Observar as concordâncias nas frases!!

Relatório Técnico (50%)

Pad: Padrão (o trabalho contém todos os itens requeridos): 10,0
Ling: Linguagem (o texto está bem escrito, correto gramaticalmente e as ideias bem expressas): 7,0
Mod:Modelo (o modelo está bem descrito e sua instanciação feita adequadamente):10.0
Res: Resultados (os resultados foram devidamente descritos, inclusive utilizando gráficos): 10.0
Aval: Avaliação (os resultados foram adequadamente apresentados): 10,0

Implementação (50%)

Com: Comentários: (o código está devidamente documentado): 10,0
Cod: Codificação: (o código está bem escrito e roda adequadamente): 10,0
Result: Resultados: (os resultados foram produzidos adequadamente, inclusive utilizando gráficos): 10,0

Nota: 9.9

Relatório Técnico
sobre Algoritmos Genéticos

Thiago M. de Sousa *Luana G. B. Martins*
Ruan C. Rodrigues

Technical Report - RT-INF_000-19 - Relatório Técnico
July - 2019 - Julho

Instituto de Informática
Universidade Federal de Goiás
www.inf.ufg.br

Relatório Técnico
sobre Algoritmos Genéticos

Thiago M. de Sousa
thiagomontelesofc@gmail.com

Luana G. B. Martins
luanagbmartins@gmail.com

Ruan C. Rodrigues
ruanchaves93@gmail.com

Abstract. *This report describes which were the decisions, resolution process and results for the problem proposed during the Artificial Intelligence subject. A challenge was introduced to generate a recommendation of the best group of courses to be enrolled by the students on each semester, as a means to attain the highest grades, through Genetic Algorithm techniques.*

Keywords: Technical Report, Genetic Algorithm.

Resumo. *Este relatório descreve quais foram as decisões, processo de resolução e resultados do problema proposto na matéria de Inteligência Artificial. Foi introduzido o desafio de criar uma recomendação de qual o melhor grupo de disciplinas a serem cursados pelos alunos a cada semestre, de forma a obterem o maior sucesso acadêmico, utilizando técnicas de Algoritmos Genéticos.*

Palavras-Chave: Relatório Técnico,Algoritmos .

Procurem usar frases menores...

1 Introdução

Este Relatório Técnico consiste na documentação de uma estratégia assumida para se obter um processo de tomada de decisões, buscando resolver o problema de se criar uma recomendação de quais matérias devem ser feitas em cada semestre no intuito de maximizar o desempenho do aluno do Bacharelado em Ciência da Computação da Universidade Federal de Goiás, tendo em mãos apenas o histórico de disciplinas já cursadas.

O processo de decisão se deu por meio da utilização de um algoritmo genético, as quais por sua vez tiram proveito de uma abordagem baseado nos Processos Genéticos de Mendel (1865) e da Teoria da Evolução de Darwin (1859).

em

a

O problema consiste no primeiro momento receber lista de matérias já cursadas pelo o aluno, consultar a disponibilidade de novas matérias e selecionar, a cada semestre, um conjunto de matérias a serem cursadas pelo aluno com fim de maximizar o seu desempenho.

No restante deste documento estão definidas a forma abordada na base de dados (Seção 2), descrição geral da solução proposta contendo ~~nela~~ a descrição geral do modelo utilizado e dos dados selecionados para o modelo (Seção 3), dos resultados obtidos (Seção 4), das propostas para como utilizar os resultados obtidos (Seção 5), conclusões finais (Seção 6) e referências.

2 Descrição da base de dados

Os dados consistem em um arquivo no formato csv (Comma-separated values) que é representado por uma matriz de 22361 linhas por 66 colunas, onde existe em cada coluna um determinado atributo referente a relação de um aluno com as disciplinas que cursou durante os anos e períodos.

	id	ano_nascimento_discente	idade_conclusao_ensino_medio	idade_ingresso_universidade	idade_colacao_grau	uf_naturalidade_discente	c
0	1	1989	17.0	19	26.0	GO	
1	1	1989	17.0	19	26.0	GO	
2	1	1989	17.0	19	26.0	GO	
3	1	1989	17.0	19	26.0	GO	
4	1	1989	17.0	19	26.0	GO	

Figura 1: Exemplo da base de dados de alunos de Ciência da Computação utilizada.

na

Os atributos nas colunas contém dados referente ao aluno e sua passagem no curso. Dados como ano de nascimento, idade de ingresso à universidade e nota do Enem são exemplos de dados relacionados ao aluno anteriormente ao ingresso na faculdade. Já atributos como quantidade de reprovações, média global e ano de conclusão estão ligados ao aluno após a entrada na universidade.

3 Descrição da solução

Com o objetivo de encontrar uma boa recomendação sobre quais matérias o aluno deve selecionar, a cada semestre, a fim de obter a maior média global possível e o mínimo de reprovações, será proposto um modelo de Algoritmo Genético.

3.1 Descrição do modelo utilizado

Nesta seção serão discutidos os princípios do funcionamento de um Algoritmo Genético e como foram aplicados como solução ao problema proposto.

3.1.1 Algoritmos Genéticos

Algoritmos Genéticos (AG) são uma técnica de busca de soluções, com características de busca cega. Foram inspirados pelos modelos biológicos de Processos Genéticos de Gregor Mendel (1865), que consistem nas transmissões de características hereditárias; e na Teoria da Evolução de Charles Darwin (1859), com a ideia da evolução dada pela sobrevivência, adaptação e reprodução dos indivíduos.[1]

O algoritmo é implementado com base em uma simulação de um ecossistema em que uma população de representações de soluções são selecionadas a fim de obter as melhores soluções ou indivíduos. A evolução geralmente se inicia com um conjunto de soluções aleatórias e é realizada por meio de gerações. A cada geração é avaliada a adaptação de cada solução na população, assim alguns indivíduos são selecionados para próxima geração, e recombinados ou mutados para formar uma nova população de soluções. A nova população é utilizada como entrada para próxima iteração do algoritmo.[3]

O processo se baseia então nos seguintes passos:

Figura 2: Processo de um Algoritmo Genético.

1. Geração de uma população inicial com indivíduos escolhidos aleatoriamente.
2. Avaliação dos indivíduos com o cálculo da função *fitness* (função objetivo).
3. Seleção dos indivíduos mais aptos.
4. Geração de uma nova população a partir do cruzamento e mutação dos indivíduos selecionados.
5. É avaliado o critério de parada e, caso não seja o desejado, é repetido o processo com a nova população gerada.

Os principais termos utilizados durante a descrição desse processo são:

- Gene: É uma decisão específica, como, por exemplo, subir ou descer, incluir ou não incluir.
- Cromossomo: É a representação de uma sequência de genes, sendo ela um indivíduo que representa uma solução.
- Geração: é o conjunto de indivíduos (cromossomos) em uma determinada iteração.
- População: É o conjunto de indivíduos que composto pelo conjunto de gerações.
- Genótipo: São os aspectos físicos que caracterizam um cromossomo: a sequência de genes.

- Fenótipo: São os aspectos resultantes da manifestação das informações do genótipo: a solução gerada a partir de um cromossomo.

Além disso, um termo importante no processo é a função de *fitness*, onde para cada indivíduo é calculado uma avaliação (fitness). Esse valor diz o quanto o indivíduo está adaptado ao seu ambiente. Na representação computacional, essa medida indica o quanto uma solução (indivíduo) é capaz de resolver o problema (ambiente). Uma boa definição para essa função é essencial para que o processo evolutivo seja capaz de fornecer uma solução que consiga resolver satisfatoriamente o problema proposto.

3.1.2 Representação do Cromossomo

A representação clássica de um cromossomo é feita como um vetor de caracteres de tamanho fixo, formado pela concatenação de caracteres 0 e 1 (representação binária). Apesar disso, existem variações para esta representação, como a representação de como valores de ponto flutuante. No caso de uma representação binária, ela é recomendada quando a solução pode ser mapeada na alternância de ações, como por exemplo, faça (1) ou não faça (0).

3.1.3 Seleção do Cromossomo

A etapa de seleção consistem na fase onde é escolhido os indivíduos para a reprodução. Existem vários modelos de seleção que podem ser aplicadas, entre os modelos mais utilizados está o da Roleta (*roulette wheel*) e Torneio.

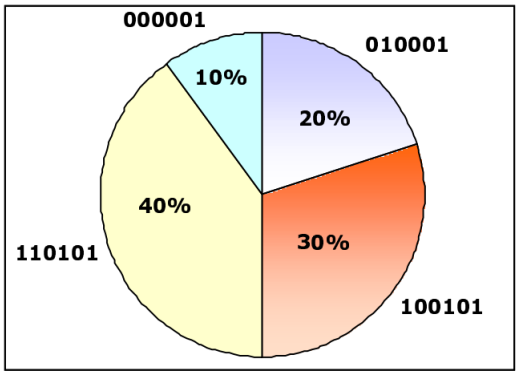

Figura 3: Método de seleção *roulette wheel*.

No método de *seleção por Roleta*, cada cromossomo é representado em uma roleta, tendo como proporção o seu índice de aptidão. Desse modo, os cromossomos com alta aptidão tem maior porção da roleta, enquanto os de menor aptidão mais baixa, é dada uma porção menor.[3]

No método de *seleção por Torneio*, é dado que um número *n* de indivíduos da população é escolhido de forma aleatória para formar um grupo temporária. Neste grupo o cromossomo escolhido dependerá somente do grau de adaptação dada a cada indivíduo do grupo.

O método de *seleção por classificação* foi feita de maneira semelhante ao de Torneio, mas ele utiliza de uma nova mecânica que corrige um possível problema com o modelo de Torneio que acontece quando existe uma grande diferença entre os valores de adequação para cada cromossomo do grupo, isso diminui muito as chances dos menores serem escolhidos e assim a população não fica diversificada. Para isso, a seleção de Classificação inicialmente classifica a população e atribui a cada indivíduo uma valor com base em sua colocação. Assim, todos os cromossomos tem chances de serem selecionados.[2]

Não tem nada acima!!!!!

O método de *Elitismo* se dá inicialmente pela cópia dos melhores cromossomos para a nova população, posteriormente o resto da população é feita utilizando os métodos citados acima. Deste modo o método previne a perda da melhor solução encontrada e deixa a nova população diversificada.[4]

3.1.4 Operadores Genéticos

O objetivo dos operadores genéticos é fazer transformações através de várias gerações, assim podendo chegar a uma formação que seja o desejado para o problema. Esse processo é necessário para que a população possa ficar diversa e mantendo características mudanças feitas pelas gerações anteriores. As principais abordagens são o método de cruzamento (crossover) e mutação.

- Cruzamento: O operador *crossover* possibilita a passagem da genética predominante. Esse processo é feito através do cruzamento de dois indivíduos, onde os trechos das características de um indivíduo são trocadas pelo trecho do segundo indivíduo. Como resultado, a operação produz um indivíduo que pode ter a combinação das melhores características de seus pais. Esse cruzamento pode ser feito em um ponto, dois pontos ou de maneira uniforme.

  – *crossover* de um ponto:

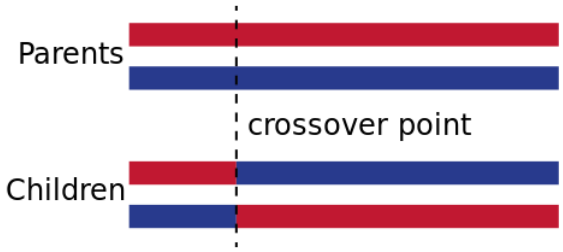

Figura 4: *crossover* de um ponto.

Tem como característica o corte em um posição aleatória e posteriormente recombinando as partes dos indivíduos pais.

  – *crossover* de dois pontos:

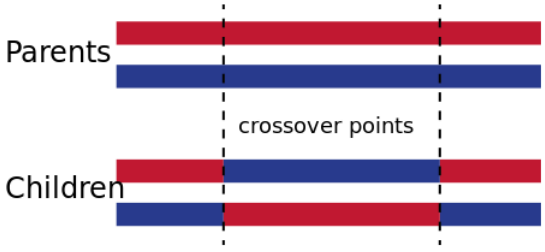

Figura 5: *crossover* de dois pontos.

Tem como característica o corte em duas posições aleatórias e posteriormente recombinando as partes dos indivíduos pais.

  – *crossover* uniforme: É gerado uma máscara de bits aleatórios e combinar os bits dos pais de acordo com a máscara gerada.

uniformizar as fontes das figuras

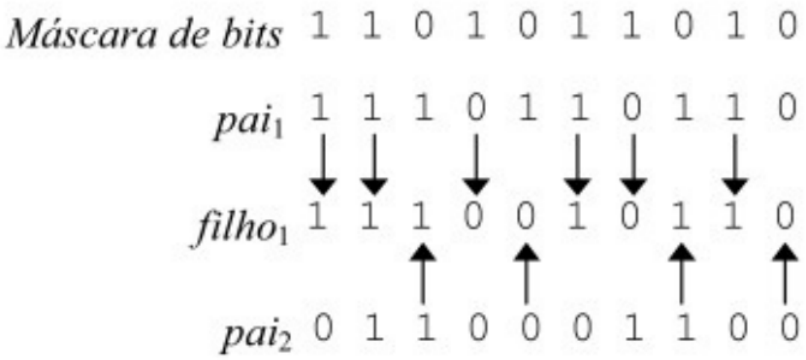

Figura 6: *crossover* uniforme.

- Mutação: A operação de mutação é a forma mais simples de se modificar um cromossomo. Ela é feita modificando de maneira aleatória algum gene do indivíduo. Apesar de ser simples, o operador ajuda a manter uma diversidade de indivíduos na população, de modo que ao fazer a troca, produz-se uma possibilidade de surgirem novas características que até então não existiam ou apareciam em uma pequena quantidade na população.

3.2 Descrição dos dados selecionados

Optamos por utilizar integralmente a base de dados fornecida, sem descartar informações de nenhum aluno. Foram agregados à base de dados já existente as informações de pré-requisito entre as disciplinas. Dado que as disciplinas na base de dados pertenciam a dois projetos pedagógicos cronologicamente distintos, também foi criado um mapa de equivalências entre as disciplinas do antigo e do novo plano. Após essa inclusão de dados, foi calculada a média geral dos alunos em cada uma das disciplinas na base de dados.

3.3 Descrição do modelo final

Dado o problema de selecionar, a cada semestre, um conjunto de k matérias a serem cursadas que maximizem as chances do aluno obter a maior média global possível e o mínimo de reprovações. Sendo assim, partindo de uma lista inicial de matérias já cursadas, consultando as tabelas de dependências e equivalências, produzimos uma lista com as matérias possíveis de serem cursadas pelo aluno no próximo semestre. Nesta lista de matérias possíveis, cada matéria está acompanhada pela média geral de todos os alunos nesta matéria.

Sabemos que as médias gerais de todas as disciplinas, consideradas em conjunto, seguem uma distribuição normal. Portanto, caso a média de nossa seleção esteja próxima da média de todas as disciplinas possíveis de serem cursadas pelo aluno naquele semestre, sabemos que temos em nossa seleção quantidades equilibradas de matérias fáceis e difíceis. Uma seleção de matérias muito difíceis aumenta as chances de reprovação, e uma seleção de matérias muito fáceis ~~irá nos~~ obrigar a fazer uma seleção de matérias muito difíceis em um momento posterior do curso.

Para solucionar o problema, portanto, devemos fazer uma seleção equilibrada. Devido à distribuição normal das médias, se a cada semestre realizamos uma seleção equilibrada, este equilíbrio será mantido no decorrer dos semestres e não iremos cair em situações de desequilíbrio. Sendo assim, para cada semestre, selecionamos k matérias a serem cursadas pelo aluno,

de tal modo que a média das matérias selecionadas seja o mais próximo possível da média das matérias possíveis.

3.4 Definição formal do modelo

Seja E a entrada do algoritmo, com:

$$E = \{(i_1, m_1), ..., (i_k, m_k)\} \tag{-1}$$

sendo i_n o índice associado à n-ésima matéria possível de ser cursada no próximo semestre, e m_n a média geral dos alunos na n-ésima matéria.

Cada cromossomo é representado como uma lista de genes da forma:

$$C = [b_1, ..., b_k] \tag{-2}$$

Onde cada cromossomo contém k genes, o mesmo tamanho da entrada, e cada gene pode assumir valores booleanos: $b_n \in \{0, 1\}$. Caso um gene assuma valor verdadeiro, a matéria cujo índice corresponde à posição ocupada pelo gene no cromossomo deve ser incluída na solução proposta pelo cromossomo. Caso assuma valor falso, ela não deve ser incluída.

Seja z a quantidade de matérias que o aluno deseja cursar no próximo semestre.

Definidas as variáveis:

$$A = \left| \frac{\sum_{i=1}^{k} m_i - \sum_{i=1}^{k} (m_i * b_i)}{k} \right| \tag{-3}$$

$$B = \left| 2^z - \prod_{i=1}^{k} 2^{b_i} \right| \tag{-4}$$

Então queremos selecionar, a cada geração do algoritmo, os cromossomos que maximizam a seguinte função de *fitness*:

$$f(b_1, ..., b_k) = \frac{1}{A + 1} * \frac{1}{B + 1} \tag{-5}$$

Onde:

- O melhor cromossomo possível é aquele que obtém $A = 0$ e $B = 0$, alcançando assim $f(b_1, ..., b_k) = 1$.
- $A = 0$ significa que a média das médias na seleção do cromossomo é exatamente igual à média das médias de todas as matérias possíveis. A função de A é penalizar cromossomos que sugerem seleções desequilibradas de matérias, onde a média das médias da seleção se distancia da média das médias de todas as matérias possíveis.
- $B = 0$ significa que o cromossomo sugeriu exatamente z matérias. A função de B é penalizar cromossomos que sugerem uma quantidade de matérias a serem cursadas distinta da quantidade de matérias z que o aluno deseja cursar.
- Na variável B, o número 2 foi escolhido como base dos expoentes, o qual comprovamos empiricamente ser eficiente para a distinção entre cromossomos com quantidades corretas e incorretas de sugestões. Manter a base dos expoentes em um valor baixo também garante a eficiência computacional do cálculo da função de aptidão.

4 Resultados obtidos

Foi decidido com base em testes as seguintes características para o algoritmo:

- Tamanho da população : 25
- Número máximo de gerações: 50
- Probabilidade de *crossover*: 0.8
- Probabilidade de mutação: 0.01
- Parâmetro de definição do tamanho do torneio : 10

Após a definição de todos os parâmetros e funções necessárias para o auxílio do algoritmo genético, a busca por uma solução otimizada pode ser iniciada.

Foi utilizado como modelo de seleção a estratégia de torneio e para auxiliar o método elitismo, que cria uma cópia das melhores soluções para garantir que elas não sejam ignoradas. Como operador genético foi utilizado o método de cruzamento onde é feito a combinação de partes de pares de soluções e mutação para alterar a composição de algumas soluções, assim permitido a diversidade da população.

Com fim de validar os resultados, foi comparado o aluno que segue o fluxo sugerido pelo nosso algoritmo com o fluxo sugerido a partir do projeto pedagógico do curso.

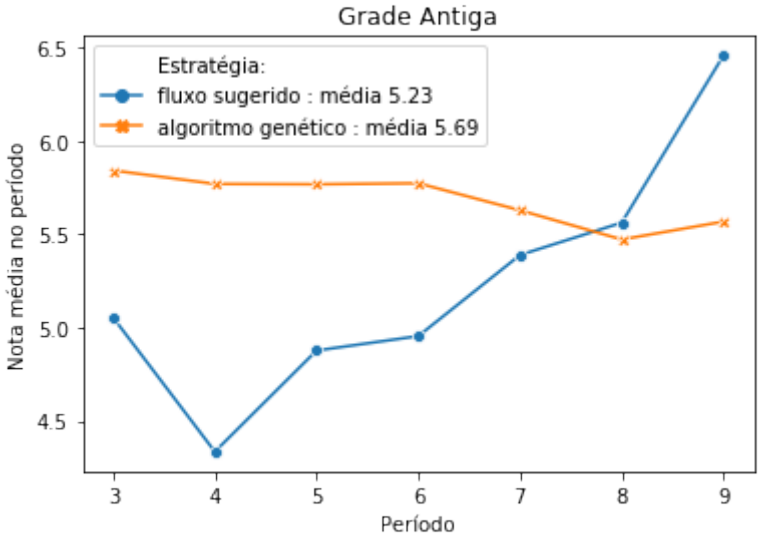

Figura 7

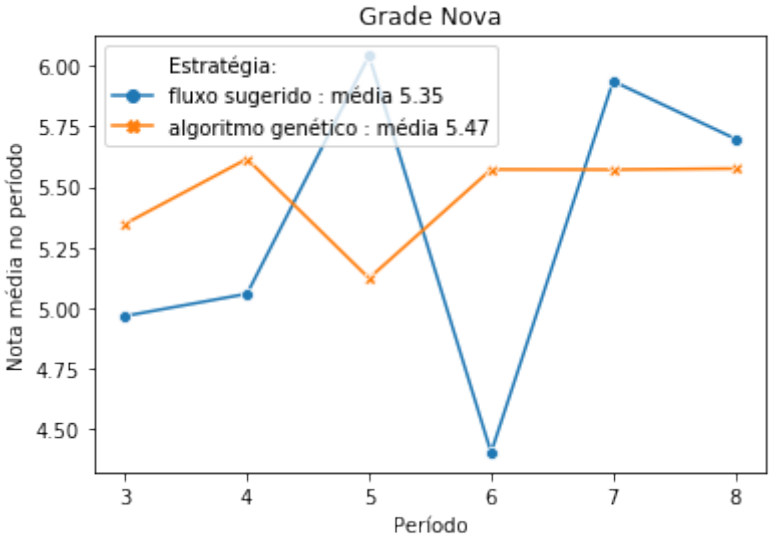

Figura 8

É possível perceber que o algoritmo consegue sugerir um fluxo que obtém uma menor variação de notas ao passar dos semestres em comparação ao fluxo sugerido pela administração do curso, isso se aplica tanto à grade antiga, como também à grade nova.

Para concluir nossos resultados, foram realizados 100 execuções do algoritmo com o objetivo de se encontrar o melhor resultado após esses testes:

	0	1	2	3	4
3º período	PESQUISA OPERACIONAL	TEORIA DOS GRAFOS	MULTIMÍDIA	ENGENHARIA DE REQUISITOS	COMPUTADOR E SOCIEDADE
4º período	PROBABILIDADE E ESTATÍSTICA	TÓPICOS 2	ARQUITETURA DE COMPUTADORES	SEGURANÇA E AUDITORIA DE SISTEMAS	DIREITO
5º período	REDES DE COMPUTADORES 1	ESTRUTURAS DE DADOS I	COMPUTAÇÃO GRÁFICA	ENGENHARIA DE SOFTWARE	PROJETO FINAL DE CURSO 1
6º período	LINGUAGENS FORMAIS E AUTÔMATOS	PROGRAMAÇÃO ORIENTADA A OBJETOS	ESTRUTURAS DE DADOS II	PROJETO DE SOFTWARE	TÓPICOS 1
7º período	ANÁLISE E PROJETO DE ALGORITMOS	TEORIA DA COMPUTAÇÃO	BANCO DE DADOS	COMPILADORES	EMPREENDEDORISMO
8º período	REDES DE COMPUTADORES 2	LINGUAGENS DE PROGRAMAÇÃO	SISTEMAS GERENCIADORES DE BD	SISTEMAS DISTRIBUÍDOS	PROJETO FINAL DE CURSO 2
9º período	SISTEMAS OPERACIONAIS 1	INTELIGÊNCIA ARTIFICIAL	0	0	0

Figura 9: Matérias sugeridas pelo Algoritmo Genético.

5 Como a solução proposta pôde resolver o problema

Ao comparar o aluno que segue o modelo proposto com um aluno que segue o fluxo sugerido no projeto pedagógico, visivelmente se percebe que o algoritmo genético consegue uma variação na dificuldade entre os semestres muito menor do que o fluxo. Sendo assim, as chances de reprovação ou de notas baixas seguindo o algoritmo genético são muito menores do que seguindo o fluxo sugerido.

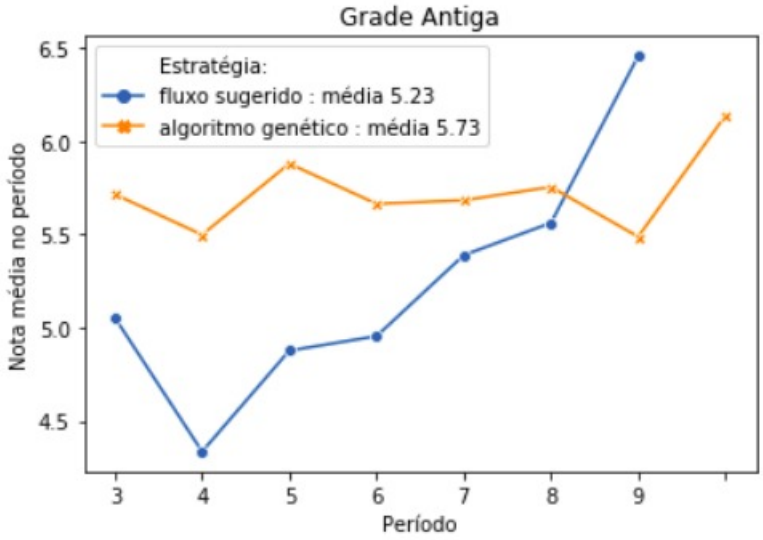

Figura 10

A visualização gráfica dos resultados mostra que o nosso modelo foi capaz de obter um fluxo de sugestões bem equilibrada ao longo do curso. O *heatmap* abaixo mostra cada linha como um período do curso de Ciência da Computação, e cada coluna é uma das matérias selecionadas. Uma tonalidade mais escura indica que a matéria selecionada possui uma dificuldade mais alta, e uma tonalidade mais clara indica uma menor dificuldade. A tonalidade totalmente branca, quando ocorre na última linha, indica uma célula vaga, não preenchida por matéria.

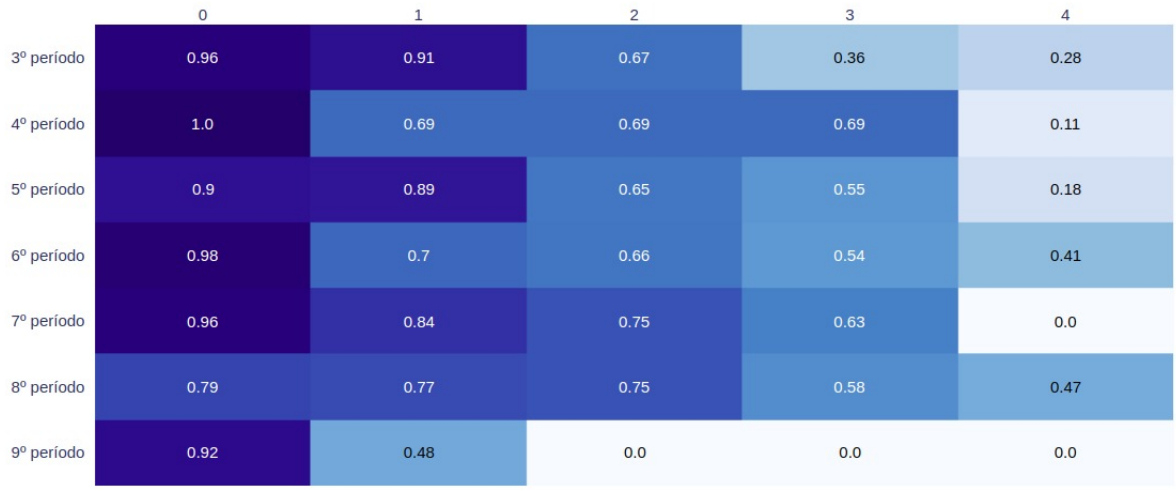

Figura 11: Representação gráfica da distribuição dificuldade das matérias selecionadas pelo Algoritmo Genético.

6 Conclusões

Os Algoritmos Genéticos buscam soluções para problemas de otimização, de forma análoga ao processo de evolução natural. Sua resolução se dá a partir da busca em uma população inicial, que efetuando o processo de evolução da mesma, obtêm uma nova população que apresenta melhores soluções para o problema em questão. É importante destacar que os Algoritmos Genéticos possuem a capacidade de resolver problemas com funções de alto grau de complexidade, porém a otimização tem um caráter local, e não existe maneira de saber quando uma solução ótima foi alcançada. Entretanto, como mostra os gráficos apresentados, percebe-se que o Algoritmo Genético consegue uma variação na dificuldade entre os semestres muito menor do que o fluxo sugerido pelo projeto pedagógico do curso de Ciência da Computação, e, em cada período, temos um equilíbrio na dificuldade das matérias selecionadas. Portanto, conseguimos constatar que o algoritmo apresentou uma distribuição equilibrada em cada período, sendo assim, as chances de reprovação ou de notas baixas seguindo o Algoritmo Genético são muito menores do que seguindo o fluxo sugerido.

Referências

[1] **Genetic algorithm**. https://en.wikipedia.org/wiki/Genetic_algorithm,.

[2] Obitko, M. **Genetic Algorithms**. https://www.obitko.com/tutorials/genetic-algorithms/a/selection.php, último acesso em Agosto de 2019.

[3] Pozo, A. **Computação Evolutiva**. inf.ufpr.br/aurora/tutoriais/Ceapostila.pdf, último acesso em Agosto de 2019.

[4] Silva, A. G. **Algoritmos Genéticos**. https://www.inf.ufsc.br/~alexandre.goncalves.silva/courses/14s2/ine5633/slides/aulaAG.pdf, último acesso em Agosto de 2019.