UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXV SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2021

SELEÇÃO GENÉTICA DE CARACTERÍSTICAS DE BASE DE DADOS DE ORGANISMOS MODELO

José Gabriel Gomes dos S. 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, DEXA, Universidade Estadual de Feira de Santana, e-mail: fcbertoni@uefs.br

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

INTRODUÇÃO

Os mecanismos do processo biológico do envelhecimento dos organismos têm sido um grande mistério para os pesquisadores, uma vez que se tem conhecimento de que o envelhecimento é inevitável, mas não se sabe ao certo os fatores que possivelmente levam a esse acontecimento nos organismos. Entretanto, apesar de algumas pesquisas revelarem alguns fatores que estão envolvidos com o envelhecimento, como, por exemplo, a restrição calórica que estende a longevidade de algumas espécies, esse processo biológico ainda é pouco entendido [1].

Assim, recentemente os dados de organismos relacionados ao envelhecimento tem crescido exponencialmente, disponíveis de forma pública na *The Gene Ontology Resources* (GO) [2] e *Human Ageing Genomic Resources* (HAGR) [3], além de outros domínios em que esses dados podem ser encontrados. Desse modo, com o objetivo de entender o processo de envelhecimento, aplicam-se abordagens de *Mineração de Dados* [1, 5] nos dados para extrair informações que possam explicar este processo.

Portanto, neste contexto, este trabalho utiliza as bases de dados GO e HAGR dos organismos modelo *Drosophila melanogaster* (mosca), *Mus musculus* (camundongo), *Caenorhabditis elegans* (minhoca) e *Saccharomyces cerevisiae* (levedura), com o objetivo de selecionar as características desses organismos, para descobrir quais características possuem impacto no processo de envelhecimento. A seleção é realizada por um *Algoritmo Genético* [4], em conjunto com um algoritmo de *Aprendizado de Máquina* [5], o qual tem o papel de avaliar o desempenho de um classificador ao ser utilizado na classificação dos organismos modelo em pró-longevidade ou anti-longevidade, considerando as características selecionadas pelo algoritmo genético.

METODOLOGIA

Para a elaboração deste trabalho foi necessária a composição do conjunto de dados dos genes e instâncias dos organismos modelo *Drosophila melanogaster* (mosca), *Mus musculus* (camundongo), *Caenorhabditis elegans* (minhoca) e *Saccharomyces cerevisiae*

(levedura). A metodologia utilizada para essa composição foi juntar os dados de duas bases de dados: GO e HAGR [1,6]. Essas bases de dados fornecem as informações dos quatro modelos de organismos supracitados.

Para realizar a seleção das características foi usado o algoritmo genético *Non-dominated Sorting Genetic Algorithm II* (NSGA-II) [10], implementado no *framework JMetal [7]*. Cada cromossomo foi definido para representar todas as características dos organismos modelo. O valor 0 indica a ausência de um determinado gene (característica) e o valor 1 indica a presença. Para trabalhar em conjunto com o algoritmo genético na seleção de características foram utilizados três classificadores do pacote de *software Waikato Environment for Knowledge Analysis* (*Weka*) [8]. Estes classificadores foram usados para avaliar o conjunto de características selecionado. Os classificadores foram *Naive Bayes* (NB), K-Nearest Neighbor (KNN) e *J48*.

Por conseguinte, foram determinados dois objetivos para serem otimizados pelo NSGA-II: a média geométrica entre a taxa de verdadeiros positivos e verdadeiros negativos, e a taxa de redução de características. Dessa forma, o NSGA-II teve a responsabilidade de maximizar os dois objetivos, de modo que, ao final da execução do algoritmo, obtivesse a menor quantidade de características e a maior média geométrica. Em outras palavras, isso significa que a tarefa é selecionar o menor conjunto de características possível e obter classificadores que conseguissem classificar corretamente os organismos modelos em duas classes pré-definidas: pró-longevidade ou anti-longevidade, sem prejuízo em relação à média geométrica.

A configuração dos hiperparâmetros do NSGA-II foi realizada usando a metodologia conhecida como *grid search* [9], que é uma forma de escolher os melhores parâmetros realizando uma pesquisa em grade, testando um conjunto de parâmetros e escolhendo os melhores. Os parâmetros avaliados foram a taxa de cruzamento, seleção e o tamanho da população. O operador de cruzamento definido foi o *Half Uniform Crossover* (HUX), o de seleção foi o Torneio Binário e a mutação foi o *Bit-Flip Mutation*. Todos estes operadores são amplamente utilizados na literatura.

Na realização dos experimentos, todos os conjuntos de dados foram divididos em mais 8 conjuntos baseados em um *threshold* que visa limitar os genes que aparecem pelo menos uma quantidade igual ou maior que esse *threshold* [1]. Essa técnica é utilizada para tentar diminuir a quantidade de características, considerando que esses genes são irrelevantes para a classificação.

Além disso, cada conjunto de dados foi dividido em 10 *folds* de treino e 10 *folds* de teste, a fim de realizar a técnica de *cross validation*, a qual os classificadores são treinados e testados para 10 *folds*. Esse processo visa evitar a geração de resultados equivocados, assegurando um resultado confiável. Vale ressaltar que durante o período em que o algoritmo genético está procurando um conjunto de soluções, apenas os *folds* de treino são utilizados e redivididos, impossibilitando que durante o processo de avaliação dos cromossomos do NSGA-II tenha contato com os *folds* de teste, garantindo um resultado sem enviesamento.

Finalmente, é extraída a solução com a maior média geométrica do conjunto de soluções ótimas encontradas pelo NSGA-II. Assim, esses genes são testados e é extraída sua média geométrica e um desvio padrão advinda da *cross validation* entre os *folds* de treino e teste, indicando se os classificadores estão sendo precisos ou estão apenas

fazendo decisões aleatórias. Dessa forma, todo esse processo é feito com todos os 9 conjuntos de dados (8 a partir da técnica de *threshold* e um com os dados originais). Como são quatro organismos modelo, há um total de 36 arquivos de dados.

RESULTADOS E DISCUSSÃO

Com todos os conjuntos de dados prontos para serem utilizados, classificadores, medidas de qualidade da solução e o NSGA-II implementados, o trabalho ainda se encontra na fase de *grid search*. Para executar essa fase, todo o trabalho precisa ser executado muitas vezes e isso leva muito tempo de execução.

Além disso, há ainda um ajuste que precisa ser feito, que é a definição de um bom critério de parada do NSGA-II. Existe na literatura a proposta de vários índices de progresso para analisar a evolução de um algoritmo genético multiobjetivo, e ainda não concluímos esse estudo, que impacta diretamente no critério de parada.

Desse modo, é necessário terminar a fase de *grid search* para que seja possível realizar as experimentações com os melhores parâmetros possíveis do NSGA-II, visando obter bons resultados.

CONSIDERAÇÕES FINAIS

O presente trabalho possui como resultados concretos a etapa de pré-processamento e seleção dos dados da *Gene Ontology Resources* e *Human Ageing Genomic Resources* para a composição dos conjuntos de dados de quatro organismos modelos: *Drosophila melanogaster* (mosca), *Mus musculus* (camundongo), *Caenorhabditis elegans* (minhoca) e *Saccharomyces cerevisiae* (levedura). Há também a implementação do NSGA-II e seus operadores genéticos para a seleção de características, além da integração com os classificadores Naive Bayes, K-Nearest Neighbor e J48, utilizando o pacote WEKA.

O trabalho ainda não possui resultados sobre a seleção das características e a análise do desempenho dos classificadores, pois a metodologia *grid search* é muito custosa computacionalmente, e esta etapa ainda não foi finalizada.

REFERÊNCIAS

[1] 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.

[2] Ashburner et al. Gene ontology: tool for the unification of biology, Nat Genet., vol.25, pp. 25-29, 2000.

[3] Tacutu et al. Human Ageing Genomic Resources: new and updated databases, Nucleic Acids Research, vol.46, pp. (D1) D1083-D1090.

[4] Mitchell, M, An Introduction to Genetic Algorithms, The MIT Press, 1996.

[5] Witten et al. Data Mining: Practical Machine Learning Tools and Techniques, USA: Morgan Kaufmann, pp. 3-9, 2011.

[6] 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.

[7] 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.

[8] Eibe Frank, Mark A. Hall, e Ian H. Witten, The WEKA Workbench, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Fourth Edition,

2016.

[9] 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.

[10] 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.