UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto n°17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVI SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2022

Proposta de Operadores Genéticos na Seleção de Características de Bases de Dados de Organismos Modelo

José Gabriel Gomes dos Santos Oliveira[1]; Fabiana Cristina Bertoni[2]

1. Bolsista PROBIC/UEFS, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: engcgabrielgomes@gmail.com

2. Orientadora, Departamento de Ciência Exatas, Universidade Estadual de Feira de Santana, e-mail: fcbertoni@uefs.br

PALAVRAS-CHAVE: algoritmos genéticos multiobjetivos; organismos modelo; seleção de características.

INTRODUÇÃO

O envelhecimento é um processo biológico natural de todos os organismos e vêm sendo muito estudado recentemente. Entretanto, não se sabe ao certo todos os mecanismos que influenciam nesse processo, tanto na longevidade quanto na anti-longevidade. Porém, é certo de que muitas pesquisas estão cada vez mais descobrindo processos biológicos que estão atrelados ao envelhecimento. Um exemplo deles é a restrição calórica que pôde estender a longevidade de muitas espécies.

Com o aumento dos estudos nessa área, houve um grande crescimento dos dados genéticos dos organismos que envolvem aspectos da sua longevidade e anti-longevidade. A partir delas, é possível aplicar técnicas de mineração de dados que permitam extrair as informações genéticas dos organismos, aplicando metodologias que visam extrair as suas características, tornando viável encontrar correlações entre os genes presentes nos organismos com o processo do envelhecimento.

O objetivo deste trabalho é utilizar Algoritmos Genéticos Multiobjetivos (AGMO) na seleção de características de organismos modelo, contribuindo para o estudo dos fatores que contribuem para o envelhecimento de organismos modelos.

METODOLOGIA

Os dados utilizados neste trabalho foram obtidos a partir de duas bases de dados: *Gene Ontology Resources* (GO) [1,2] e *Human Ageing Genomic Resources* [3] (HAGR). Os dados foram filtrados em apenas termos GO que representam um processo biológico [4, 5] dos seguintes organismos modelo: *Drosophila melanogaster, Mus musculus, Caenorhabditis elegans e Saccharomyces cerevisiae*. Mais especificamente, um exemplo de um organismo modelo é constituído por genes, os quais são termos GO. A combinação destes genes determina se o exemplo de um determinado organismo modelo possui tendência de pró-longevidade ou anti-longevidade. Em outras palavras, exemplos classificados como pró-longevidade indicam que os seus genes favorecem a longevidade do organismo, em contrapartida, exemplos anti-longevidade indicam que os seus genes favorecem a diminuição da vida útil.

A principal questão que envolve este trabalho é saber se realmente todos os genes de um organismo modelo são necessários para determinar a pró-longevidade ou a anti-longevidade. Ou seja, será que não existe um subconjunto de genes que melhor determine a pró-longevidade ou a anti-longevidade. Sendo assim, para encontrar o subconjunto destes genes, neste trabalho foi aplicado o algoritmo *Non-dominated Sorting Genetic Algorithm II* (NSGA-II) [6], que é um Algoritmo Genético Multiobjetivo.

A codificação dos cromossomos está representada na Figura 1. O valor 0 indica a ausência de um gene e o valor 1 indica a sua presença. Desta forma, cada cromossomo define um conjunto de genes, e o tamanho de um cromossomo é igual a quantidade total de genes de um determinado organismo modelo.

Figura 1. Exemplo da codificação de um cromossomo.

1	0	1	1	0	1	0	1	...	1

Um fator importante para a utilização dos dados oriundos do *Gene Ontology Resources* e *Human Ageing Genomic Resources* foi a frequência dos genes no conjunto. Para a seleção dos dados foi necessário definir thresholds, ou seja, um limite mínimo de frequência que cada gene deveria estar presente no conjunto de dados para ser considerado como um gene válido para este conjunto de dados. Os thresholds variaram de 3 a 10, desta forma, para cada valor de threshold foi gerado um conjunto de dados para cada organismo modelo.

Para avaliar a qualidade (*fitness*) dos cromossomos é necessário encontrar um balanceamento entre a acurácia do classificador, ao classificar os exemplos como pró-longevidade ou anti-longevidade, e a taxa de redução de características. A acurácia foi calculada a partir da média geométrica das medidas de sensibilidade e especificidade [7]. A taxa de redução de características foi calculada a partir da quantidade de genes que estão codificados com zero no cromossomo, dividida pela quantidade total de genes.

Em relação à definição dos parâmetros do algoritmo genético, foi aplicado o método de pesquisa em grade (*grid search*) [8]. Os parâmetros são as taxas de mutação e cruzamento e o tamanho da população. Para todos esses parâmetros foram definidos 27 testes, para que a melhor combinação pudesse ser alcançada. Na Tabela 1 são apresentados os valores testados para cada parâmetro.

O operador de mutação utilizado foi o *Bit-Flip Mutation*, o de cruzamento foi o *Half Uniform Crossover* (HUX) e o de seleção foi o torneio binário.

Após a execução do grid search, a melhor combinação de parâmetros obtida está descrita na Tabela 2. A quantidade de gerações foi fixada empiricamente em 100. Uma quantidade maior de gerações apenas consumia mais tempo de processamento, sem melhorar os resultados de acurácia e taxa de redução.

Tabela 1. Conjunto de valores dos parâmetros testados usando o grid search.

Mutação	20%	40%	60%
Cruzamento	50%	70%	90%
População	100	150	200

Tabela 2. Valores dos parâmetros definidos após o grid search.

Base de dados	Taxa de Mutação	Taxa de Cruzamento
Drosophila melanogaster	20%	50%
Mus musculus	20%	90%
Caenorhabditis elegans	20%	90%
Saccharomyces cerevisiae	20%	50%

É importante ressaltar que os valores de acurácia foram obtidos utilizando a técnica **10-folds cross-validation**. Essa técnica divide o conjunto de dados original em 10 partes iguais (folds). Assim, os classificadores são treinados e testados 10 vezes, e em cada par treinamento-teste, usa-se nove folds para treinamento e um para teste. Em cada par treina-teste, os dados de treinamento e teste são diferentes do par treinamento-teste anterior. Desta forma, ao final de 10 execuções, é calculada a média geométrica do desempenho do classificador e a média da taxa de redução de redução das características.

A implementação do NSGA-II foi obtida através do *framework jMetal* [9] e a implementação dos classificadores foi obtida pelo pacote de *software Waikato Environment for Knowledge Analysis* (Weka) [10].

RESULTADOS E/OU DISCUSSÃO

Os resultados obtidos neste trabalho foram obtidos a partir da aplicação do NSGA-II na seleção dos melhores genes de cada organismo modelo. Lembrando que esta seleção deve considerar o melhor balanceamento entre acurácia e taxa de redução. Os classificadores utilizados foram *Naive Bayes*, *K-Nearest Neighbor* e *J48*.

Os resultados estão descritos na Tabela 3, os quais são a média e desvio padrão dos 10 folds. O termo Normal significa que é o conjunto de dados original. O termo T(X) significa que foi aplicado um threshold de X nos dados, ou seja, para que um determinado gene de um organismo modelo seja considerado na base de dados, é necessário que este gene tenha frequência de no mínimo X. Exemplificando. Um threshold T(3) significa que apenas genes com frequência 3 ou mais foram considerados na base de dados.

Vale destacar que os valores alcançados sem seleção de características são importantes para serem comparados e medir se o método de seleção foi efetivo ou se houve diminuição considerável nos valores de acurácia.

Tabela 3. Resultados de acurácia.

Base de dados	Método	Classificador		
		J48	KNN	NB
Drosophila melanogaster	Sem seleção	Normal 37,75 ± 15,8	T(3) 57,63 ± 22,2	T(3) 55,01 ± 7,5
	Com seleção	T(3) 25,89 ± 18,3	Normal 56,83 ± 17,1	Normal 48,29 ± 22,2
Mus musculus	Sem seleção	T(4) 43,35 ± 23,9	Normal 56,01 ± 12,4	T(3) 47,29 ± 18,5
	Com seleção	Normal 38,0 ± 28,1	Normal 55,85 ± 11,1	Normal 44,91 ± 25,9
Caenorhabditis elegans	Sem seleção	Normal 43,39 ± 10,7	Normal 58,66 ± 7,1	T(3) 50,34 ± 6,0
	Com seleção	Normal 41,26 ± 7,9	Normal 57,25 ± 5,3	Normal 45,98 ± 7,8
Saccharomyces cerevisiae	Sem seleção	Normal 25,84 ± 23,3	T(5) 48,93 ± 21,3	Normal 22,54 ± 24,4
	Com seleção	Normal 16,27 ± 21,4	Normal 37,33 ± 27,7	Normal 24,31 ± 26,7

A partir dos resultados da Tabela 3, foi aplicado o teste estatístico de Wilcoxon com nível de confiança de 95%, a fim de verificar se há diferença significativa entre os resultados. A análise estatística foi realizada apenas nos melhores resultados, com e sem seleção de características, para cada base de dados. Estes resultados estão destacados em cinza na Tabela 3. Os resultados da análise estatística são: Drosophila melanogaster (p=0,9218); Mus musculus (p=0,4989); Caenorhabditis elegans (p=0,6250); Saccharomyces cerevisiae (p=0,1386).

Analisando estes resultados, constata-se que o valor de p obtido em todas as bases de dados foi maior que o nível de significância de 0,05. Isto quer dizer que as diferenças dos resultados não são significativamente diferentes, ou seja, não se deve rejeitar a hipótese nula. Em

outras palavras, a aplicação do algoritmo genético para selecionar as características dos organismos modelo não contribuiu para aumentar o desempenho dos classificadores.

No entanto, ainda é necessário analisar a taxa de redução obtida pelo algoritmo genético. As seguintes taxas de redução foram obtidas (somente para os melhores valores para cada base de dados): Drosophila melanogaster (Normal+KNN) 56%; Mus musculus (Normal+KNN) 54%; Caenorhabditis elegans (Normal+KNN) 53%; Saccharomyces cerevisiae (Normal+KNN) 53%.

Os resultados de redução da quantidade de características em cada base de foi maior que 50%, ou seja, mais da metade das características foram eliminadas e mesmo assim, o desempenho dos classificadores não foi prejudicado, como demonstrado na Tabela 3. Esta redução pode contribuir para a diminuição do custo computacional dos classificadores, além de contribuir para a análise dos genes mais propensos à longevidade ou a anti-longevidade de cada organismo modelo.

CONSIDERAÇÕES FINAIS

Este trabalho considerou o problema de seleção de características dos organismos modelo Drosophila melanogaster, Mus musculus, Caenorhabditis elegans e Saccharomyces cerevisiae, utilizando um algoritmo genético multiobjetivo, o NSGA-II, o qual possuía dois objetivos a serem otimizados, a taxa de redução das características e a acurácia dos classificadores.

De acordo com os resultados, a seleção genética de características não melhorou o desempenho dos classificadores, mas por outro lado, conseguiu reduzir em mais de 50%, para todas as bases de dados, a quantidade de características. Além disso, como trabalhos futuros, será realizada uma análise detalhada dos genes mais selecionados em cada base de dados, com o objetivo de descobrir quais deles são mais determinantes à longevidade ou à anti-longevidade.

REFERÊNCIAS

[1] Ashburner et al. Gene ontology: tool for the unification of biology. **Nat Genet.** vol.25, n.1, pp. 25-9, 2000.

[2] The Gene Ontology resource: enriching a GOld mine. **Nucleic Acids Research**. vol.49, n.D1, pp.D325-D334, 2021.

[3] Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., Diana, E., Lehmann, G., Toren, D., Wang, J., Fraifeld, V. E., de Magalhaes, J. P. Human Ageing Genomic Resources: new and updated databases. **Nucleic Acids Research**. vol.46, n.D1, pp.D1083-D1090, 2018.

[4] Wan et al. Predicting the Pro-Longevity or Anti-Longevity Effect of Model Organism Genes with New Hierarchical Feature Selection Methods. **IEEE/ACM Trans. Comput. Biol. Bioinformatics**. vol.2, n.2, pp. 262–275, 2015.

[5] L. R. Campos, M. G. Pires. An Empirical Comparison of Hierarchical and Ranking-Based Feature Selection Techniques in Bioinformatics Datasets. **V Symposium on Knowledge Discovery, Mining and Learning**. pp. 113-120, 2017.

[6] Deb et al, A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. **IEEE Transactions on Evolutionary Computation**. vol. 6, n. 2, pp. 182-197, 2002.

[7] Tom Fawcett, An introduction to ROC analysis. **Pattern Recognition Letters**, vol. 27, pp.861–874, 2006.

[8] A. M. Coroiu. Tuning model parameters through a Genetic Algorithm approach. **IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)**. pp. 135-140, 2016.

[9] J.J. Durillo e A.J. Nebro. jMetal: a Java Framework for Multi-Objective Optimization. **Advances in Engineering Software**. vol. 42 pp. 760-771, 2011.

[10] Eibe Frank, Mark A. Hall, e Ian H. Witten. The WEKA Workbench, Data Mining: Practical Machine Learning Tools and Techniques. **Morgan Kaufmann.** ed. 4, 2016.