ESTG

Engenharia Informática

Projeto realizado por Rui Soares, nº 8150289

Inteligência artificial

Descobrir características de um vírus com um algoritmo genético

25/05/2021- Marco de canaveses

# Algoritmo genético

Numa primeira parte do projeto, foi decidido dividir o algoritmo genético em duas parte uma para os node e outra para as edges, o que foi verificado depois que se a área de procura das coordenadas dos node for muito grande vai ser mais difícil de encontrar as mesmas e quando as encontrarmos se tiverem o tipo errado iriam ser descartadas então foi decidido que as coordenadas e o tipo tinham que ser encontradas separadamente para ser muito mais rápido a encontrar totalmente os nodes. O mesmo se verificou para as edges e também foi decidido separar encontrar as edges de encontrar o weight

## Nodes

### Encontrar X e Y

* Para encontrar os X e Y é gerado uma percentagem de nodes na matriz com os limites da width e height do vírus que queremos encontrar.
* Depois de preencher uma percentagem da matriz é calculado o fitness e uma ordenação da lista de nodes a partir da qual é feito um elitismo, são tirados melhores nodes a quantidade de nodes tirados é definido pelo utilizador.
* Apos o elitismo é gerada a próxima população com copias dos nodes no elitismo, onde os primeiros nodes são os nodes do elitismo.
* Depois de gerada a população é feita uma mutação em que cada node de cada população tem uma percentagem de ser alterado, essa percentagem é introduzida pelo utilizador. Caso as coordenadas dos node mutado já exista na lista o node permanece igual.
* É feito outravez o elitismo, geração de população, a mutação e o calculo do fitness, isto repete-se até ser alcançado o critério de paragem.
* No final todos os nodes com o x ou y errado são retirados da lista do cromossoma com o fitness maior.

### Encontrar o Type

* Para encontrar o Type é gerada uma lista de nodes em cada elemento da população com copias do melhor cromossoma de X e Y encontrado no passo anterior.
* É mutado, em cada node dos cromossomas há uma probabilidade do tipo ser alterado, esta probabilidade é introduzida pelo utilizador.
* Depois é calculado o fitness.
* O elitismo é feito da mesma forma (é usado o mesmo método).
* E é gerada uma lista de nodes com o tipo dos melhores do elitismo em que os nodes no elitismo ficam na população gerada.
* É feito outra vez o elitismo, geração de população, a mutação e o cálculo do fitness, isto repete-se até ser alcançado o critério de paragem.
* No final todos é guardada a população com maior fitness.

## Edges

### Edges

* Depois de encontrar completamente os nodes a população de edges é gerada com uma percentagem das edges possíveis cada edge é gerada por 2 nodes para o início e o fim da edge, as edges nunca são repetidas.
* A população é mutada, é calculado o fitness e feito o elitismo.
* A segunda população é feita com o crossover uniforme dos nodes do elitismo, os primeiros nodes são os nodes do elitismo inalterados
* É feita mais uma vez a mutação, calculo de fitness e o elitismo.
* Isto repete-se até ser alcançado o critério de paragem.
* No final são retiradas as edges inúteis da lista da população com maior fitness e a mesma é guardada.

Numero de edges possíveis [ ( nº nodes \* (nº nodes -1))/2 ]

### Weights

* Para encontrar o Weight é gerada uma lista de nodes em cada elemento da população com copias do melhor cromossoma de edges encontrado no passo anterior.
* É mutado, em cada edge dos cromossomas há uma probabilidade do weight ser alterado, esta probabilidade é introduzida pelo utilizador.
* Depois é calculado o fitness.
* O elitismo é feito da mesma forma.
* E é gerada uma lista de edges com o weight dos melhores do elitismo em que as edges no elitismo ficam na população gerada.
* É feito outra vez o elitismo, geração de população, a mutação e o cálculo do fitness, isto repete-se até ser alcançado o critério de paragem.
* No final todos é guardada a população com maior fitness.

## Fitness

A é calculado da mesma forma para cada parte, sempre que é encontrada uma coisa correta é somado 0.01, o melhor fitness é o que tem o fitness maior.

## Mutação

A mutação foi estruturada da mesma forma para todas as partes, recebe a frequência e em cada elemento de cada cromossoma há uma probabilidade de haver uma mudança, a frequência é decidida pelo utilizador.

## Crossover

O crossover é feito nas edges porque o número de edges possível é bastante elevado então haver um crossover faz com que encontrar as edges possa ser mais rápido. O crossover usado é um crossover uniforme, em que há 50% de probabilidade de uma edge de cada cromossoma ser inserida, se não existir.

A probabilidade de crossover é sempre 1.

## Elitismo

O elitismo recebe um número, por exemplo 3, e retorna os 3 elementos com o melhor fitness na população.

## Critério de paragem

A procura para quando o elitismo mais alto é igual durante X gerações, X é decidido pelo o utilizador, este critério é igual para todas as partes do algoritmo.

# Analise

## Virus 1

Quantidade de vírus encontrados para a amostra = 100

Tamanho de população = 20

Critério de paragem = 10

Mutação = 0.15 Mutação 0.30

|  |  |
| --- | --- |
| Media | 56.37 |
| Mediana | 56 |
| Desvio | 3.506574 |
| Mínimo | 49 |
| Máximo | 70 |

|  |  |
| --- | --- |
| Media | 53.74 |
| Mediana | 54 |
| Desvio | 2.468468 |
| Mínimo | 48 |
| Máximo | 61 |

Mutação = 0.50 Mutação = 0.51

|  |  |
| --- | --- |
| Media | 53.66 |
| Mediana | 53 |
| Desvio | 2.782412 |
| Mínimo | 49 |
| Máximo | 63 |

|  |  |
| --- | --- |
| Media | 53.4 |
| Mediana | 53 |
| Desvio | 2.601476 |
| Mínimo | 48 |
| Máximo | 61 |

Mutação 0.52 Mutação = 0.60

|  |  |
| --- | --- |
| Media | 53.44 |
| Mediana | 53 |
| Desvio | 2.637415 |
| Mínimo | 48 |
| Máximo | 63 |

|  |  |
| --- | --- |
| Media | 53.76 |
| Mediana | 54 |
| Desvio | 3.008809 |
| Mínimo | 49 |
| Máximo | 65 |

O vírus 1 é pequeno, fazer um reset total a alguns cromossomas pode ajudar.

A media de iterações em 100 virus encontrados é menor em 52% de mutação

## Virus 2

Quantidade de vírus encontrados para a amostra = 100

Tamanho de população = 20

Mutação = 0.04 Mutação = 0.20

|  |  |
| --- | --- |
| Media | 125.48 |
| Mediana | 125 |
| Desvio | 9.333853 |
| Mínimo | 105 |
| Máximo | 148 |

|  |  |
| --- | --- |
| Media | 108.25 |
| Mediana | 108 |
| Desvio | 5.854076 |
| Mínimo | 98 |
| Máximo | 129 |

Mutação = 0.21 Mutação 0.22

|  |  |
| --- | --- |
| Media | 107.77 |
| Mediana | 108 |
| Desvio | 5.746945 |
| Mínimo | 96 |
| Máximo | 125 |

|  |  |
| --- | --- |
| Media | 108.47 |
| Mediana | 107 |
| Desvio | 5.80222 |
| Mínimo | 98 |
| Máximo | 125 |

Mutação = 0.23 Mutação 0.30

|  |  |
| --- | --- |
| Media | 108.75 |
| Mediana | 108 |
| Desvio | 5.841985 |
| Mínimo | 98 |
| Máximo | 127 |

|  |  |
| --- | --- |
| Media | 109.35 |
| Mediana | 109 |
| Desvio | 5.890199 |
| Mínimo | 97 |
| Máximo | 126 |

A mutação melhor é 21 % este vírus é mais complexo que o vírus 1, mudar tanto os nodes e edges começa a ser mau.

## Virus 3

Quantidade de vírus encontrados para a amostra = 100

Tamanho de população = 20

|  |  |
| --- | --- |
| Media | 222.07 |
| Mediana | 219.5 |
| Desvio | 18.22433 |
| Mínimo | 187 |
| Máximo | 297 |
|  |  |

Mutação = 0.04 Mutação = 0.10

|  |  |
| --- | --- |
| Media | 235.15 |
| Mediana | 233.5 |
| Desvio | 16.41899 |
| Mínimo | 194 |
| Máximo | 294 |

Mutação = 0.11 Mutação 0.12

|  |  |
| --- | --- |
| Media | 221.74 |
| Mediana | 221.5 |
| Desvio | 15.19478 |
| Mínimo | 193 |
| Máximo | 267 |

|  |  |
| --- | --- |
| Media | 221.09 |
| Mediana | 220 |
| Desvio | 14.27302 |
| Mínimo | 188 |
| máximo | 254 |

Mutação = 0.13 Mutação = 0.20

|  |  |
| --- | --- |
| Media | 228.31 |
| Mediana | 226 |
| Desvio | 17.11618 |
| Mínimo | 195 |
| Máximo | 275 |

|  |  |
| --- | --- |
| Media | 245.73 |
| Mediana | 247 |
| Desvio | 22.44358 |
| Mínimo | 188 |
| Máximo | 296 |

No vírus 3 a melhor mutação é de 0.12, o número de iterações desce até 0.12 e começa a subir a partir de 0.12.

## VIrus 4

Quantidade de vírus encontrados para a amostra = 100

Tamanho de população = 20

Mutação = 0.04 Mutação = 10

|  |  |
| --- | --- |
| Media | 189.66 |
| Mediana | 188 |
| Desvio | 13.75634 |
| Mínimo | 156 |
| Máximo | 223 |

|  |  |
| --- | --- |
| Media | 172.18 |
| Mediana | 171 |
| Desvio | 9.962941 |
| Mínimo | 148 |
| Máximo | 210 |

Mutação = 13 Mutação = 14

|  |  |
| --- | --- |
| Media | 170.74 |
| Mediana | 170 |
| Desvio | 10.18874 |
| Mínimo | 153 |
| Máximo | 200 |

|  |  |
| --- | --- |
| Media | 170.06 |
| Mediana | 169 |
| Desvio | 9.895229 |
| Mínimo | 148 |
| Máximo | 207 |

Mutação = 15 Mutação = 16

|  |  |
| --- | --- |
| Media | 171.63 |
| Mediana | 170 |
| Desvio | 10.29137 |
| Mínimo | 150 |
| Máximo | 206 |

|  |  |
| --- | --- |
| Media | 171.75 |
| Mediana | 171.5 |
| Desvio | 10.06281 |
| Mínimo | 151 |
| Máximo | 202 |

Mutação = 20 Mutação = 30

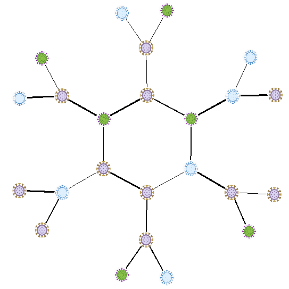
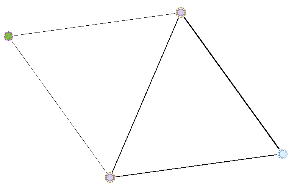
|  |  |
| --- | --- |
| Media | 173.72 |
| Mediana | 173 |
| Desvio | 11.1129 |
| Mínimo | 146 |
| Máximo | 205 |

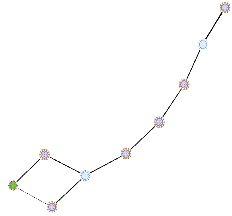
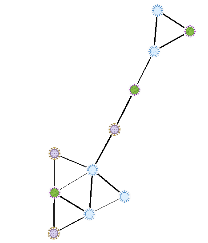
|  |  |
| --- | --- |
| Media | 179.2 |
| Mediana | 178 |
| Desvio | 12.51827 |
| Mínimo | 156 |
| Máximo | 217 |

A media de iterações da mutação desce até 14% depois aumenta, sendo assim a mutação ótima 14%.

## Conclusão da análise da mutação

Se ordenarmos os vírus com mutação ótima:

Virus1 – 52% -> Virus2 – 21% -> Virus4 – 14% ->Vírus3 – 12%



Podemos verificar que os vírus tambem ficam numa ordem de complexidade, onde o vírus mais pequeno tem a mutação ótima maior e o complexo tem a mutação ótima mais pequena.

Podemos concluir que quanto maior for a complexidade do vírus ou de encontrar o vírus mudar em menor quantidade os cromossomas é melhor, porque a quantidade de nodes ou edges ou área de encontrar nodes vai ser maior então uma percentagem de por exemplo 40% num vírus mais pequeno vai efetuar menos mudanças nos nodes e nas edges do que num vírus mais complexo.

Mutação = 40%

**Vírus pequeno** **Vírus grande**

4 nodes 20 nodes

4 edges 50 edges

No vírus pequeno vai afetar só 1 ou 2 nodes e 1 ou 2 edges

No vírus grande vai afetar 8 ou 9 nodes ou edges.

Se o fitness do vírus grande estiver a 98% de fitness ao mutar 40% vai quase fazer um reset a meio cromossoma, o que é muito mau porque vamos precisar de um número maior de iterações.

No vírus pequeno fazer uma mutação grande vai ajudar muito

# Analise da população

## Virus 3

Mutação = 0.20 Critério de paragem = 30

População = 10 População = 20

|  |  |
| --- | --- |
| Media | 281.97 |
| Mediana | 278 |
| Desvio | 39.48288 |
| Mínimo | 203 |
| Máximo | 390 |

|  |  |
| --- | --- |
| Media | 245.73 |
| Mediana | 247 |
| Desvio | 22.44358 |
| Mínimo | 188 |
| Máximo | 296 |

População = 30 População = 40

|  |  |
| --- | --- |
| Media | 206.06 |
| Mediana | 205.5 |
| Desvio | 14.27212 |
| Mínimo | 175 |
| Máximo | 244 |

|  |  |
| --- | --- |
| Media | 220.96 |
| Mediana | 222 |
| Desvio | 15.58484 |
| Mínimo | 180 |
| Máximo | 258 |

No vírus 3 , com a população 20 há uma diferença maior na média de iterações do que com 30 e 40 , o poder computacional vai precisar de maior do que numa população de 10.

Podemos concluir que no vírus 2 a partir de 20 de população não há uma mudança de media de iterações grande e o poder computacional está sempre a aumentar seria melhor usar uma população de 20.

## Virus 2

População = 10 população = 20

|  |  |
| --- | --- |
| Media | 127.36 |
| Mediana | 125.5 |
| Desvio | 10.84465 |
| Mínimo | 108 |
| Máximo | 160 |

|  |  |
| --- | --- |
| Media | 108.25 |
| Mediana | 108 |
| Desvio | 5.854076 |
| Mínimo | 98 |
| Máximo | 129 |

População = 30

|  |  |
| --- | --- |
| Media | 102.9 |
| Mediana | 103 |
| Desvio | 4.147532 |
| Mínimo | 94 |
| Máximo | 114 |

## Virus 1

População = 10 População = 20

|  |  |
| --- | --- |
| Media | 62.12 |
| Mediana | 62 |
| Desvio | 5.731853 |
| Mínimo | 49 |
| Máximo | 80 |

|  |  |
| --- | --- |
| Media | 56.37 |
| Mediana | 56 |
| Desvio | 3.506574 |
| Mínimo | 49 |
| Máximo | 70 |

População = 30

|  |  |
| --- | --- |
| Media | 53.82 |
| Mediana | 54 |
| Desvio | 2.375719 |
| Mínimo | 48 |
| Máximo | 60 |

## Conclusão da análise do tamanho da população

Nos vírus 1, 2 e 3 a maior diferença entre os tamanhos de população é no tamanho 20, a partir de 20 de população começa a ter menos diferença, como não é suposto sabermos o tamanho da área onde estamos a procurar nem se o vírus é muito complexo ou não é melhor usar sempre 20 de população até para os vírus 7, 8, e 9..

O algoritmo está a gerar uma percentagem de nodes, ou seja, se a área de procura for 100 por 100 e a percentagem for 60% vai gerar mais de 6000 nodes, com 20 de população encontramos os nodes mais rápido do que com 10 e a diferença de poder computacional é mais aceitável do que se gerarmos populações de 30, a diferença de iterações tambem vai ser muito maior de 10 para 20 e a partir de 20 a diferença vai diminuindo e poder computacional aumentando.

O mesmo acontece nas edges onde é gerada uma percentagem do total de edges possíveis.