**NOME**

Edna de Carvalho Andrade <ednacarvalhosempre@gmail.com>

Kaic de Oliveira Barros < [kaicbarros@gmail.com](mailto:kaicbarros@gmail.com)>

Marcos Neto Santos <marcos\_nto@hotmail.com>

Universidade Federal de Sergipe (UFS) - Curso de Sistemas de Informação – Campus Itabaiana Av. Vereador Olímpio Grande, S/N – Bairro Centro – CEP 495000-000 – Itabaiana - SE

Resumo

Abstract

**Title:**

1. Introdução

Aprendizado por Reforço é um ramo estudado pela ciência da computação. É um método de programação que consiste em um agente aprender como se comportar num ambiente dinâmico através de interações do tipo “tentativa e erro”. Nesse trabalho iremos demostrar os conceitos de Aprendizado por Reforço criando um simulador. Nele iremos utilizar a planta do Shopping Peixoto como local para o aprendizado de um robô, que por sua vez irá se locomover utilizando a base de aprendizado.

1. Aprendizado por Reforço

O aprendizado por reforço é um paradigma computacional de aprendizagem em que um agente aprendiz procura maximizar uma medida de desempenho baseada nos reforços que recebe ao interagir com um ambiente desconhecido. O agente tem como objetivo aprender de maneira autônoma uma política ótima de atuação, através da interação com o ambiente.

No ambiente de aprendizado por reforço, um agente é inserido em um ambiente e interage com ele através de percepções e ações. A cada passo o agente recebe como entrada uma indicação do estado atual do ambiente. O agente então escolhe uma ação a tomar, e gera a sua saída. A ação altera o estado do ambiente, e uma medida dessa mudança de estado é informada ao agente através de um valor de sinal de reforço. O comportamento do agente deve tomar ações que maximizem o valor final da soma dos reforços recebidos em um intervalo de tempo ou tentativas. Tal politica deve ser aprendida através de um processo de tentativa e erro.



Formalmente, o aprendizado por reforço pode ser identificado alguns elementos como por exemplo:

* Uma política responsável por definir o padrão de comportamento do agente, ou seja, uma política π determina como o agente deve decidir por certas ações, em detrimento de outras.
* Uma função valor que associa um valor a um estado (par estado-ação).
* O modelo do ambiente no qual o agente aprendiz será inserido.

A aprendizagem no aprendizado por reforço se dá sem a presença de um professor que ensina através de exemplos. Sua utilização é recomendada quando não se dispõe de modelos a priori, ou quando não se consegue obter exemplos apropriados das situações as quais o agente aprendiz irá enfrenta.

2. 1 algoritmo de q-learning

O algoritmo Q-learning desenvolvido por Watkins é considerado uma das mais importantes contribuições em aprendizado por reforço. Ele é um dos algoritmos de aprendizagem por reforço que não necessitam de uma modelagem completa do ambiente, ou seja, não necessita conhecer a matriz de probabilidades de transição e r (s, a) para todos os possíveis estados e ações do ambiente.

O Q-learning para valores ótimos de Q não depende da política que está sendo utilizada, a função ação-valor Q se aproxima diretamente a função ação-valor ótima, através de atualizações dos pares estado-ação, que são feitas à medida que estes pares são visitados. Assim, aprender a função ação-valor Q corresponde a aprender a política ideal (π\*). A expressão de atualização do Q-valores no algoritmo de Q-learning fundamenta-se na função ação-valor e é denotada por:

C:\Users\Edna\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ss.png

Uma característica importante deste algoritmo é que a escolha das ações a serem executadas durante o processo de aproximação interativa da função Q pode ser feita através de qualquer critério de exploração/explotação, inclusive de forma aleatória. Isso causa um dilema de quando deve-se explorar (escolher a ação randômica) e quando deve-se usufruir (escolher a ação que atualmente está com maior valor).

1. Projeto

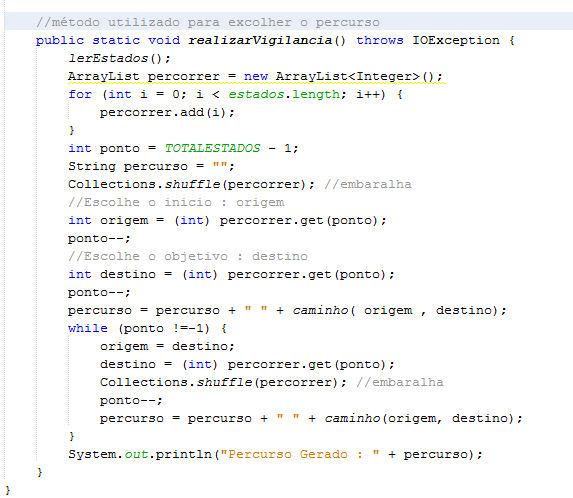
O objetivo do projeto é simular o aprendizado de máquina em um ambiente, depois utilizar a base de dados aprendida para percorrer os caminhos do ambiente, de forma que simule um vigia. Para o desenvolvimento do simulador, utilizamos uma planta de um ambiente (Shopping Peixoto) para aprendizado do mesmo. Nela foram inseridos pontos que representam os locais de acessos possíveis.

O software foi implementado na linguagem java, utilizando o software Netbeans IDE. Apresenta dentre suas classes a Dado, a Estado e a MetodosPrincipais. As duas primeiras classes são para estanciar um dado, composto do valor da ação e o identificador de um estado, e um estado do ambiente de aprendizado. Já a classe MetodosPrincipais como o próprio nome possui os métodos principais, dentre eles o método aprendizado e o realizarVigilancia.

O método aprendizado possui como parâmetros uma única variável que indica o estado objetivo dentre o conjunto de estados existente. O aprendizado é baseado no algoritmo Q-learning anteriormente apresentado. Ele realiza um laço de repetição, onde o ponto de parada foi definido como chegar 50000 de vezes ao objetivo, dentro dele existe outro laço de repetição que é executado até o estado atual, estado selecionado aleatoriamente na primeira execução, seja igual ao objetivo passado como parâmetro. Destro dele ocorre, a escolha aleatória (exploração), ou especifica (usufruir), de uma ação dentre as do estado atual. Depois disso, a soma da recompensa do próximo estado, dado a partir da ação escolhida do estado atual que está no momento, somado com o peso (0,9) multiplicado por um valor (o maior valor dentre os valores das ações do próximo estado), é atribuída a ação escolhida do estado atual. Após o estado atual recebe o valor do próximo estado. Após finalização do laço de repetição principal, as informações obtidas pelo processo são salvas em um arquivo especifico do estado objetivo.



O método realizarVigilancia realiza o processo de gerar o percurso o qual o rodô deverá percorrer. Esse percurso é gerado a partir da escolha aleatoriamente de um estado para ser a origem do processo de vigilância e um estado destino, esses e os demais pertencentes ao conjunto de estados (pontos) os quais ele deve monitorar. Em seguida, o caminho da origem ao destino é atribuído ao percurso, e o destino se torna a origem e é escolhido um novo estado para se tornar o destino, sendo que os destinos anteriores não pertencem mais a lista de estados a serem visitados para que não apresente no final locais sem monitoração. O processo é realizado até que a lista de pontos a serem monitorados esteja vazia.



1. conclusão

Este projeto teve como objetivo auxiliar na compreensão das técnicas e conceitos aprendizagem discutidas em sala, sendo o aprendizado por reforço acompanhado do algoritmo de Q-learning utilizados no desenvolvimento da ferramenta. Estes foram utilizados para fornecer dados para a aplicação dado o ambiente de aplicação.

1. Referências

LIMA JUNIOR, F. C. **Algoritmo Q-learning como Estratégia de Exploração e/ou Explotação para as Metaheurísticas GRASP e Algoritmo Genético**. Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, 2009. Disponível em: < ftp://ftp.ufrn.br/pub/biblioteca/ext/bdtd/FranciscoCLJ.pdf>. Acesso em 15 de set. 2017.

**Aprendizagem por Reforço**. Disponível em: <http://professor.ufabc.edu.br/~ronaldo.prati/InteligenciaArtificial/reinforcement-learning.pdf.> Acesso em 15 de set. 2017

**Aprendizado por reforço.** Disponível em: <https://www.maxwell.vrac.puc-rio.br/19637/19637\_4.PDF>. Acesso em 15 de set. 2017.