

APLICAÇÃO DO ALGORITMO SARSA NA COLETA DE LIXO - ESTUDO DO PARÂMETRO ALPHA¹

Darlinton Prauchner², Rogério Martins³, Edson Padoin⁴.

¹ Trabalho de conclusão de curso realizado no curso de Ciência da Computação da Unijuí

² Aluno do curso de Ciência da Computação

³ Professor do curso de Ciência da Computação da Unijuí

⁴ Professor do curso de Ciência da Computação da Unijuí

INTRODUÇÃO

Nos últimos anos, de acordo com dados da Organização Panamericana de Saúde (OPAS), o problema da coleta adequada de resíduos sólidos tem aumentado em vários países, principalmente onde existe crescimento populacional e concentração desta população nas áreas urbanas.

Uma das novas abordagens para este problema envolve a utilização de robôs coletores autônomos, em um serviço de coleta porta a porta. Neste sistema, o robô coletor possui um certo nível de inteligência artificial que possibilita: navegação entre pessoas, automóveis e obstáculos dos mais variados tipos; coleta do resíduo sólido na residência do cidadão e; descarregamento do mesmo em uma estação centralizada. Em 2011 na Itália, a cidade de Péciole foi recipiente de um prolongado teste deste sistema, onde o robô DustCart interagiu com um ambiente urbano real e realizou as tarefas acima descritas com êxito.

Atualmente, dentro da área da inteligência artificial, existe uma categoria de algoritmos de aprendizado não supervisionado, composta por métodos que aprendem unicamente através de iteração com o ambiente no qual estão inseridos, mudando o comportamento do agente de acordo com sinais de reforço positivo ou negativo que ele obtém a cada ação executada neste ambiente. O SARSA é um destes algoritmos, sendo muito aplicado em problemas relacionados a robótica, ele é capaz de desenvolver um comportamento inteligente quando existe uma representação adequada do ambiente e que o mesmo seja suficientemente explorado pelo robô. Conforme especificado por [1], o SARSA possui uma garantia de conversão a uma solução ótima contanto que alguns requisitos sejam atendidos, porém tais resultados ainda não foram publicados.

Sendo assim, com este trabalho pretendemos descobrir a melhor configuração de alpha no algoritmo de aprendizado por reforço SARSA (aquela que traz mais recompensa em longo prazo) segundo as especificações definidas por [1] enquanto aplicamos o mesmo ao problema da coleta de lixo 2D.

METODOLOGIA

Desenvolvemos um mundo controlado, onde disponibilizamos um robô que deve coletar lixo e leva-lo até a lixeira. O robô não sabe onde está o lixo ou onde está a lixeira, mas consegue saber sua localização atual baseado em sua memória, ainda que esta posição seja apenas uma provável posição.

Neste mundo também disponibilizamos uma base carregadora, onde o robô deve acoplar-se caso sua bateria esteja abaixo de determinado nível. O desgaste na bateria do robô se dará em função do tempo, sendo que de tempos em tempos, o mesmo necessitará ir até a base carregadora para recarregar.

Utilizando as capacidades sensoriais deste robô geramos uma representação estadual, o robô em questão deve aprender a encontrar a posição do lixo, da lixeira e da base carregadora, além de entender as dinâmicas do mundo no qual está inserido. Ele deve levar o lixo até a lixeira no menor tempo e distância possíveis; ele também deve encontrar a base carregadora de energia quando considerar que sua bateria está fraca.

Para cumprir estas tarefas, utilizaremos um mundo controlado, onde podemos executar testes repetidamente e obter dados para análise com um maior grau de precisão. Neste mundo, a tarefa que o robô deverá aprender a desempenhar será a navegação entre estados onde ele encontra o lixo ou a lixeira, bem como o estado onde está a base carregadora de energia.

Este mundo será uma versão linearizada do problema da coleta de lixo, onde utilizaremos cores para representar os elementos envolvidos nesta tarefa. O lixo será representado pela cor prata e a cor vermelha representará a lixeira e a base carregadora, sendo assim, o robô em questão não vai obter o lixo fisicamente, esta será apenas uma variável dentro de sua memória. Quando o mesmo encontrar a cor prata entenderá que encontrou o lixo e agora está carregando o mesmo. Quando encontrar a cor vermelha entenderá que pode descarregar o lixo neste local.

Considerando que a base carregadora também é representada por uma cor, assim que o robô encontrar a cor que representa a base carregadora, o seu nível de energia é elevado ao máximo, e ele retorna a sua atividade de coleta de lixo.

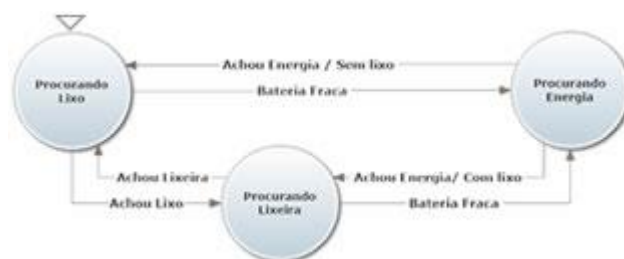


Fig. 1.

Diagrama de estados

Sendo assim, o robô estará sempre executando uma dentre três tarefas possíveis: procurando por lixo; procurando por lixeira ou procurando por energia, conforme pode ser visualizado na figura 1.



Fig. 3. Robô no mundo da coleta de lixo

O mundo conta com uma área de 3136cm², dentro do qual disponibilizamos 4 quadrados de diferentes cores, medindo 18.6cm de base por 18.6cm de altura. Estes quadrados representam (dependendo de sua cor) o lixo, a lixeira e a base carregadora, mas também podem ser utilizados como informação extra para ajudar o robô a formar a sua representação estadual.

Para a tarefa descrita, utilizaremos o robô iRobot Create série 4400, exibido na figura 3, tirando proveito de sua confiabilidade, flexibilidade e robustez. É um robô desenvolvido para ser aplicado a programação, atendendo usuários básicos e avançados.

Utilizando como base as informações obtidas pelos sensores do robô, criamos algumas propostas de funções de estado contendo o sinal do sensor de infravermelho, indicador do nível de bateria/presença de lixo e os indicadores de impacto frontal direito e esquerdo.

Desenvolvemos uma função de recompensa bastante simples, com foco em recompensar o robô ao realizar ações corretas e também recompensar a exploração no mundo. A tabela abaixo mostra a recompensa recebida pelo robô de acordo com a tarefa sendo realizada e o estado atual.

Tarefa	Estado atual	Recompensa
Procurando lixo	Achou lixo	1.00
Procurando lixo	Não achou lixo	0.00
Procurando lixeira	Achou lixeira	1.00
Procurando lixeira	Não achou lixeira	0.00
Procurando energia	Achou energia	1.00
Procurando energia	Não achou energia	0.00

TABELA I. FUNÇÃO RECOMPENSA

Para incentivar o robô a explorar o mundo, e assim acelerar a velocidade com a qual ele descobre estados e aprende, estamos dando um bônus para quando o robô escolhe a ação “andar para frente” e não encontra o limite do mundo.

RESULTADOS E DISCUSSÃO

Conforme pode ser verificado na figura abaixo, o alpha está afetando a velocidade do aprendizado, onde conforme maior o alpha, mais rápido será o aprendizado. Porém, ao chegar ao alpha 0.3, percebemos que o mesmo aprende mais rápido no começo da execução mas tem dificuldade em atingir uma política ótima em longo prazo.



Fig. 4. Influência do alpha na recompensa acumulada média por passo

O gráfico abaixo também demonstra a capacidade do alpha em afetar a velocidade do aprendizado. Para o alpha 0.2, percebemos que o mesmo recebe mais recompensa na primeira metade da execução, e também, que é a configuração de alpha que mais aumenta a quantidade de recompensa recebida na segunda metade da execução em relação a primeira.

Modalidade do trabalho: Relatório técnico-científico
Evento: XIX Jornada de Pesquisa

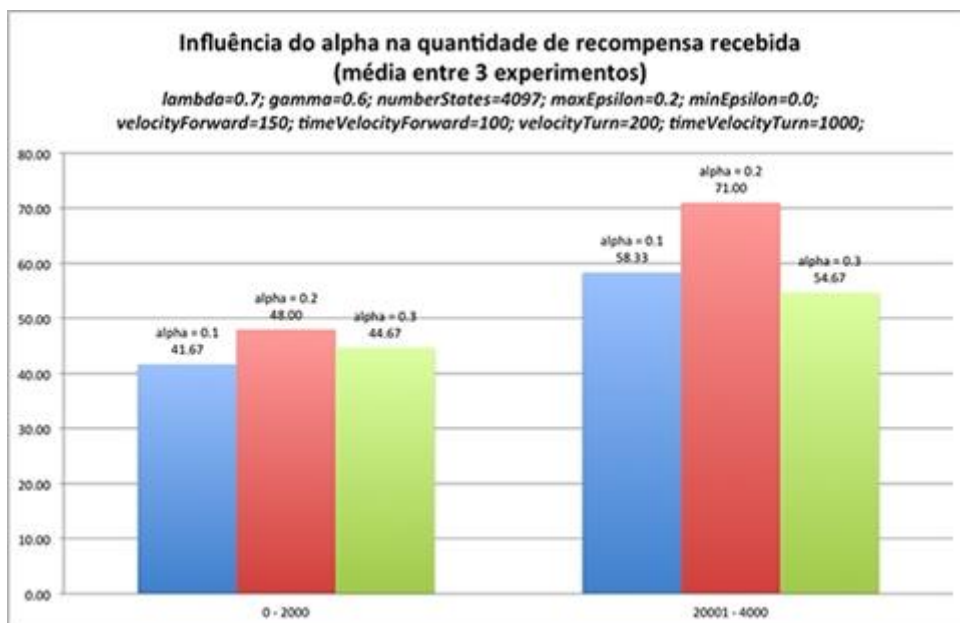


Figura 5 – Influência do alpha na quantidade de recompensa recebida

CONCLUSÕES

Concluimos que para a nossa configuração da tarefa da coleta de lixo, o melhor valor do parâmetro alpha no tempo de execução determinado em nossos testes é 0.2. A configuração de alpha 0.1 tende a atingir um desempenho igual ou maior que o alpha 0.2 (contanto que seja executado por um tempo maior), tanto na curva de aprendizado quanto na quantidade de recompensa recebida.

Também podemos perceber que o aumento do alpha não representa necessariamente uma conversão mais rápida a uma política ótima, este porém, influencia na velocidade do aprendizado nos primeiros momentos da execução, onde quanto mais alto o valor de alpha, mais rápida é a conversão a uma boa política.

Além disso, pode-se reparar que quando o objetivo é uma política ótima (melhor comportamento possível) é recomendável escolher um alpha menor, sacrificando a velocidade de aprendizado mas garantindo convergência a um resultado mais adequado em longo prazo. Por outro lado, quando o objetivo é obter um comportamento razoável (ainda que longe do ótimo) o mais rápido possível, pode ser mais adequado utilizar um valor maior de alpha.

PALAVRAS-CHAVE: Sarsa; Agente robótico; Curva de aprendizagem; Coleta de lixo; Função de recompensa.

Modalidade do trabalho: Relatório técnico-científico
Evento: XIX Jornada de Pesquisa

REFERÊNCIAS BIBLIOGRAFICAS: [1] G. A. Barto, R. S. Sutton, “Reinforcement Learning: An Introduction”, Bradford book, 2012.