradas em torno da racionalidade (que combinam matemática e engenharia). Cada grupo tem ao mesmo tempo desacreditado e ajudado o outro (Russel & Norvig, 2004).

Tabela 3.1 – Definições de Inteligência Artificial (Russel & Norvig, 2004)

Sistemas que pensam como humanos	**Sistemas que pensam racionalmente**
"O novo e interessante esforço para fazer os computadores pensarem... máquinas com mentes, no sentido total e literal." (Haugeland, 1985 *apud*. Russel & Norvig, 2004)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985 *apud*. Russel & Norvig, 2004)
"[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978 *apud*. Russel & Norvig, 2004)	"O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992 *apud*. Russel & Norvig, 2004)
Sistemas que agem como humanos	**Sistemas que agem racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzeil, 1990 *apud*. Russel & Norvig, 2004)	"A inteligência computacional é o estudo do projeto de agentes inteligentes." (Poole et. al., 1998 *apud*. Russel & Norvig, 2004)

Dentro da IA, existe um sub-campo dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender possibilitando-o aperfeiçoar seu desempenho em uma tarefa, chamado aprendizagem de máquina. Aprendizagem de máquina denota mudanças, em um sistema, que são adaptativas no sentido de que elas capacitam o sistema a fazer a mesma tarefa, ou tarefas similares, mais eficientemente na próxima vez. Assim, esse campo visa a construção de algoritmos que possam adquirir conhecimento sobre a tarefa executada de modo automático.

Existem três tipos de aprendizagem (Osório *et. al.*, 1999 *apud*. Serra, 2004): supervisionada, não supervisionada e aprendizagem por reforço. Na aprendizagem supervisionada, o princípio básico é que deve-se conhecer, através de pares entrada-saída, quais as respostas que devem ser fornecidas pelo sistema para determinadas entradas ou impulsos externos. Quem terá esse conhecimento será uma espécie de supervisor, o qual

através da diferença entre os valores esperados e os valores obtidos será capaz de saber o erro que está sendo produzido, realizando então os ajustes dos parâmetros. O aprendizado estará completo quando o erro não mais existir, ou assumir valores satisfatoriamente pequenos. A partir desse momento, pode-se dizer que o sistema adquiriu o conhecimento passado pelo supervisor, estando então treinado para o problema apresentado.

Na aprendizagem não supervisionada, não são requeridas saídas desejadas e, por isso, é conhecido pelo fato de não precisar de "professores" para o seu treinamento. O treinamento da rede utiliza apenas os valores de entrada. A rede trabalha essas entradas e as organiza em categorias, usando para isso, os seus próprios critérios. Para uma entrada aplicada à rede, será fornecida uma resposta indicando a classe a qual a entrada pertence. Se o padrão de entrada não corresponde às classes existentes, uma nova classe é gerada.

Já na aprendizagem por reforço, o usuário possui apenas indicações imprecisas sobre o comportamento final desejado. Nesse tipo de aprendizagem dispõe-se apenas de uma avaliação qualitativa do comportamento do sistema sem, no entanto, poder medir quantitativamente o erro (desvio do comportamento em relação ao comportamento de referência desejado). A aprendizagem por reforço é um método por tentativa e erro, baseando suas ações somente em um índice de desempenho, chamado de "sinal de reforço", que é utilizado para otimização. Aprendizagem por reforço será detalhada o item 3.2 que segue.

3.2 Aprendizagem por Reforço (AR)

Segundo (Sutton & Barto, 1998), aprendizagem por reforço é aprender o que fazer – como mapear situações e ações – de modo a maximizar um sinal de reforço numérico. Ao que aprende, não e dito que ações tomar, como na maioria das formas de aprendizagem de máquina, mas ao contrario, deve-se descobrir qual ação produz maior reforço por tentativa e erro. Aprendizagem por Reforço é, antes de tudo, indicado quando se deseja obter uma política ótima (comportamento que o agente segue para alcançar um objetivo) nos casos em que não se conhece *a priori* a função que modela esta política. O agente deve interagir com seu ambiente diretamente para obter essas informações, que serão processadas através de um algoritmo apropriado, a fim de executar as ações que levem o agente a atingir os seus objetivos.

3.2.1 Características Gerais da AR

Os elementos principais que caracterizam a Aprendizagem por Reforço, diferenciando-a de outras abordagens de aprendizagem, são descritas abaixo.

Aprendizado por Interação: a interação entre o agente e o ambiente é a característica principal que define um problema de AR. O agente executa uma ação e, após isso, aguarda um sinal de reforço do ambiente em resposta à ação tomada, assimilando através do aprendizado o valor de reforço obtido para tomar decisões posteriores.

Retorno Atrasado: em um sistema de AR, busca-se alcançar objetivos globais no ambiente em longo prazo. Assim, as ações tomadas têm como objetivo maximizar o retorno total, isto é, a qualidade das ações tomadas é avaliada pelas soluções encontradas não como resultado imediato.

Orientado pelo Objetivo: como em AR, é considerado apenas um ambiente que dá respostas perante ações efetuadas, não é necessário conhecer detalhes da modelagem desse ambiente. Existe, simplesmente, um agente que interage dentro desse ambiente desconhecido tentando alcançar um objetivo. O objetivo é, geralmente, otimizar algum comportamento dentro do ambiente.

Investigação x Exploração: o dilema conhecido na literatura como "*The Exploration x Exploitation Dilemma*" – **Dilema da Investigação x Exploração**, que consiste em decidir quando se deve aprender e quando não se deve aprender sobre o ambiente, mas usar a melhor informação já obtida até o momento também é encontrado nos agentes de AR. Para que um sistema seja realmente autônomo, esta decisão deve ser tomada pelo próprio sistema, sem a interferência humana.

A decisão é fundamentalmente uma escolha entre agir baseado na melhor informação de que o agente dispõe no momento ou agir para obter novas informações sobre o ambiente que possam permitir níveis de desempenho ainda maiores no futuro. Isto significa que o agente deve aprender quais ações maximizam os valores dos ganhos obtidos no tempo, mas também, deve agir de forma a atingir esta maximização, explorando ações ainda não executadas ou regiões pouco visitadas do espaço de estados. Como ambas as formas trazem, em momentos específicos, benefícios à solução dos problemas, uma boa estratégia é mesclar os modos de investigação (*exploration*) e aproveitamento (*exploitation*).

Este é um problema crucial no contexto da aprendizagem por reforço, pois agir para obter informação pode aumentar o desempenho em longo prazo, embora faça com o desempenho em curto prazo diminua. Tomando-se estes cuidados, quanto mais tempo o agente estiver atuando no ambiente, mais corretas serão suas ações no decorrer de sua tarefa.

3.2.2 Problema de AR

Basicamente, os sistemas de Aprendizagem por Reforço são constituídos por um agente que interage com um ambiente por meio de ação e percepção. Assim, o agente toma uma ação e, como conseqüência desta ação, percebe qual a reação do sistema. A ação tomada muda de alguma forma o ambiente, afetando o seu estado na tentativa de alcançar o objetivo relacionado. As mudanças são comunicadas ao agente através de um sinal de reforço. Como pode ser visto na Figura 3.1.

Na Figura 3.1, o agente executa uma ação e, como consequência, recebe do ambiente o reforço por ter tomado esta ação e o próximo estado que se encontra o ambiente, que muda por causa da ação tomada. O agente deve monitorar o ambiente frequentemente e reagir apropriadamente, pois os efeitos das ações não podem ser perfeitamente antecipados. Em um sistema de AR, o estado do ambiente é representado por: 1) um conjunto de variáveis de estado percebidas pelo agente, onde o conjunto das combinações de valores dessas variáveis forma o conjunto de estados discretos do agente (S); 2) um conjunto de ações discretas possíveis, que escolhidas por um agente muda o estado do ambiente (A(s)) e 3) valores de transições de estados, que são passados ao agente através de um sinal de reforço, denominado ganho (valores tipicamente entre 0 e 1) (Govasi, 2003).

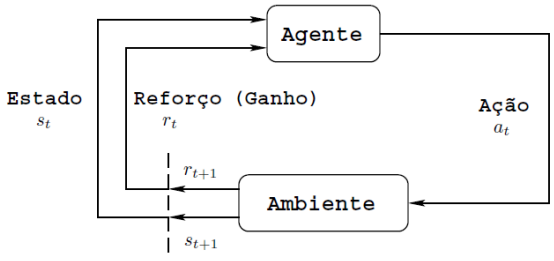

Figura 3.1 - Interação Agente x Ambiente (Sutton & Barto, 1998)

O objetivo do método é levar o agente a escolher a seqüência de ações que tendem a aumentar a soma de valores de reforço, ou seja, é encontrar a política π, definida por Sutton & Barto (1998) como o mapeamento de estados e ações que maximize as medidas do reforço acumuladas no tempo.

A Aprendizagem por Reforço apresenta cinco partes fundamentais, que são:

O Ambiente: tudo que está fora do agente, compreendendo os estados, as ações possíveis de serem tomadas, o conjunto de reforços, é chamado de ambiente em um sistema de AR. O ambiente é o espaço onde está inserido o agente e deve ser, pelo menos, parcialmente observável através de sensores, descrições simbólicas ou situações mentais.

A Política: expressa pelo termo π, as ações a serem tomadas pelo agente para alcançar o objetivo. Em outras palavras, uma política π é um mapeamento de estados *S* e ações *A* em um valor π(s, a). Assim, se um agente AR muda a sua política, então as probabilidades de seleção de ações sofrem mudanças e conseqüentemente, o comportamento do sistema apresenta variações à medida que o agente vai acumulando experiência a partir das interações com o ambiente. A política é quem define quais ações serão executadas em um determinado instante, e o objetivo do sistema AR é aprender até chegar a uma política de controle ótimo que permita ao sistema maximizar seus rendimentos.

Reforço e Retorno: O Reforço é um sinal do tipo escalar (r_{t+1}), que é devolvido pelo ambiente ao agente assim que uma ação tenha sido efetuada e uma transição de estado ($s_t \rightarrow s_{t+1}$) tenha ocorrida. Existem diferentes formas de definir o reforço para cada transição no ambiente, gerando-se funções de reforço que, intrinsecamente, expressam o objetivo que o sistema AR deve alcançar. O agente deve maximizar a quantidade total de reforços recebidos chamado de retorno, que nem sempre significa maximizar o reforço imediato a receber, mas o reforço acumulado durante a execução total.

De modo geral, o sistema AR busca maximizar o valor esperado de retorno, com isso, o retorno pode ser definido como uma função da seqüência de valores de reforço até um tempo *T* final. No caso mais simples é um somatório como aparece na Equação 3.1.

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + ... + r_T \tag{3.1}$$

Em muitos casos a interação entre agente e ambiente não termina naturalmente em um episódio (seqüência de estados que chegam até o estado final), mas continua sem limite, como, por exemplo, em tarefas de controle contínuo. Para essas tarefas a formulação do retorno é um problema, pois $T = \infty$ e o retorno que se deseja também tenderá ao infinito ($R_t = \infty$). Para estes problemas foi criada a taxa de amortização (γ), a qual determina o grau de

influência que têm os valores futuros sobre o reforço total. Assim, a expressão do retorno aplicando taxa de amortização é expressa pela Equação 3.2

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} = \sum \gamma^k r_{t+k+1} \tag{3.2}$$

Onde, $0 \leq \gamma \leq 1$. Se $\gamma \rightarrow 0$, o agente tem uma visão míope dos reforços, maximizando apenas os reforços imediatos, e se $\gamma \rightarrow 1$, a visão do reforço abrange todos os estados futuros dando maior importância ao estado final, desde que a seqüência R_t seja limitada.

Função de Reforço: As funções de reforço podem ser bastante complicadas, porém existem pelo menos três classes de problemas freqüentemente usadas para criar funções adequadas a cada tipo de problema:

- **Reforço só no estado final:** Nesta classe de funções, as recompensas são todas zero, exceto no estado final, em que o agente recebe uma recompensa real (ex: +1) ou uma penalidade (ex: −1). Como o objetivo é maximizar o reforço, o agente irá aprender que os estados correspondentes a uma recompensa são bons e os que levaram a uma penalidade devem ser evitados.

- **Tempo mínimo ao objetivo:** Funções de reforço nesta classe fazem com que o agente realize ações que produzam o caminho ou trajetória mais curta para um estado objetivo. Toda ação tem penalidade (−1), sendo que o estado final é (0). Como o agente tenta maximizar valores de reforço, ele aprende a escolher ações que minimizam o tempo que leva a alcançar o estado final.

- **Minimizar reforços:** Nem sempre o agente precisa ou deve tentar maximizar a função de reforço, podendo também aprender a minimizá-las. Isto é útil quando o reforço é uma função para recursos limitados e o agente deve aprender a conservá-los ao mesmo tempo em que alcança o objetivo.

Função Valor-Estado/Valor-Ação: Define-se uma função valor-estado como o mapeamento do estado, ou par estado-ação em um valor que é obtido a partir do reforço atual e dos reforços futuros.

Se a função valor-estado considera só o estado *s* é denotada como *V(s)*. Neste caso, o agente recebe um valor de reforço por estar no estado *s*. Já quando é considerado o par estado-ação *(s,a)*, então a função valor-estado é denotada como função valor-ação *Q(s,a)*, o que

significa que em um estado s, cada ação a ser tomada terá um valor de reforço diferente a ser recebido pelo agente.

- **Função valor-estado:** Uma vez que os reforços futuros mantêm dependências das ações futuras, as funções valor dependem também da política π que o AR segue. Em um Processo de Decisão Markoviano se define uma função valor-estado $V\pi(s)$ dependente da política π como a Equação 3.3:

$$V\pi(s) = \mathrm{E}\pi\{R_t \mid s_t = s\} = \mathrm{E}\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right\} \tag{3.3}$$

onde a função $V\pi(s)$ é o valor esperado do retorno para o estado $s_t = s$. Isto é, o somatório dos reforços aplicando a taxa de amortização γ.

- **Função valor-ação:** Se consideramos o par estado-ação, a equação para a função valor-estado $Q\pi(s,a)$ será, como apresentado na Equação 3.4:

$$Q^{\pi}(s,a) = \mathrm{E}\pi\{R_t \mid s_t = s, a_t = a\} = \mathrm{E}\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right\} \tag{3.4}$$

que é semelhante à Equação 3.3, só que considerando o reforço esperado para um estado $s_t = s$ e uma ação $a_t = a$.

As equações 3.3 e 3.4 apresentam as funções valor-estado e valor-ação respectivamente, que dependem dos valores de reforço, o qual implica o conhecimento completo da dinâmica do ambiente como um Processo de Decisão Markoviano, o qual será detalhado na sessão 3.2.3.

3.2.3 Fundamentos matemáticos

Existem dois conceitos que devem ser conhecidos para facilitar a modelagem de um problema como um sistema de Aprendizagem por Reforço. A seguir, apresentamos uma breve descrição destes conceitos.

3.2.3.1 Propriedade de Markov

Um estado é formado pelo conjunto de variáveis que representam a situação do sistema em um determinado momento. Assim, quando a probabilidade de transição de um estado *s* para um estado *s'* depende apenas do estado *s* e da ação *a* adotada em *s*, isso significa que o estado corrente fornece informação suficiente para o sistema de aprendizagem decidir

que ação deve ser tomada. Quando o sistema possui essa característica, diz-se que ele satisfaz a Propriedade de Markov.

No caso mais geral, se a resposta em *t + 1* para uma ação efetuada em *t* depende de todo o histórico de ações até o momento atual, a dinâmica do ambiente é definida pela especificação completa da distribuição de probabilidades, como mostra a Equação 3.5:

$$\Pr\{s_{t+1}=s', r_{t+1}=r \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, ..., r_1, s_0, a_0\} \quad (3.5)$$

onde a probabilidade (*Pr*) do estado s_{t+1} ser o estado s' e o reforço r_{t+1} ser igual a *r* é uma função que depende de todos os estados, ações e reforços passados. Se a resposta do ambiente em *t + 1* depende apenas dos estados e reforços em *t*, então, a probabilidade da transição para o estado s' é dada pela expressão da Equação 3.6.

$$P^a_{s,s'} = \Pr\{s_{t+1}=s' \mid s_t=s, a_t=a\} \quad (3.6)$$

A probabilidade de transição satisfaz às seguintes condições: 1) $P^a_{s,s'} \geq 0, \forall s, s' \in S, \forall a \in A(s)$ e 2) $\sum_{s' \in S} P^a_{s,s'} = 1, \forall s \in S, \forall a \in A(s)$

A Propriedade de Markov é de fundamental importância na AR, uma vez que tanto as decisões como os valores são funções apenas do estado atual, abrindo a possibilidade de métodos de soluções incrementais, onde pode-se obter soluções a partir do estado atual e para cada um dos estados futuros, como é feito no método de Programação Dinâmica.

3.2.3.2 Processos Markoviano de Decisão (PMD)

Segundo Bellman apud. Serra (2004), um Processo Markoviano de Decisão é definido como um conjunto de estados $S, \forall s \in S$, um conjunto de ações *A(s)*, um conjunto de transições entre estados associadas com as ações e um conjunto de probabilidades *P* sobre o conjunto *S* que representa uma modelagem das transições entre os estados. Assim, dado um par de estado e ação, a probabilidade do estado *s* passar a um estado s' é dado na Equação 3.7

$$P^a_{s,s'} = \Pr\{s_{t+1}=s' \mid s_t=s, a_t=a\} \quad (3.7)$$

onde *Pr* é o operador de probabilidade; neste caso representa-se a probabilidade do estado s_{t+1} ser s', sempre que o estado s_t for igual a *s* e a ação a_t for igual a *a*. Desta forma, a dependência que o estado seguinte s_{t+1} seja o estado s' está relacionada a tomar a ação *a* no instante *t*.

De forma análoga, dados um estado e ação atuais e um estado seguinte s′, o valor esperado do retorno é dado na Equação 3.8.

$$R^{a}_{s,s'} = E\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\} \tag{3.8}$$

onde E é o valor esperado do retorno r_{t+1}, sempre que o estado s_t no instante t passe a ser o estado s' no instante $t + 1$.

Os valores de probabilidade $P^{a}_{s,s'}$ e retorno esperado $R^{a}_{s,s'}$ determinam os aspectos mais importantes da dinâmica de um PDM finito. Podemos caracterizá-lo como: 1) um ambiente evolui probabilisticamente baseado num conjunto finito e discreto de estados; 2) para cada estado do ambiente, existe um conjunto finito de ações possíveis; 3) cada passo que o sistema de aprendizado executar uma ação, é verificado um custo positivo ou negativo para o ambiente em relação à ação; e 4) estados são observados, ações são executadas e reforços são relacionados.

Assim para quase todos os problemas de Aprendizagem por Reforço é suposto que o ambiente tenha a forma de um Processo Markoviano de Decisão, desde que seja satisfeita a Propriedade de Markov no ambiente. Nem todos os algoritmos de AR necessitam uma modelagem PMD inteira do ambiente, mas é necessário ter-se pelo menos a visão do ambiente como um conjunto de estados e ações (Govasi, 2003).

3.2.4 Métodos de Solução

Para solucionar o problema de Aprendizagem por Reforço, que são: 1) avaliação de política; e 2) encontrar a política de controle ótimo, existem três classes de métodos fundamentais (Sutton & Barto, 1998): Programação Dinâmica, Monte Carlo e Diferença Temporal, que apresentaremos e analisaremos nas subseções a seguir.

3.2.4.1 Programação Dinâmica (PD)

A Programação Dinâmica tem a vantagem de ser matematicamente bem fundamentada, mas exige uma modelagem bem precisa do ambiente como um Processo Markoviano de Decisão. Programação Dinâmica é uma coleção de algoritmos que podem obter políticas ótimas sempre que exista uma modelagem perfeita do ambiente como um PMD, isto é, como um conjunto de estados, ações, retornos e probabilidades da transição em todos os estados. Os algoritmos clássicos de PD são usados de forma limitada, uma vez que a

modelagem perfeita do ambiente como PDM exige um grande custo computacional, porém, fornece um bom fundamento para o conhecimento dos outros métodos usados na solução do problema de AR e um padrão de comparação.

A dinâmica do sistema é dada por um conjunto de probabilidades de transição de estado, $P^a_{s,s'} = \Pr\{s_{t+1} = s' \mid s_t = s, a_{t+1} = a\}$, e por um conjunto de reforços imediatos esperados, $R^a_{s,s'} = E\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\}$, para todo par $s, s' \in S$ e $a \in A(s)$.

Avaliação da Política

As escolhas das ações do agente são feitas a partir de uma função do estado, chamada política ($\pi : S \rightarrow A$). O valor de utilidade de um estado, dada uma política, é o reforço esperado partindo do estado e seguindo a política apresentada na Equação 3.3. E paralelamente a essa função valor-estado, existe uma função valor-ação para a política π, que é definida pela Equação 3.4.

As funções valores $V\pi$ e $Q\pi$ podem ser estimadas por experiências. Por exemplo, se um agente seguir uma política π e mantiver uma média, para cada estado encontrado, dos atuais retornos que tem seguido aquele estado, então a média convergirá ao valor-estado $V\pi$. Se médias diferentes forem mantidas para cada ação feita em um estado, então estas médias convergirão similarmente aos valores da ação $Q\pi$.

Uma propriedade fundamental das funções de valor usadas durante a Aprendizagem por Reforço e Programação Dinâmica é que elas satisfazem particularidades recursivas:

$$\begin{aligned} V\pi &= E\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right\} \\ &= E\pi\left\{r_{t+1+\gamma}\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s\right\} \\ &= \sum_{a \in A(s)} \pi(s,a) \sum_{s' \in S} P^a_{s,s'}\left[R^a_{s,s'} + \gamma E\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s'\right\}\right] \\ &= \sum_{a \in A(s)} \pi(s,a) \sum_{s' \in S} P^a_{s,s'}\left[R^a_{s,s'} + \gamma V\pi(s')\right], \forall s \in S \end{aligned} \tag{3.9}$$

A Equação 3.9 é a equação de Bellman para $V\pi$ expressa um relacionamento entre o valor de um estado e os valores de seus estados sucessores. Calcula todas possíveis médias excessivamente, ponderando cada um por sua probabilidade de ocorrer. Indica que o valor do estado *s* inicial deve igualar ao valor amortizado do estado seguinte previsto, mais a

recompensa esperada ao longo do caminho. O algoritmo para resolver a avaliação de política pode ser visto abaixo.

Avaliação Iterativa de Política

Entrar π, a política a ser avaliada
Inicializar V(s) = 0 para todo s ∈ S
Repete
$\Delta \leftarrow 0$
Para cada $s \in S$*:*
$v \leftarrow V(s)$
$V(s) \leftarrow \sum_{a \in A(s)} \pi(s,a) \sum_{s' \in S} P^a_{s,s'} \left[R^a_{s,s'} + \gamma V(s') \right]$
$\Delta \leftarrow \max \{\Delta, | v - V(s) |\}$
Até $\Delta < \theta$ (um pequeno número positivo)
Sair $V \approx V^{\pi}$

Dadas duas políticas π e π', qual delas é a mais desejável? Para compararmos as políticas devemos calcular as funções valor-estado induzidas por elas. Dizemos que π domina π′, denotado por $\pi < \pi'$, se $V\pi(s) \geq V\pi'(s)$ para todo $s \in S$ Se $V\pi(s) > V\pi'(s)$ para algum $s \in S$. Além de permitir a comparação entre políticas, avaliação de política pode levar a um algoritmo para busca de uma política de controle ótima, o que será objeto de estudo mais adiante.

Política de Controle Ótimo

A Programação Dinâmica organiza e estrutura a busca de boas políticas a partir das funções de valor-estado e valor-ação. Deste modo, políticas ótimas são obtidas sempre que funções valor ótimas são obtidas. Usualmente as funções valor-estado e valor-ação ótimas são denotadas por *V*(s)* e *Q*(s)*, respectivamente, como é expresso nas equações 3.10 e 3.11:

$$
\begin{aligned}
V^*(s) &= \max_{a \in A(s)} Q^{\pi^*}(s,a) \\
V^*(s) &= \max_a E_{\pi^*} \{ R_t \mid s_t = s, a_t = a \} \\
V^*(s) &= \max_a E_{\pi^*} \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a \right\} \\
V^*(s) &= \max_a E_{\pi^*} \left\{ r_{t+1} \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s, a_t = a \right\} \\
V^*(s) &= \max_a E \{ r_{t+1} + \gamma V^*(s_t + 1) \mid s_t = s, a_t = a \} \\
V^*(s) &= \max_a \sum_{s'} P^a_{s,s'} \left[R^a_{s,s'} + \gamma V^*(s') \right]
\end{aligned}
\tag{3.10}
$$

$$Q^*(s,a) = E\left\{ r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') \mid s_t = s, a_t = a \right\}$$
$$Q^*(s,a) = \sum_{s'} P^a_{s,s'} \left[R^a_{s,s'} + \gamma \max_{a'} Q^*(s', a') \right] \quad (3.11)$$

Na Equação 3.10 a função valor ótimo *V*(s)* é encontrada como o máximo das funções de valor esperadas segundo a ação selecionada. E a partir de *Q**, pode-se determinar uma política ótima simplesmente como $\pi^*(s) = arg\ max_{a \in A(s)}\ Q^*(s,a)$.

Para atualizar a função valor com a finalidade de melhorar a política, utiliza-se a iteração de valores. A função valor $V_{k+1}(s)$ do estado *s* para o passo *k+1* de avaliação de política como é encontrada na Equação 3.12:

$$V_{k+1}(s) = \max E\{ r_{t+1} + \gamma V_k(s_{t+1}) \mid s_t = s, a_t = a \}$$
$$V_{k+1}(s) = \max_a \sum_{s'} P^a_{s,s'} \left[R^a_{s,s'} + \gamma V_k(s') \right], \forall s \in S \quad (3.12)$$

onde o valor atualizado $V_{k+1}(s)$ é encontrado a partir dos valores armazenados no passo k da sequencia de iterações, aplicando a equação de otimalidade de Bellman. Esta sequência de iterações deve alcançar no ponto final a política ótima *V*(s)*.

O método de procura da política ótima da Programação Dinâmica exige a varredura de todos os estados no espaço do modelo de PMD, fazendo com que exista um grande custo computacional para modelagens complexas, resultando em uma desvantagem do método.

O termo gulosa (*greedy*) é usado para descrever um procedimento de busca ou de decisão, que seleciona alternativas baseada somente em considerações locais e/ou imediatas, sem considerar a possibilidade que tal seleção poder encontrar no futuro alternativas melhores. Consequentemente descreve as políticas que as ações baseiam somente em consequências em curto prazo.

Valor de Iteração

Inicializar V de forma arbitrária, por exemplo V(s) =0 para todo s ϵ S
Entrar π, a política a ser avaliada
Inicializar V(s) = 0 para todo s ϵ S
Repete
  $\Delta \leftarrow 0$
  Para cada s ϵ S:
  $v \leftarrow V(s)$
  $V(s) \leftarrow \max_{a \in A(s)} \sum_{s' \in S} P^a_{s,s'} \left[R^a_{s,s'} + \gamma V(s') \right]$
  $\Delta \leftarrow \max\{\Delta, | v - V(s) |\}$
Até $\Delta < \theta$ (um pequeno número positivo)
Sair uma política determinística, π, *tal que*
$\pi(s) = \arg\max_{a \in A(s)} \sum_{s' \in S} P^a_{s,s'} \left[R^a_{s,s'} + \gamma V(s') \right]$

3.2.4.2 Método de Monte Carlo (MC)

O método de Monte Carlo não precisa da modelagem do ambiente e se apresenta de forma simples em termos conceituais, baseia-se no cálculo da média de retornos obtidos em seqüências. Para assegurar-se que exista um valor de retorno bem definido, o método de Monte Carlo é utilizado apenas para tarefas episódicas, isto é, a experiência é dividida em episódios que de algum modo alcançam o estado final sem importar as ações que foram selecionadas (exemplo: jogo de xadrez). Desta forma, somente depois da conclusão de um episódio o valor de retorno é obtido e as políticas são atualizadas. Entretanto, não são viáveis quando a solução do problema é possível apenas de forma incremental, porque para se atualizar, o método de Monte Carlo exige que seja alcançado o estado final no processo e com isso o mesmo pode se apresentar lento.

Uma vantagem do método de Monte Carlo é que, diferente do método de Programação Dinâmica, não necessita de informação completa do ambiente, apenas necessita das amostras da experiência como seqüências de dados, ações e reforços a partir de uma interação real ou simulada com o ambiente.

O aprendizado a partir de experiência real é notável, uma vez que não exige o conhecimento a priori das dinâmicas do ambiente, e ainda, pode levar a um comportamento ótimo. Embora seja requerida uma modelagem, esta deve apenas gerar transições de estados, sem precisar todo o conjunto de distribuições de probabilidade para todas as possíveis transições, como é exigido pela Programação Dinâmica.

Avaliação da Política

Supondo-se que o método de Monte Carlo é considerado para obter a função valor sob uma dada política, que é representada pelo retorno esperado, isto é, a acumulação amortizada dos futuros reforços desde o estado *s* até o estado desejado. Uma forma de se aproximar o valor de retorno esperado a partir da experiência é calcular a média dos retornos observados após visitar esse estado. Na medida em que mais retornos são observados, a média deve se aproximar do valor real esperado, sendo esta conseqüência o princípio básico do método de Monte Carlo.

Seja $V\pi(s)$ a função valor-estado sob a política π. Dados um conjunto de episódios obtidos sob a mesma política passando pelo estado s (cada ocorrência de *s* em um episódio é chamada de visita a s), existem duas variantes do método de Monte Carlo: A primeira,

chamada de *Every-Visit MC Method*, estima a função de valor como a média dos retornos após todas as visitas ao estado *s*, enquanto a segunda, chamada de *First-Visit MC Method*, estima a função de valor como a média dos retornos após a primeira visita ao estado s.

De qualquer forma, se o número de visitas for infinito, ambas as variantes do método de Monte Carlo, convergem ao valor $V\pi(s)$. Podemos ver através do algoritmo abaixo.

MC com exploração no início

Inicializar, para todo $s \in S$, $a \in A(s)$
Q(s,a) $\leftarrow$ de forma arbitraria
$\pi(s) \leftarrow$ de forma arbitraria
Retornos(s) $\leftarrow$ uma lista vazia, para todo $s \in S$
Repetir infinitamente:
(a) Gerar um episodio usando π explorando novos estados

(b) Para cada par ***(s,a)*** gerado no episodio:
$R \leftarrow$ retorno após a primeira ocorrência de ***(s,a)***
Adiciona ***R*** a lista ***Retornos(s)***
Q(s,a) $\leftarrow$ media de (***Retornos(s)***)
$V(s) \leftarrow$ media de (***Retornos(s)***)
(c) Para cada s do episódio:
$\pi(s) \leftarrow \arg\max_a Q(s,a)$

Política Ótima

A fim de melhorar a política é necessário fazer com que esta seja mais gulosa para a função valor-estado $V\pi(s)$ atual. Neste caso é conveniente assumir como valor de retorno a função valor-ação $Q\pi(s,a)$. Assim, uma política gulosa para uma função valor-ação $Q(s,a)$ é aquela que para um estado *s* toma a ação que maximiza o valor Q como na Equação 3.13 que segue.

$$\pi(s) = \arg\max_{a \in A(s)} Q(s,a) \tag{3.13}$$

Desta forma, uma melhora na política pode ser obtida fazendo a política π_{k+1} ser gulosa em respeito à função valor-ação $Q\pi_k$, logo após a avaliação da função valor-ação $Q\pi_k$ de π_k, podemos gerar uma seqüência de avaliação da função e melhora da política, conforme Equação 3.14:

$$\pi_0 \xrightarrow{A} Q^{\pi_0} \xrightarrow{M} \pi_1 \xrightarrow{A} Q^{\pi_1} \ldots \xrightarrow{A} Q^{\pi_k} \xrightarrow{M} \pi_{k+1} \ldots \xrightarrow{M} \pi^* \xrightarrow{A} Q^* \tag{3.14}$$

Onde A indica avaliação de política e *M* indica processo guloso de melhora de política. Segundo este processo se o número de episódios é muito grande, a função valor se aproximará à função valor-ação ótima *Q**.

Primeira visita ao método MC para estimar V^{π}

Inicializar
$\pi \leftarrow$ política a ser avaliada
$V \leftarrow$ uma função valor-estado arbitraria
Retornos(s) $\leftarrow$ uma lista vazia, para todo $s \in S$
Repetir sempre:
(a) Gerar um episodio usando π

(b) Para cada estado *s* que aparece no episódio:
$R \leftarrow$ quantidade de retorno após a primeira ocorrência de *s*
Adiciona ***R*** a lista ***Retornos(s)***
$V(s) \leftarrow$ media de (***Retornos(s)***)

3.2.4.3 Método da Diferença Temporal (DT)

Os métodos de Diferenças Temporais não exigem um modelo exato do sistema e permitem ser incrementais, da mesma forma que os métodos de Monte Carlo. Eles são uma combinação de características dos métodos de Monte Carlo com as idéias da Programação Dinâmica, que buscam estimar valores de utilidade para cada estado no ambiente. Em outras palavras, quanto mais próximo da convergência do método, mais o agente tem certeza de qual ação tomar em cada estado.

O aprendizado é feito diretamente a partir da experiência, sem a necessidade de uma modelagem completa do ambiente, como característico do método de Monte Carlo, mas leva vantagem em cima deste por atualizar as estimativas da função valor a partir de outras estimativas já aprendidas em estados sucessivos (*bootstrap*), sem a necessidade de alcançar o estado final de um episódio antes da atualização. Neste caso a avaliação de uma política é abordada como um problema de predição, isto é, estimar a função valor $V\pi$ sob a política π.

Avaliação da Política - Predição DT

Tanto DT como MC utilizam a experiência para resolver o problema da predição. Dada certa experiência sob a política π, se é visitado um estado intermediário s_t, ambos os métodos atualizam suas estimativas $V\pi\ (s_t)$ baseando-se no acontecido depois de visitado o estado. Sendo que o método de Monte Carlo espera até que o retorno total seja conhecido e usa esse retorno como objetivo para a atualização de $V\pi\ (s_t)$, como aparece na equação abaixo.

$$V\pi(s_t) \leftarrow V\pi(s_t) + \alpha[R_t - V\pi(s_t)] \quad (3.15)$$

onde R_t representa o retorno atual no instante t, e o símbolo α é uma constante de atualização (taxa de aprendizagem), $\alpha \in [0,1]$.

Os métodos de Diferenças Temporais não necessitam alcançar o estado final de um episódio, e sim o estado seguinte no instante $t + 1$. Em DT são utilizados, o valor de reforço imediato r_{t+1} e a função de valor estimada $V\pi(s_{t+1})$ para o próximo estado ao invés do valor real de retorno R_t como no método de Monte Carlo, executando a atualização imediatamente após cada passo. Com estas condições, nos métodos de Diferenças Temporais a Equação 3.10 converte-se na Equação 3.16.

$$V\pi_{st} \leftarrow V\pi_{st} + \alpha[r_{t+1} + \gamma V\pi(s_{t+1}) - V\pi(s_t)] \quad (3.16)$$

onde o objetivo para atualização é o valor $r_{t+1} + \gamma V\pi(s_{t+1}) - V\pi(s_t)$ que precisamente define a diferença no tempo t e $t + 1$, característica esta que neste método recebe o nome de Diferenças Temporais. Como a atualização é feita a partir do estado seguinte, os métodos DT são conhecidos como métodos *single-step*.

Predição DT para estimar V^{π}

Inicializar ***V(s)*** de forma arbitraria, e π (política a ser avaliada)
Repete:
    Inicializar ***s***
    Repete (para cada passo do episodio)
        $a \leftarrow$ ação dada por π para ***s***
        Tomar a ação ***a***, observar retorno ***r*** e próximo estado ***s'***
        $V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)]$
        $s \leftarrow s'$
    Até ***s*** ser o estado final

Vantagens dos Métodos de Predição DT

A vantagem mais notável do método DT é a relacionada com o método de Programação Dinâmica, onde esta não necessita da modelagem completa do PDM do ambiente, de seus reforços e das distribuições de probabilidade das transições dos seus estados.

A vantagem seguinte diz respeito ao método de Monte Carlo, visto que DT pode ser implementado de forma totalmente incremental para aplicações On-Line; o método de Monte

Carlo deve aguardar até o final de um episódio para obter o retorno verdadeiro, enquanto DT só necessita aguardar até o estado seguinte. Em aplicações em que os ambientes são definidos como sendo contínuo, o conceito de episódio não é aplicável com facilidade.

Embora as atualizações das funções valor não sejam feitas a partir de reforços reais, mas de valores supostos, é garantida a convergência até a resposta correta. Tanto em DT como em MC a convergência às predições corretas tem forma assintótica. Dentre os dois métodos, algum deles deve convergir mais rápido; a resposta ainda não é dada formalmente, uma vez que até este momento não existe uma demonstração matemática de qual dos métodos é o mais rápido. Mesmo assim, é mostrado experimentalmente que os métodos DT são mais rápidos para tarefas estocásticas (Sutton & Barto, 1998).

3.2.5 Q-learning

Um dos maiores avanços na área de AR foi o desenvolvimento de um algoritmo baseado em Diferenças Temporais que dispensa a política, (off-policy methods) conhecido como Q-learning. A versão mais simples, *One-step Q-learning* (Watkins & Dayan, 1992), é definida pela Equação 3.17:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\alpha[r_{t+1} +\gamma \ max\, Q(s_{t+1},a)-Q(s_t,a_t)] \tag{3.17}$$

onde a função de valor-ação $Q(s_t,a_t)$ é atualizada a partir do seu valor atual, o reforço imediato r_{t+1}, e a diferença entre a máxima função valor no estado seguinte (encontrando e selecionando a ação do estado seguinte que a maximize), menos o valor da função valor-ação no tempo atual. O fato de selecionar a ação que maximize a função valor no estado seguinte permite achar de uma forma simples a função valor-ação estimada.

Uma característica do Q-learning é que a função valor-ação Q aprendida, aproxima-se diretamente da função valor-ação ótima Q^* sem depender da política que está sendo utilizada. Este fato simplifica bastante a análise do algoritmo e permite fazer testes iniciais da convergência. A política ainda mantém um efeito ao determinar quais pares estado-ação devem ser visitados e atualizados, porém, para que a convergência seja garantida, é necessário que todos os pares estado-ação sejam visitados continuamente e atualizados, por isso Q-learning é um método dito *off-policy*.

Algoritmo Q-learning

Inicializar ***Q(s,a)*** de forma arbitraria
Repete (para cada episódio):
    Inicializar ***s***
    Repete (para cada passo do episodio)
        Escolher ***a*** para ***s*** usando política obtida dado Q (p. e. ε-gulosa)
        Tomar a ação ***a***, observar retorno ***r*** e próximo estado ***s'***
        $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a' \in A(s)} Q(s',a') - Q(s,a)\right]$
        $s \leftarrow s'$
    Até ***s*** ser o estado final

A política ε-gulosa é definida no algoritmo pela escolha da ação que possui o maior valor esperado, com a probabilidade definida por (1 − ε), e de ação aleatória, com probabilidade ε. Este processo permite que o algoritmo explore o espaço de estados e esta é uma das condições necessárias para garantir que algoritmos RL encontrem a ação ótima. Q-learning foi o primeiro método AR a possuir fortes provas de convergência. É uma técnica muito simples que calcula diretamente as ações sem avaliações intermediárias e sem uso de modelo.

Dados os valores *Q*, existe uma política definida pela execução da ação *a*, quando o agente esta em um estado *s*, que maximiza o valor *Q(s,a)*. A convergência do algoritmo Q-learning não depende do método de exploração usado. Um agente pode explorar suas ações a qualquer momento, não existem requisitos para a execução de ações estimadas como as melhores. No entanto, para melhorar o desempenho do sistema é necessária, durante o aprendizado, a busca das ações que maximizam o retorno.

Resumidamente, pode-se enumerar os mais importantes aspectos do algoritmo Q-learning:

- O objetivo do uso do algoritmo Q-learning é achar uma regra de controle que maximize cada ciclo de controle;
- O uso do reforço imediato é indicado sempre que possível e necessário, desde que ele contenha informação suficiente que ajude o algoritmo a achar a melhor solução;
- Q-learning é adotado quando o número de estados e ações a serem selecionados é finito e pequeno.

3.3 Considerações Finais

Neste capítulo foi apresentado que os problemas em Aprendizagem por Reforço são caracterizados por um agente que deve aprender comportamentos através de interações de tentativa e erro em um ambiente dinâmico, ou seja, se uma ação desse agente é seguida de estados satisfatórios, ou por uma melhoria no estado, então a tendência para produzir esta ação é reforçada. Foi visto também que AR não é definido como um conjunto de algoritmo de aprendizagem, mas como uma classe de problemas de aprendizagem e que todo algoritmo que resolver bem esse problema será considerado um algoritmo de AR.

Apresentou-se a seus fundamentos matemáticos, através da Propriedade de Markov e do Processo de Decisão Markoviano que é quando uma tarefa de AR satisfaz as Propriedades de Markov. Foi visto que a AR dispõe de vários métodos de aprendizagem, apresentaram-se as características desses métodos. Foi apresentado o algoritmo Q-learning, algoritmo adotado no estudo dessa dissertação, por ter uma facilidade de aproximação da função ótima sem depender da política que está sendo utilizada, simplificando a analise e permitindo testes iniciais de convergência.

Capítulo 4 - Modelagem do algoritmo de CACC proposto

Neste capítulo, serão apresentadas as características levadas em conta no momento da modelagem do sistema de comunicação e da solução de CACC proposta. Estas informações são importantes para a avaliação do algoritmo de CACC e para que o mesmo possa ter bom desempenho, visando melhorar a qualidade de serviço prestada aos usuários da rede e, ao mesmo tempo, proporcionar maior rendimento ao operador de rede.

4.1 Considerações de projeto

Para utilizar o algoritmo de AR no problema de CACC, é necessário identificar os estados, as ações e os reforços do sistema. O agente de AR é o próprio algoritmo de CACC, que desempenha a decisão de aceitação ou não de uma chamada, e que, consequentemente, mudará ou não o estado do sistema. O ambiente corresponde a todo o restante de parâmetros do sistema, compreendendo as tecnologias disponíveis, as chamadas em curso, etc.

O sistema em consideração é uma rede sem fio heterogênea composta por T RATs (*Radio Access Technologies* – Tecnologias de Acesso a Rádio), numeradas de 1 até T, com capacidade de largura de banda (taxa de bits) finita de B_t unidades de largura de banda (ulb) cada, onde a capacidade total da RH é a soma das larguras de banda das redes componentes, conforme mostrado na Equação 4.1. Estas redes podem suportar I classes de serviço representando aplicações multimídia como VoIP (*Voice over Internet Protocol* – Voz sobre protocolo de Internet), fluxos de vídeo, dentre outros, caracterizados por transmissão com taxa fixa de dados.

$$B = \sum_{t=1}^{T} B_t \tag{4.1}$$

O significado físico para unidade de recurso de rádio (como *time slot*, sequência de código, etc.) – é dependente da implementação tecnológica específica feita na interface de rádio (Pla *et. al.*, 2004). Entretanto, não importa qual tecnologia de acesso múltiplo (FDMA, TDMA ou CDMA) é usada, pode-se interpretar a capacidade do sistema em termos de largura de banda efetiva ou largura de banda equivalente (Kesidis *et. al*., 1993) (Nasser & Hassanein, 2004). Portanto, toda vez que referir-se a largura de banda de uma chamada, significa dizer o número de *ulb* (unidade de largura de banda) que é adequado para garantir a qualidade desejada para esta chamada.

Para fins da modelagem de Markov, é considerado que as chamadas chegam ao sistema seguindo processos de Poisson mutuamente independentes com taxas $\lambda_1, \lambda_2, ..., \lambda_I$, onde λ_I é a taxa média de chegada de chamadas à rede para a classe de serviço *I*. O tempo de serviço destas classes são variáveis aleatórias exponencialmente distribuídas com parâmetros $\mu_1, \mu_2, ..., \mu_I$, respectivamente.

Se uma chamada da classe *I* chega à rede heterogênea e é aceita, então será reservada para ela uma taxa fixa de b_i *ulb*. Assim, pode-se modelar a largura de banda total ocupada na rede heterogênea através da Equação 4.2, onde Bo_t é a largura de banda utilizada na tecnologia *t* (calculado através da Equação 4.3), e $n_{t,i}$ é o numero de chamadas em curso da classe *i* na tecnologia *t*.

$$Bo = \sum_{t=1}^{T} Bo_t \tag{4.2}$$

$$Bo_t = \sum_{i=1}^{T} n_{t,i} * b_i \tag{4.3}$$

4.2 Espaço de estados

Mais formalmente, o sistema de alocação de recursos é modelado como um Processo Semi Markoviano de Decisão (Govasi, 2003), cujo espaço de estados é dado pela Equação 4.4.

$$\Phi = \left\{ \left(M_{t,i}, e\right) : \sum_{i=1}^{I} n_{t,i} * b_i \le B_t \forall\, 1 < t \le T; 0 < i \le I \right\} \tag{4.4}$$

Onde $M_{t,i}$ é uma matriz contendo o numero de chamadas em curso de todas as classes de serviço em todas as tecnologias componentes da RH, e *e* é o último evento acontecido, que identifica as chegadas ou partidas de chamadas para cada classes de serviço do sistema.

4.3 Ações possíveis

No momento da chegada de uma chamada, o CACC pode executar as ações de aceitação ou rejeição de chamada, e definir qual tecnologia receberá esta chamada. No encerramento das chamadas nenhuma ação é requerida. A Tabela 4.1 ilustra as ações possíveis no momento da chegada de uma chamada em uma RH com 3 tecnologias.

Tabela 4.1 – Tabela de ações possíveis no momento da chegada de uma chamada, em uma rede com 3 tecnologias

Nº da ação	Descrição
0	rejeitar chamada
1	aceitar chamada na tecnologia 1
2	aceitar chamada na tecnologia 2
3	aceitar chamada na tecnologia 3

Pode-se observar que, o número de ações é igual ao número de tecnologias acrescido de 1, onde o número da ação escolhida corresponde ao número da tecnologia que receberá a chamada, e a ação 0 (zero) indica a rejeição da chamada pela rede heterogênea.

4.4 Função reforço

Um dos fatores mais importantes na resolução de um problema através de AR é a definição da função de reforço. Na abordagem proposta, foram geradas duas funções de reforço para cada par *rede-classe de serviço*, a primeira para calcular o valor de reforço recebido por aceitar uma chamada que chega e a segunda para calcular o valor de reforço recebido por rejeitar a mesma chamada, conforme também usado em Mignanti *et. al.* (2009).

Vale ressaltar que, apesar da solução proposta ter semelhança com a abordagem de Migananti *et. al.* (2009), as duas diferem em alguns aspectos. Em primeiro lugar, o foco é diferente. Enquanto o trabalho de Mignanti foca em redes de próxima geração e convergentes, o CACC proposto tem foco em redes sem fio heterogêneas abrangendo mais de uma tecnologia no mesmo mecanismo de gerenciamento. Outro fator, é que a solução de Mignanti é apresentada em modelo, mas a validação é feita em termos de bits, kilobits e/ou megabits

por segundo, enquanto que a solução proposta aqui usa o conceito de ulb, permitindo que o modelo possa ser adaptado, independente do meio físico e da unidade usada como medida de tráfego.

Além disso, a solução de Mignanti produz resultados e propõe que o treinamento seja feito online, o que aumenta consideravelmente os requisitos de recursos computacionais e pode ser tendencioso no momento que estiver executando continuamente por muito tempo porque embora reflita o comportamento em longo prazo, pode não refletir o que acontece em momentos em determinadas horas do dia, como é o caso das horas de pico de utilização das redes, enquanto que a apresentada aqui é feita através de treinamento *off-line*, permitindo que sejam aferidas as características da rede para cada momento e geradas políticas específicas.

As funções geradas são tratadas separadamente cujo resultado é um valor baseado na largura de banda utilizada na rede, nas taxas médias de chegada de chamadas, no tempo médio de duração dessas chamadas e em um preço atribuído às classes de serviço suportadas.

As duas funções são mostradas em detalhes nas Equações 4.5 e 4.6. A função $rA_{t,i}$ calcula o reforço recebido pelo CACC por aceitar uma chamada da classe *i* na tecnologia *t*, enquanto a função $rR_{t,i}$ calcula o reforço recebido por rejeitar a mesma chamada nesta tecnologia.

$$rA_{t,i}\left(Bo_t, \vec{\lambda}, \vec{\mu}, \vec{\rho}\right) = f\left(Bo_t\right) - \Delta_i\left(\vec{\lambda}, \vec{\mu}, \vec{\rho}\right) \tag{4.5}$$

$$rR_{t,i}\left(Bo_t, \vec{\lambda}, \vec{\mu}, \vec{\rho}\right) = f(0) - rA_{t,i}\left(Bo_t, \vec{\lambda}, \vec{\mu}, \vec{\rho}\right) \tag{4.6}$$

Nas duas equações, o valor da função *f(.)* é a contribuição ligada à largura de banda, e é calculada de acordo com a Equação 4.7; *Δi(.)* é a contribuição de inversão para a classe de serviço *i* e é calculada conforme a Equação 4.8; Bo_t é a largura de banda ocupada na rede *t*.

O vetor $\vec{\lambda} = (\lambda_1, ..., \lambda_I)$ contém as taxas médias de chegadas de chamadas de todas as classes de serviço, $\vec{\mu} = (\mu_1, ..., \mu_I)$ é o vetor que contém os tempos médios de duração das chamadas para cada classe de serviço; e $\vec{\rho} = (\rho_1, ..., \rho_I)$ é um vetor que contém os preços atribuídos a cada classe de serviço e pode ser baseado na largura de banda ocupada por uma chamada desta classe, no tempo de uso do canal, ou em outros parâmetros que podem ser definidos a critério do operador de rede.

A Equação 4.7 apresenta como o cálculo é realizado para garantir que o agente aceite uma chamada somente se houver largura de banda disponível suficiente para comportá-la, caso contrario a chamada será rejeitada. Neste caso, o valor de B_t representa a largura de banda total suportada pela tecnologia e *B0* é um parâmetro de ajuste para a curva de aceitação.

$$f(Bo_t) = \frac{1}{1 + e^{(Bo_t - B_t)/B0}} \tag{4.7}$$

Na Figura 4.1 é mostrada também a importância de *B0* em um par de funções (aceitação e rejeição) para uma classe genérica de serviço e mostrando sua influência na curva desta função. Nota-se que dependendo do valor de *B0*, a curva da função é mais acentuada ou não, sendo que o ponto 0,5 sempre é atingido para $Bo_t/B_t=1$, quando a banda ocupada corresponde a toda a capacidade da rede. Este ponto é chamado de ponto de inversão, pois é onde o reforço por aceitar determinada chamada é o mesmo por rejeitar esta chamada ($rA_{t,i}=rR_{t,i}$): antes dele, o reforço por aceitar as chamadas que chegam será maior, depois dele, prevalece o reforço por rejeitar chamadas nesta tecnologia. A função com maior valor irá determinar qual ação será tomada pelo agente - aceitação ou rejeição da chamada que chega.

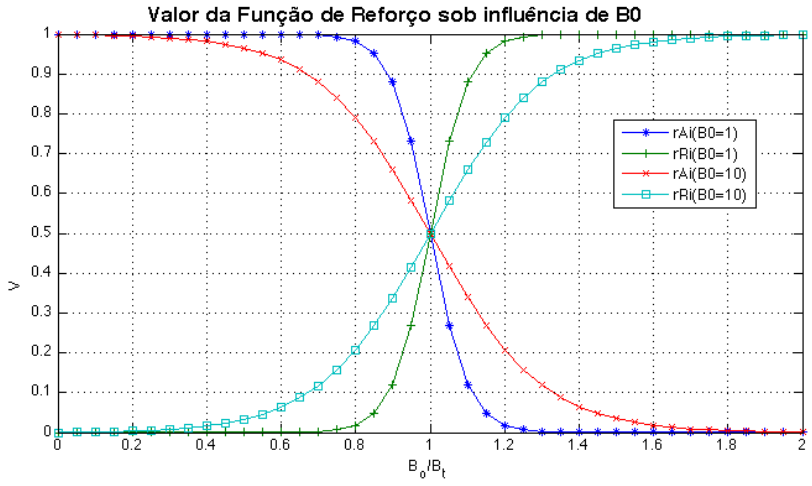

Figura 4.1 – Influência de B0 na curva das funções de aceitação e rejeição de chamadas. (Mignanti *et. al.*, 2009)

O segundo termo na Equação 4.5, termo de inversão, vêm da necessidade do operador de rede maximizar seu rendimento em longo prazo. O valor de $\Delta i(.)$, calculado através da Equação 4.8, deve ser tão alto quanto a classe de serviço for considerada inconveniente pelo operador de rede, e mais baixo quanto a classe for considerada apropriada.

$$\Delta_i = c \cdot g_i(.) \tag{4.8}$$

Neste caso, *c* é um parâmetro livre que pode ser usado para obter um melhor ajuste do algoritmo, de acordo com a necessidade do operador de rede, e $g_i(.)$ é a chamada função de inversão. Ela leva em conta a conveniência em aceitar outras classes de serviço em relação à classe *i* e é calculado através da Equação 4.9.

$$g\left(\vec{\lambda},\vec{\mu},\vec{\rho}\right) = \frac{\sum_{k\neq i}^{N} \lambda_k \cdot \mu_k \cdot \rho_k}{\sum_{k=1}^{N} \lambda_k \cdot \mu_k \cdot \rho_k} \tag{4.9}$$

Na Figura 4.2 são apresentados gráficos que demonstram o efeito do valor de Δ no resultado das funções de aceitação e rejeição de chamadas, alterando o ponto de inversão e a diferença de valores entre as duas funções. É possível observar que, dependendo do valor de Δ, o ponto onde o valor por rejeitar uma chamada é maior que o valor por aceitar esta chamada seja antecipado, fazendo, com isso, uma reserva de banda para as chamadas que são priorizadas no sistema. Assim, de acordo com a função de inversão, certa classe de serviço *i* é mais conveniente se:

- O seu preço é maior;
- A sua freqüência é maior (λ);
- A sua duração é maior (μ);

O parâmetro preço é algo intuitivo na decisão de aceitação ou não de chamadas, pois quanto maior for ele, melhor será o rendimento do operador de rede, o que faz com que as chamadas que são mais bem taxadas sejam priorizadas em relação às outras.

Por outro lado, o parâmetro de taxa média de chegada de chamadas foi escolhido por indicar que é mais rentável para o operador aceitar as chamadas que chegam com maior frequência. Além disso, os usuários que requisitam esse tipo de chamada são maioria o que contribui para melhorar a satisfação dos usuários.

O parâmetro duração média de chamada foi adicionado para indicar ao operador de rede que é mais importante aceitar as chamadas que duram mais, pois estas vão gerar maior rendimento no longo prazo, principalmente se a tarifação for feita por tempo de uso do canal e, além disso, esse tipo de chamada é efetuado pelos usuários que utilizam a rede por mais

tempo e, priorizando esses usuários, a rede será melhor avaliada por esse tipo de usuários. A lógica é fazer um produto pelas três características aferidas da rede e, a partir daí, gerar um valor que possa medir a importância de aceitar ou não uma classe de chamadas.

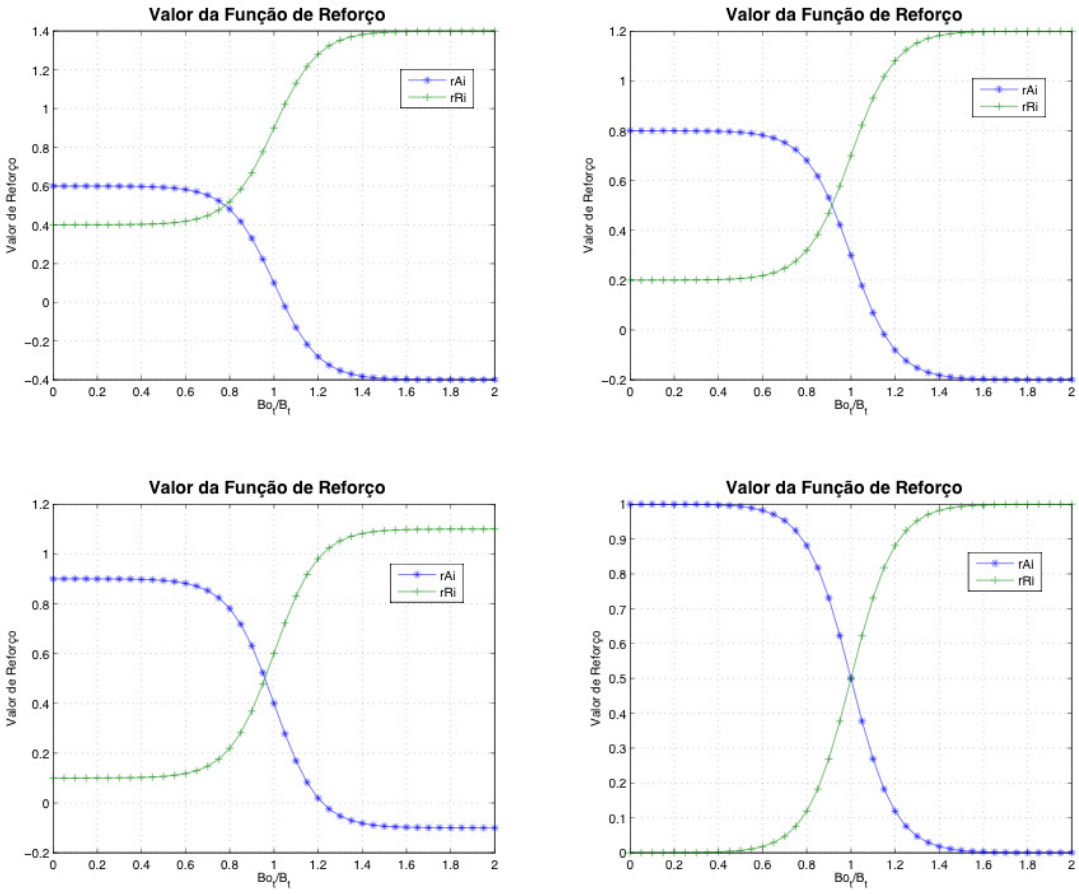

Figura 4.2 – Funções de reforço para Δi=0.4, Δi=0.2, Δi=0.1 e Δi=0.

Assim, através das duas funções de reforço – de aceitação e de rejeição – é possível traçar uma política onde o CACC, que gerencia a rede heterogênea, contenha a melhor ação a ser tomada de modo a beneficiar o operador de rede, mas levando também em consideração os anseios dos usuários. É importante notar que nessas funções existem parâmetros de ordem do operador de rede (preço), de ordem da rede em si (ocupação), e de características dos usuários que utilizam esta rede (taxa média de chegada e tempo médio de duração das chamadas), o que faz com que sejam balanceados parâmetros de interesse conflitantes, como é o caso do rendimento do operador e qualidade do serviço prestado ao usuário, para se chegar a um limiar que possa atender a ambos.

4.5 Treinamento, operação e implementação do algoritmo de CACC proposto

A solução proposta nesse trabalho é a elaboração de um mecanismo de alocação de chamadas e gerenciamento de recursos em redes heterogêneas, que foi desenvolvido em duas fases: treinamento e operação; e são descritas a seguir.

De acordo com o diagrama da Figura 4.3, no início são inseridos os dados de entrada, que são os parâmetros do algoritmo de treinamento, parâmetros da rede, parâmetros do modelo e parâmetros de simulação. De posse desses dados, então é realizado o treinamento do CACC, a partir do qual é gerada a política de decisão de aceitação ou rejeição de chamadas e, de posse dessa política, é então feita a validação do algoritmo através da simulação da rede em operação.

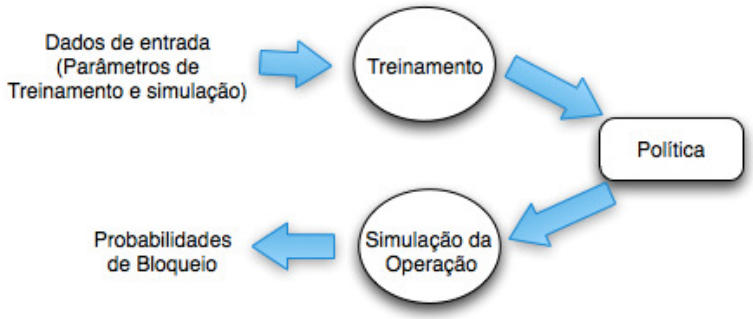

Figura 4.3 – Diagrama de implementação do algoritmo e simulação

4.5.1 Treinamento

Caracterizado por ser desenvolvido baseado no algoritmo de aprendizagem por reforço, o mecanismo de CACC proposto executa a tarefa de treinamento com o objetivo de gerar uma política de decisão que possa ser utilizada na fase de operação do algoritmo. Esse treinamento pode ser feito de forma *on-line* ou *off-line*. Quando o treinamento é feito durante a fase de operação, este é dito *on-line*, e faz com que o agente, mesmo em operação, continue aprendendo sobre o comportamento do sistema. Para a aplicação em redes heterogêneas, significa dizer que o CACC, mesmo em execução na rede, continua a aprender os comportamentos das chamadas, e qual é a melhor ação a ser tomada em determinado momento de acordo com suas funções de reforço.

Já no treinamento *off-line*, o treinamento do algoritmo de CACC é feito anteriormente, a partir de características da própria rede e, a partir dos dados coletados, gera uma política de

decisão a qual usará durante a fase de operação. No caso das RH significa dizer que primeiramente serão obtidas as características da rede e, de posse destes dados, será executado o treinamento para geração de uma política de decisão.

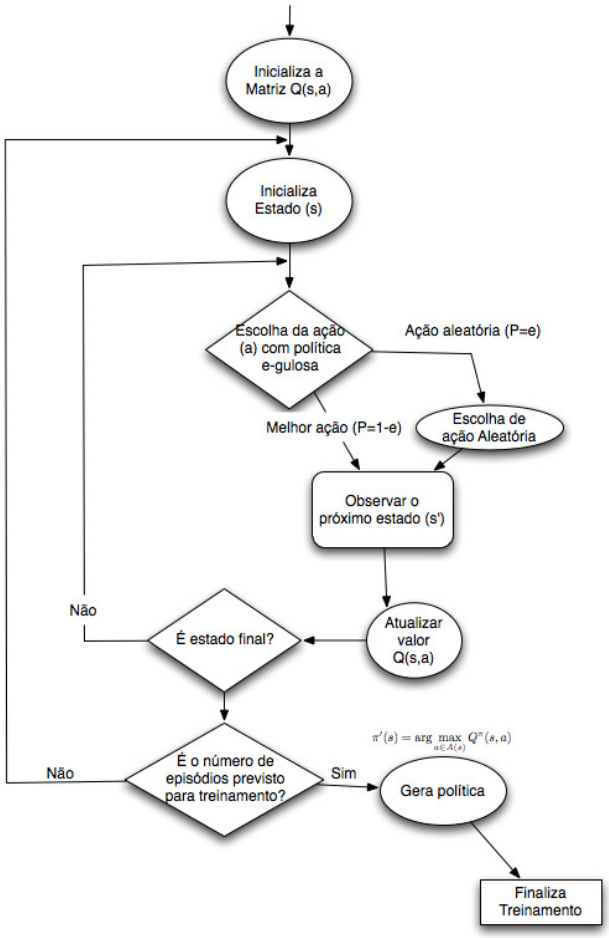

Figura 4.4 – Diagrama do algoritmo de treinamento do Q-Learning adotado

Como uma rede heterogênea pode ter alterado seu comportamento (número de usuários, número de chamadas, taxa de chegada de chamadas e tempo médio de duração) ao longo do tempo (dia, semana, mês, etc.) de acordo com a região de operação, da classe social, dentre outros fatores que podem influenciar, optou-se por fazer o treinamento *off-line* pois, desta forma, é possível refletir as características da rede para momentos específicos, e treiná-la para isso gerando políticas a ser aplicadas em cada momento específico, dependendo do comportamento da rede. Outro fator que inviabiliza a implementação do algoritmo de modo *on-line* é o fato de que o algoritmo de inteligência computacional necessita ficar em execução no momento da operação da rede, o que aumenta consideravelmente o custo computacional sem falar que, por se tratar de um sistema de tempo real onde o fator tempo é um recurso

crítico o que, neste caso, pode gerar um maior atraso no início de chamadas, afetando assim a percepção do usuário.

Nas Figuras 4.4 e 4.5. é possível observar o diagrama e seu respectivo algoritmo, usado para treinamento do CACC proposto neste trabalho. Nela é dada uma visão geral de como funciona o treinamento do algoritmo de geração da política.

Inicialmente, o algoritmo de treinamento é inicializado com a matriz Q(s,a) contendo valor nulo para todos os pares *estado-ação* possíveis. O estado, conforme definido na Equação 4.4, possui uma matriz $M_{t,i}$ contendo o mapeamento de todas as classes de chamadas nas tecnologias que fazem parte da rede heterogênea. Essa matriz é inicializada com valor zero em todos os seus elementos, significando que não há nenhuma chamada em curso na RH. Também faz parte da variável estado o último evento ocorrido. Este é inicializado estatisticamente como um evento de chegada de chamada, calculado de acordo com as taxas de chegada de chamadas de todas as classes de serviço (conforme Equação 4.10). As ações possíveis para o momento de chegada de chamadas, conforme descrito na Tabela 4.3. são, rejeitar ou aceitar a chamada em uma das tecnologias disponíveis.

1. Inicializa a matriz Q(s,a) como uma rede que não contém nenhuma chamada em curso.
2. Repete (para cada episódio)
3. Inicializar *s*
4. Repete (para cada passo do episódio)
5. Escolher *a* para *s* usando política *e*-gulosa
6. Tomar a ação *a*
7. Observar o próximo estado *s'* e o retorno *r*
8. $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a) \right]$
9. $s \leftarrow s'$
10. Até s ser o estado final
11. Até o número de episódios definido

Geração da política de decisão

12. Repete (para cada estado)
13. Repete (para cada ação possível neste estado)
14. Verifica se é a ação que possui maior valor de Q(s,a);
15. Se for, define como melhor ação para este estado
16. Até todas as ações possíveis para este estado serem visitadas
17. Até todos os pares *estado-ação* serem visitados.
18. Retorna a política com todos os pares *estado-melhor ação* para o CACC

Figura 4.5 – Algoritmo de treinamento do CACC usado na solução proposta.

$$Pch_k = \frac{\lambda_k}{\sum_{i=0}^{I} \lambda_i} \tag{4.10}$$

Em aprendizagem por reforço, um episódio é definido como sendo a sequência de estados que chegam até o estado final (Sutton & Barto, 1998). No caso da aplicação proposta, é o ciclo que leva o agente, do momento em que não possui nenhuma chamada em curso na rede até o momento que a rede heterogênea está totalmente ocupada, não restando recursos disponíveis para comportar novas chamadas. Este último momento é definido como estado final do episódio.

Os passos do episódio são ciclos que correspondem à chegada de chamada e sua aceitação ou rejeição na rede heterogênea. A política ε-gulosa é definida no algoritmo de AR pela escolha da ação que possui maior valor esperado, com probabilidade definida por (1-ε) e de ação aleatória com probabilidade ε (Sutton & Barto, 1998). Matematicamente, dado Q obtemos a ação gulosa a^* para um estado s fazendo, conforme Equação 4.11.

$$\begin{aligned} a^* &= \arg\max_{a \in A(s)} Q(s,a) \\ \pi(s,a^*) &= 1 - \varepsilon + \frac{\varepsilon}{|A(s)|} \\ \pi(s,a) &= \frac{\varepsilon}{|A(s)|}, \forall \ a \in A(s) - \{a^*\} \end{aligned} \tag{4.11}$$

Esta restrição permite que o algoritmo explore o espaço de estados, sendo uma das condições necessárias para garantir que o algoritmo encontre uma política de controle ótima.

Após a execução dos passos da Figura 4.5 para um número de episódios definidos no início da execução do treinamento, é gerada a política de controle que será usada pelo CACC no momento de operação da rede. Isso é feito percorrendo o espaço de estados e armazenando, para cada estado, a melhor ação a ser executada que é a ação que possui maior valor de *Q(s,a)*.

4.5.2 Operação

A operação é a fase onde o CACC esteja desempenhando sua tarefa na rede heterogênea, que é definir se as chamadas que chegam irão ser aceitas ou rejeitas e qual tecnologia irá recebê-las. Nesse caso, é necessária a verificação de qual classe de serviço está

chegando, qual a banda necessária para comportá-la, qual o estado da rede heterogênea como um todo, e também de uma consulta à política de controle gerada pelo algoritmo de AR. De posse do retorno da política, então é tomada a ação indicada. O algoritmo pode ser visto na Figura 4.6.

1.	Chegada da chamada
2.	Verificação de qual classe de serviço pertence à chamada e qual a banda requerida
3.	Verificação do estado da rede heterogênea
4.	Consulta a política de controle para verificar a melhor ação
5.	Se melhor ação for aceitar, então verificar qual a melhor rede e direcionar a chamada
6.	Se melhor ação for rejeitar a chamada, então descarte a solicitação

Figura 4.6 – Algoritmo do CACC em operação numa rede heterogênea

4.5.3 Implementação do algoritmo

Para avaliação de desempenho do algoritmo proposto, foi implementado um cenário de simulação de um sistema de comunicação composto por mais de uma tecnologia, onde cada uma delas possui como característica, a capacidade (em número de ulb). A implementação foi feita usando a linguagem de programação Java, que foi escolhida pois é uma linguagem de programação que possui muitos recursos, relativamente independente de plataforma e permite fazer interface da solução proposta com outras implementações de aprendizagem de máquina como Weka (Witten e Frank, 2000), desenvolvida na mesma linguagem.

Entretanto, para verificar a confiabilidade dos resultados gerados pelo simulador, um algoritmo de CACC baseado em seleção aleatória de tecnologia, foi implementado e comparado a um modelo feito no simulador Arena (Drevna & Kasales, 1994) com a mesma finalidade. Os resultados obtidos foram idênticos, o que atesta a confiabilidade do simulador desenvolvido para avaliação de desempenho do CACC proposto. Assim, verificada a confiabilidade do ambiente desenvolvido, foi então inserida no simulador a política gerada através de aprendizagem por reforço e, então, obteve-se os resultados que serão apresentados no Capítulo 5.

Capítulo 5 - Estudos de Caso e Resultados

Para avaliar a efetividade do CACC proposto é necessário executá-lo em modo de operação em algum cenário real ou simulado. Como aferir em cenário real envolve um conjunto extenso de outras atividades, além de não estar em um meio totalmente controlado, optou-se por criar dois cenários de simulação para verificar o comportamento do algoritmo em uma rede heterogênea. Neste capítulo são apresentadas as descrições dos dois cenários, as configurações de rede, os parâmetros do algoritmo utilizados, a fim de mostrar como o mesmo foi avaliado, quais os resultados obtidos, seus pontos fortes e fracos e quais os benefícios em sua aplicação.

5.1 Cenários de Avaliação

Para fins de avaliação de desempenho, foram desenvolvidos dois cenários de simulação, onde cada um deles é composto de uma rede heterogênea com duas tecnologias distintas: Tecnologias 1 e 2, com capacidade de largura de banda de 88 e 160 ulb, respectivamente; que aceitam duas classes de serviço (1 e 2). Na simulação foram gerados fluxos para requisições de novas chamadas, através de eventos discretos, de acordo com processos de Poisson mutuamente independentes, conforme indicado no modelo do capítulo anterior, e com média de atendimento exponencialmente distribuída. Estes cenários podem ser considerados relativamente simples, embora acredite-se que o objetivo principal seja atingido que é a validação da proposta. Além disso, a ideia é que os cenários sejam testados de forma mais simples e depois sejam extrapolados. Com os cenários atuais geraram um total de 112640 estados. As classes de serviço foram pensadas de modo a indicar qual classe seria priorizada e, após a execução, validar se o que foi projetado aconteceria na prática.

O algoritmo proposto foi comparado a um CACC guloso que aceita chamadas enquanto possui largura de banda disponível na rede heterogênea e faz a seleção de tecnologia aleatoriamente. Duas métricas foram usadas para avaliação de desempenho do algoritmo

proposto: a primeira é a probabilidade de bloqueio de chamadas para as duas classes de serviço, calculada pela razão das chamadas bloqueadas em relação ao número total de chamadas que chegaram ao sistema e rendimento; a segunda é o rendimento da operadora em cada uma das redes, que é calculado tendo em vista o preço por aceitar uma chamada de determinada classe.

Foram criados dois cenários de simulação A e B. No cenário A, foram simuladas duas classes de serviço 1 e 2, onde uma é superior à outra tanto no preço, na duração e na taxa de chegada. Assim espera-se que esta possa apresentar melhor desempenho no decorrer da simulação. No cenário B, a taxa de chegada de chamadas da classe 2 é aumentada dez vezes, o que aumenta consideravelmente a ocupação da rede. Ambos os cenários visam avaliar o desempenho do algoritmo, e ver o comportamento da rede. Na Tabela 5.1 são apresentados os dados referentes a cada um dos cenários.

Tabela 5.1 – Tabela de parâmetros dos cenários testados

Parâmetros	**Cenário A**		**Cenário B**	
	Classe 1	Classe 2	Classe 1	Classe 2
Largura de Banda (ulb)	8	1	8	1
Duração média (em segundos)	5400	120	5400	120
Preço	8	1	8	1
Taxa de chegada (chamadas por segundo)	De: 0,00027 a 0,04166	0,00278	De: 0,00027 a 0,04166	0,0278

Para estes cenários, os eventos possíveis de acontecerem, conforme modelo apresentado no capítulo 4, são apresentados na Tabela 5.2.

Tabela 5.2 – Eventos possíveis em uma rede com 2 tecnologias e 2 classes de serviço

Nº do Evento	**Descrição do Evento**
0	Chegada de uma chamada da classe 1 na tecnologia 1
1	Chegada de uma chamada da classe 1 na tecnologia 2
2	Chegada de uma chamada da classe 2 na tecnologia 1
3	Chegada de uma chamada da classe 2 na tecnologia 2
4	Encerramento de uma chamada da classe 1 da tecnologia 1
5	Encerramento de uma chamada da classe 1 da tecnologia 2
6	Encerramento de uma chamada da classe 2 da tecnologia 1
7	Encerramento de uma chamada da classe 2 da tecnologia 2

Para a mesma rede, na Tabela 5.3 é apresentado, o conteúdo da matriz $M_{t,i}$, em um determinado momento, mostrando o número de chamadas em curso de cada classe nas duas tecnologias. Nesse caso, existem 3 chamadas da classe 1 na tecnologia 1 e 13 chamadas da mesma classe na tecnologia 2. Por sua vez, existem 5 chamadas da classe 2 na tecnologia 1 e 7 na tecnologia 2.

Tabela 5.3 – Exemplo da matriz $M_{t,i}$ para uma rede heterogênea com 2 tecnologias e 2 classes de serviço

Tecnologia	Número de chamadas em curso (Classe 1)	Número de chamadas em curso (Classe 2)
1	3	5
2	13	7

Através das Tabelas 5.2 e 5.3, é possível observar que o número de ações possíveis em uma rede heterogênea, para o algoritmo proposto, depende da quantidade de classes de serviços suportadas e do número de tecnologias componentes desta rede. Além disso, também é possível notar que o agente (CACC) tem a todo momento o controle da rede heterogênea como um todo, possuindo um registro de quantas chamadas existem em curso em cada uma das tecnologias e a quais classes pertencem, através da matriz $M_{t,i}$.

O tempo de decisão é sempre o momento de chegada de qualquer chamada (classe 1 e 2). No momento do encerramento de chamadas nenhuma decisão é requerida, caracterizando-se apenas como um evento do sistema. Além disso, as possíveis ações, para esta rede, no momento da chegada de chamadas são:

- 0: rejeitar chamada;
- 1: aceitar a chamada na rede 1;
- 2: aceitar a chamada na rede 2;

Baseado nos dados citados, foi implementado o algoritmo de aprendizagem por reforço, para obter uma política de admissão de chamadas e seleção de tecnologia para criação do CACC. Com a política gerada pelo algoritmo, foi simulado um sistema de comunicação, com chegada de chamadas, treinados em 50000 (cinquenta mil episódios), que produziram os resultados das Figuras 5.1 e 5.2, mostrando o desempenho da rede em termos

de probabilidade de bloqueio de chamadas para o CACC baseado em AR e para o CACC guloso baseado em seleção aleatória (SA).

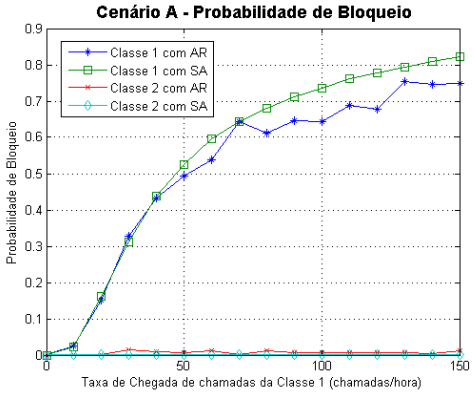

Figura 5.1 – Probabilidade de Bloqueio para o cenário A

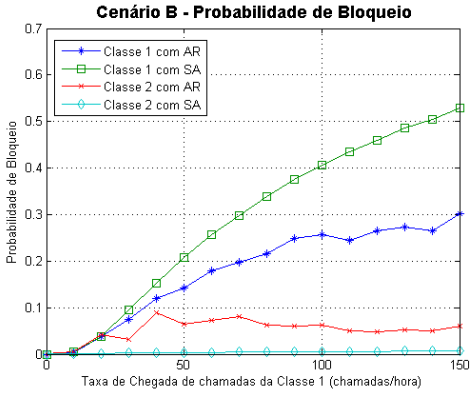

Figura 5.2 – Probabilidades de Bloqueio para o cenários B

No cenário A (Figura 5.1), onde a taxa de chegada de chamadas da classe 2 é menor (0,00278 chamadas por segundo ou 10 chamadas por hora), nota-se que a probabilidade de bloqueio da classe de serviço 1 no algoritmo baseado em AR desempenha melhor resultado à medida que a sua taxa de chegada aumenta, causando mais intrusão na rede, em relação ao algoritmo de CACC baseado em SA; o que sugere que em uma rede ociosa, ambos os algoritmos executam bem, visto que a probabilidade de bloqueio é muito baixa. Quando o foco da análise é a classe de serviço 2, ambos os algoritmos tem desempenho similar, demonstrados pela quase imperceptível variação de um para outro.

No cenário B, onde a taxa de chegada de chamadas da classe 2 é dez vezes maior (0,0278 chamadas por segundo ou 100 chamadas por hora) do que no cenário A, existe um aumento da intrusão da rede e da importância desta classe 2 no momento da decisão. Nesse cenário, o desempenho dos algoritmos é diferenciado. Neste caso, como no cenário A, a

diferença de desempenho aparece à medida que a taxa de chegada de chamadas da classe 1 é aumentada, entretanto, essa diferença é mais evidente do que no cenário A, pois a rede se torna mais ocupada.

Para a classe 2, os resultados mostram que o CACC guloso apresenta sempre melhor desempenho do que o CACC baseado em AR. Isso acontece porque o algoritmo baseado em AR dá prioridade ao tráfego da classe 1, devido sua taxa de chegada, tempo de duração e preço.

Nas Figuras 5.3 e 5.4 é possível observar os resultados de taxa de utilização de rede para os cenários A e B, respectivamente. Pode-se notar que o CACC proposto tem sempre melhor desempenho comparado ao CACC guloso. Isto acontece porque ele reserva uma parte da capacidade de largura de banda disponível na rede para as chamadas priorizadas e, consequentemente, diminui sua probabilidade de bloqueio. Essa reserva é feita porque o algoritmo dá mais importância àquelas classes que possuem maior taxa de chegadas, maior tempo de duração e maior preço atribuído pelo operador de rede.

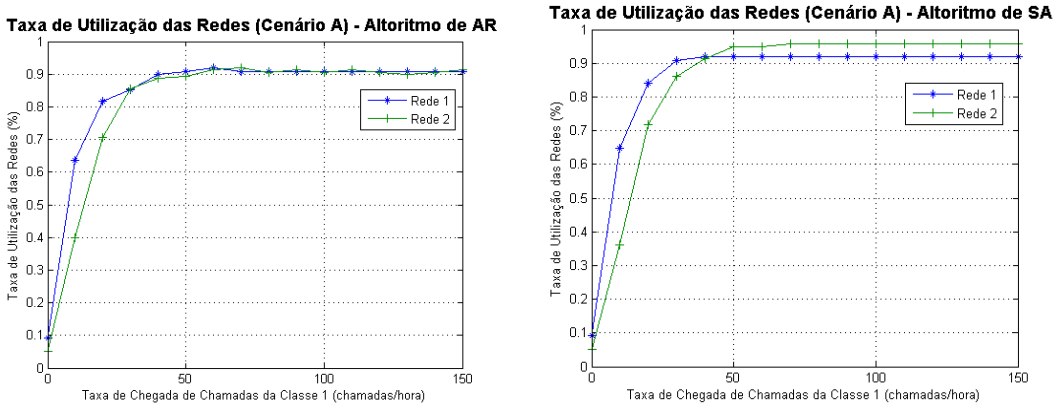

Figura 5.3 – Taxa de utilização das redes no Cenário A

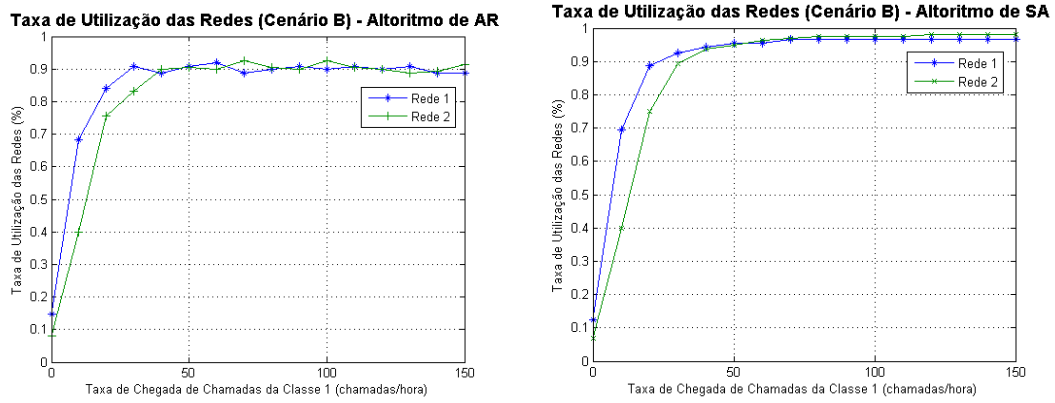

Figura 5.4 – Taxa de utilização das redes no Cenário B

Nas figuras 5.5 e 5.6 são apresentados os resultados de rendimento das Redes 1 e 2, dos algoritmos de CACC baseados em SA e em AR, para os cenários A e B. Através deles, é possível notar que, em termos de rendimento, à medida que a rede se torna ocupada (quando chegam mais chamadas da classe 1), o algoritmo proposto continua aumentando o rendimento, mesmo com alto grau de ocupação das redes, o que não acontece com os resultados apresentados pelo algoritmo de CACC guloso que vai aumentando seu rendimento até chegar um ponto onde atinge seu limite não podendo mais aumentá-lo. Isso acontece pois a reserva de banda feita pelo algoritmo apresentado permite que mais chamadas sejam aceitas.

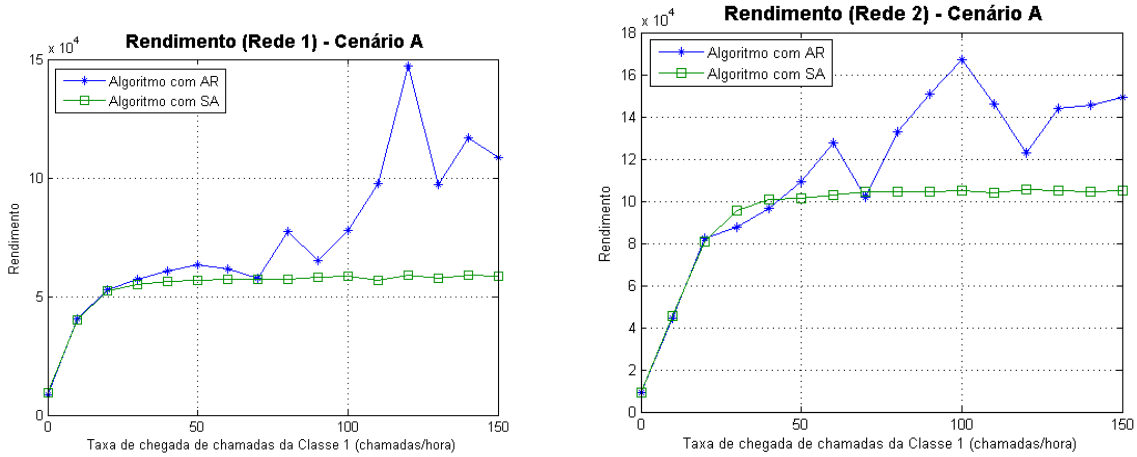

Figura 5.5 – Rendimento das Redes 1 e 2 para o Cenário A

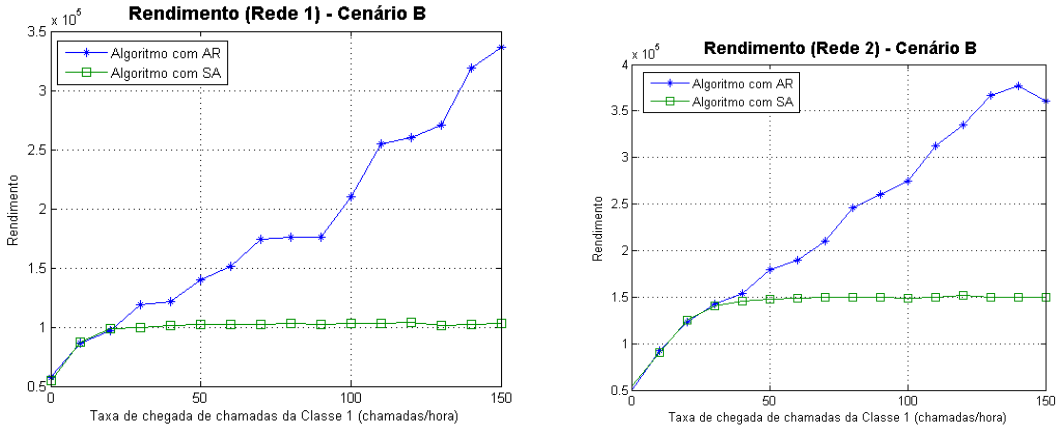

Figura 5.6 – Rendimento das Redes 1 e 2 para o Cenário B

Assim, diante dos resultados apresentados, conclui-se que, em condições onde a rede não está muito ocupada, ambos os algoritmos apresentam resultados satisfatórios, visto que a rede tem recursos suficientes para comportar a quantidade de chamadas que chega. Entretanto, à medida que a rede vai se tornando ocupada e sobrecarregada, os resultados em

termos de probabilidade de bloqueio e rendimento do algoritmo de CACC baseado em AR se mostram melhores que os do CACC baseado em SA, mantendo sempre níveis menores de probabilidade de bloqueio para as classes priorizadas, e aumentando os índices de rendimento do operador de rede, mesmo em condições de ocupação extrema da rede. Através dos gráficos de utilização é possível observar também que esses resultados acontecem pois o CACC proposto reserva uma parte da largura de banda para as classes de maior prioridade. Os gráficos de rendimento corroboram que do ponto de vista do operador o CACC proposto mantém os níveis de rendimento aumentando mesmo com a rede em condições de grande ocupação o que justifica sua utilização e permite observar sua utilidade, ao mesmo tempo que também leva em consideração características de usuário.

6. Considerações finais e trabalhos futuros

O futuro das comunicações sem fio está nas redes heterogêneas, que integram várias tecnologias distintas a trabalhar de modo conjunto cooperando entre si visando maior satisfação dos usuários da rede, melhor utilização dos recursos disponíveis e permitindo maior qualidade de serviço. Um dos pontos chaves nesse contexto é o desenvolvimento de mecanismos que gerenciem os recursos de rede conjuntamente. O controle de admissão de chamadas é um mecanismo de gerenciamento de recursos que decide se uma chamada é aceita em uma rede ou não, e em RHSF, ele deve decidir também qual tecnologia receberá uma chamada que se inicia.

Este trabalho apresentou uma proposta de controle de admissão de chamadas conjunto baseado no método de inteligência computacional conhecido como Aprendizagem por Reforço. Este método tem sido usado no contexto de rede sem fio, inclusive na área de gerenciamento de recursos e têm se mostrado eficaz. A abordagem proposta se utilizada de parâmetros da própria rede como base para tomada de decisão como largura de banda utilizada pela chamada, taxa média de chegada de chamadas, tempo médio de duração e preço atribuído a cada classe dessas chamadas para a tomada de decisão.

Os resultados apresentados mostram que o desempenho do algoritmo é satisfatório frente a algoritmos que gerenciam redes sem fio sem uma política de seleção de tecnologia definida, realizando esta tarefa de forma aleatória. Também mostram que, à medida que a rede heterogênea se torna mais ocupada, melhor é o desempenho em relação ao método de comparação e melhor é o desempenho em relação ao rendimento do operador de rede.

Esta proposta é importante pois, embora seja relativamente simples, permite extrapolar cenários e simular implantações de redes futuras de modo a verificar o funcionamento da mesma. Além disso, cenários podem ser estimados e simulados também para expansão de capacidade da rede e análise de comportamento nesses casos.

Estudos complementares são necessários para fazer uma análise do impacto deste algoritmo quando chamadas de *handoff* chegarem a uma tecnologia específica. Outro ponto que merece novas investigações é a verificação da influência de outros parâmetros, das várias camadas de comunicação para analisar sua influência em resultados desse algoritmo para redes heterogêneas. Outros estudos envolvendo diversos momentos onde os parâmetros usados para gerar a política possam variar (como em horários de pico de utilização), para que o algoritmo possa se adequar à necessidade.

Além disso, também necessita-se fazer, ainda, testes em outros ambientes de simulação que possam reafirmar os resultados aqui apresentados, o que deve requerer mais recursos no momento da implementação do algoritmo e sua comparação com simuladores de rede largamente utilizados como NS (Network Simulator).

Referências Bibliográficas

3rd Generation Partnership Project (3GPP). **About 3GPP**. Disponível em: <http://www.3gpp.org/About-3GPP> Acesso em: 12/05/2012.

ANDREWS, G. Jeffrey; GHOSH, Aranabha; MUHAMED, Rias; **Fundamentals of WiMAX: Understanding Broadband Wireless Networking**, Prentice Hall. 1ª ed. 2007.

CAO, Gen; YANG, Dacheng; ZHU, Xiaoyue; ZHANG, Xin. **A Joint Resource Allocation and Power Control Algorithm for Heterogeneous Network**. Em: 19th International Conference on Telecommunications (ICT), 2012.

DIAS, Ugo Silva. **Evolução das Redes de Comunicação Móveis**. 1º Encontro Regional de Telecomunicações, 2010. Disponível em: <http://mwsl.unb.br/index.php/pt/downloads/category/11-externas> Acesso em: 30 de abril de 2012.

DREVNA, M.; KASALES, C. **Introduction to Arena**. In: Procedings of the 1994 Winter Simulation Conference, Ed. J.D. Tew, M.S. Manivannan, D.A. Sadowski, e A.F. Seila, páginas: 431-436. Institute of Electrical and Electronics Engineers, Piscataway, NJ

FALOWO, Olabisi E.; CHAN, Anthony H. **Joint call admission control algorithms: Requirements, approaches, and design considerations**. Computer communications 31, 2008.

FALOWO, Olabisi E.; CHAN, Anthony H., **Joint Call Admission Control Algorithm for Fair Radio Resource Allocation in Heterogeneous Wireless Networks Supporting Heterogeneous Mobile Terminals.** In: Consumer Communications and Networking Conference (CCNC), 2010.

GOVASI, Abhijit. **Simulation-Based Optimization: parametric optimization techniques and reinforcement learning**. Kluwer Academic Publishers, Estados Unidos, 2003.

GRAY, Doug. **WiMax, HSPA+, and LTE: A Comparative Analysis**. WiMax Forum, 2009.

HASIB, Abdul; FAPOJUWO, Abraham O. **Cross-layer radio resource management in integrated WWAN and WLAN networks**. Computer Networks Journal, 2010.

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. **IEEE 802.16.2: Standard**: Recommended Practice for Local and metropolitan area networks, 2004.

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. **IEEE 802.22 Standard** for Information technology-- Local and metropolitan area networks-- Specific requirements-- Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and

Physical Layer (PHY) specifications: Policies and procedures for operation in the TV Bands, 2011. Disponível em: <http://standards.ieee.org/about/get/802/802.22.html> Acesso em: 28 de Junho de 2012

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. . **IEEE Standard for Local and metropolitan area networks.** Part 20: Air Interface for Mobile Broadband Wireless Access Systems Supporting Vehicular Mobility - Physical and Media Access Control Layer Specification. New York, 2008.

KASSAB, Mohamed; BONNIN, Jean-Marie; BELGHITH, Abdelfettah. **Technology Integration Framework for Fast and Low Cost Handovers – Case Study: WiFi-WiMax Networks**. Journal of Computer Systems and Communications, 2010.

KESIDIS, G.; Warlrand, J.; Chang, C. **Effective bandwidths for multiclass Markov fluides and other ATM sources**. In: IEEE Transactions on Networking, vol. 1, no. 4, Aug. 1993.

KUROSE, James F.; ROSS, Keith W. **Redes de computadores e a Internet: uma abordagem top-down**. Tradução Arlete Simille Marques. 3ª ed. São Paulo, Pearson Addison Wesley, 2006.

LANDSTROM, S., *et. al.* **Heterogeneous networks-increasing cellular capacity**. Ericsson Reveiw, 2011.

LEE, S., *et. al.* **A Probabilistic Call Admission Control Algorithm for WLAN in Heterogeneous Wireless Environment**. In: IEEE Transactions on Wireless Communications, Vol. 8, Nro. 4, April 2009.

MICHAELIS. **Moderno dicionário da língua portuguesa**, Cia. Melhoramentos, 1998.

MIGNANTI, *et. al.* **A Model Based RL Admission Control Algorithm for Next Generation Networks**. In: Eighth International Conference on Networks. March 1 - 6, Cancun, Mexico. 2009.

NASSER, N; HASSANEIN, H. **Adaptive call admission control for multimedia wireless networks with QoS provisioning**. In: Proceedings of the 2004 International Workshop on Mobile and Wireless Networking, Montreal, Canada, August 2004, pp. 30-37.

NOGUEIRA, A. D.; *et. al.* **Integração de redes UMTS e IEEE 802.11 utilizando os protocolos MIPv6 e SIP**. In: Anais do Simpósio Brasileiro de Redes de Comutadores e Sistemas Distribuídos, XXV, Maio de 2007. SBRC 2007

PLA, V.; *et. al.* **Optimal admission control using handover prediction in mobile cellular networks**. In: Proceedings of Second International Working Conference on Performance Modeling and Evaluation of Heterogeneous Networks, West Yorkshire, UK, 2004

QUALCOMM INCORPORATED. **LTE Advanced: Heterogeneous Networks**, 2011. Disponível em: <http://www.qualcomm.com/documents/files/lte-advanced-heterogeneous-networks.pdf>, Acesso em: 28 de Junho de 2012.

RACKLEY, Steve. **Wireless Networking Technology: from principles to successful implementation**. Elsevier, 2007.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência Artificial: tradução da segunda edição**. Rio de Janeiro: Campus, 2004.

SERRA, Maurício R. G. **Aplicações de Aprendizagem por Reforço em Controle de Tráfego Veicular**. Dissertação de Mestrado. Universidade Federal de Santa Catarina, 2004.

SUN, Zhuo; WANG, Wenbo. **Investigation of Cooperation Technologies in Heterogeneous Wireless Networks.** Journal of Computer Systems and Communications, 2010.

SUTTON, Richard S.; BARTO, Andrew G. **Reinforcement Learning: An Introduction**. MIT Press, Cambridge, MA, 1998.

TANENBAUM, Andrew S. **Redes de computadores**. Tradução da 4ª edição americana. Ed. Campus, 2003.

WATKINS, C.; DAYAN, P. **Q-Learning. Machine Learning**, Vol.8: pp.279.292, 1992.

WITTEN, H.; FRANK, E. **Data Mining: Practical machine learning tools with Java implementations**, Morgan Kaufmann. 2000. Software disponível em: http://www.cs.waikato.ac.nz/ml/weka, Acesso em: 10 de Outubro de 2011.

YAU, Kok-Lim Alvin; KOMISARCZUK, Peter; TEAL, Paul D. **Reinforcement learning for context awareness and intelligence in wireless: Review, new features and open issues**. In: Journal of Network and Computer Applications, 2012.

VUCEVIC, Nemanja; *et. al.* **Reinforcement learning for joint radio resource management in LTE-UMTS scenarios**. In: Computer Networks Journal, 2011.