SELEÇÃO DE INSTÂNCIAS DE GRANDES BASES DE DADOS USANDO ALGORITMOS EVOLUTIVOS MULTIOBJETIVO

João Paulo Santos Sena[1]; Matheus Giovanni Pires[2];

1. Bolsista PIBIC/CNPq, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: joaopaulo761@gmail.com

2. Orientador, Departamento de Exatas, Universidade Estadual de Feira de Santana, e-mail: mgpires@ecomp.uefs.br

PALAVRAS-CHAVE: redução de dados; seleção de instâncias; algoritmos genéticos multiobjetivo.

INTRODUÇÃO

Os Sistemas Baseados em Regras *Fuzzy* (SBRF) têm sido amplamente usados para a resolução de diversos tipos de problemas, tais como, controle (Leephakpreeda, 2011), modelagem (Pedrycz, 1996), classificação (Ishibuci, 1995). A maneira mais comum para a aquisição do conhecimento de um SBRF é a partir de dados numéricos, os quais representam amostras ou exemplos do problema. As formas mais bem-sucedidas de extração automática de conhecimento a partir de dados para a construção de SFBR são as que combinam metodologias para aprendizado de máquina com conceitos de sistemas *fuzzy*. Entre elas, destacam-se as Redes Neurais Artificiais e a Computação Evolutiva (Cordón et al, 2001).

Os Algoritmos Genéticos Multiobjetivo (AGMO), vêm demonstrando ser uma poderosa ferramenta para a construção automática (ou projeto automático) de SBRF. No entanto, este processo é fortemente influenciado pela quantidade de instâncias e características presentes nas bases de dados, que afetam o tamanho do espaço de busca e o tempo computacional. Por isso, a redução de dados é de fundamental importância para reduzir o tempo de aprendizado do SBRF e amenizar as dificuldades durante o processo de convergência dos algoritmos evolutivos.

A redução de dados, neste caso, a seleção de instâncias, é um problema multiobjetivo, pois busca-se reduzir a base de dados, e ao mesmo tempo, manter o desempenho do classificador estável ou superior, quando comparado com a base de dados original. Portanto, este trabalho possui o objetivo de investigar a aplicação de Algoritmos Genéticos Multiobjetivo para a seleção de instâncias de grandes bases de dados.

METODOLOGIA

A seleção de instâncias foi realizada por três AGMO, sendo eles: NSGA-II (Deb *et al.* 2002), NSGA-III (Deb & Jain, 2014) e o NSGA-DO (Pimenta & Camargo, 2015). A implementação destes algoritmos foi realizada utilizando o framework *jMetal* (Durillo & Nebro, 2011). O Problema foi modelado considerando uma codificação binária do cromossomo, onde 1 indica a presença de uma instância e 0 sua ausência. Os operadores escolhidos juntamente com suas taxas para cruzamento, mutação e seleção foram *Single Point Crossover (0.9)*, *Bit Flip Mutation (0.2)* e *Binary Tournament*, respectivamente. O tamanho da população foi definido como 100 e a quantidade máxima de gerações 50. No entanto, a execução poderia ser interrompida antes das 50 gerações, caso não houvesse melhora do melhor indivíduo durante cinco gerações seguidas. Os valores dos parâmetros descritos foram definidos de forma empírica. A função de *fitness*, responsável pela

avaliação de cada cromossomo, foi definida a partir de dois objetivos, acurácia e taxa de redução, que são obtidos através das equações 1 e 2 respectivamente.

$$Acurácia = \frac{classificações\ de\ instâncias\ corretas}{quantidade\ total\ de\ instâncias} \quad (1)$$

$$Redução = \frac{quantidade\ total\ - quantidade\ selecionada}{quantidade\ total} \quad (2)$$

O classificador utilizado para o cálculo da acurácia durante o treinamento foi o *k-Nearest Neighbor* (KNN) disponível no WEKA[1], com o valor de *k* igual a 5.

Após a aplicação do AGMO para a seleção de instancias foram utilizados os classificadores KNN e o C4.5, que se baseia em uma arvore de decisões, como classificadores de testes, ambos disponíveis no WEKA[1].

Como o resultado do AGMO é um conjunto de soluções, escolheu-se a solução média entre os 2 objetivos, não priorizando a acurácia nem a redução.

RESULTADOS E/OU DISCUSSÃO

Para a realização dos experimentos foram utilizadas 37 bases de dados extraídas do repositório aberto de dados do KEEL[2]. No entanto neste resumo serão apresentados os resultados apenas de 9 bases, as quais foram selecionadas aleatoriamente, devido ao espaço limitado deste resumo. Elas foram divididas em conjuntos de treinamento e teste, utilizando a abordagem *10-fold cross validation* (Kohavi, 1995). Além disso, cada *fold* foi executado três vezes, logo, para cada base de dados foram efetuadas 30 execuções. Portanto, os resultados expressam a média destas 30 execuções. Baseado no trabalho de Fazzolari *et al.* (2013), as bases também foram reduzidas utilizando o algoritmo PBIL e os resultados foram comparados com os AGMO.

Para comparar os resultados obtidos foi utilizado o teste de Friedman (Friedman, 1940) para verificar se existem diferenças significantes entre os resultados. Uma vez constatada uma diferença significante, os resultados são comparados em pares, utilizando o teste Wilcoxon Signed-Ranks (Wilcoxon, 1945).

Os resultados obtidos estão descritos nas Tabelas 1, 2 e 3, as quais, mostram a média e desvio padrão das medidas de tempo, acurácia e taxa de redução, respectivamente. O tempo de classificação foi calculado em segundos.

De acordo com os resultados da Tabela 1, é possível constatar uma redução no tempo de classificação, considerando as bases reduzidas pelos AGMO, quando comparados com a base de tamanho original. Por outro lado, ao se comparar os tempos de classificação com a base reduzida pelo algoritmo PBIL, este proporcionou uma classificação mais rápida em relação aos AGMO.

Os dados da Tabela 1 estão relacionados com os da Tabela 2, pois a redução no tempo de classificação está diretamente relacionada com a redução das bases. Ao analisar

[1] WEKA: Waikato Environment for Knowledge Analysis. https://www.cs.waikato.ac.nz/ml/weka
[2] http://sci2s.ugr.es/keel/ datasets.php

os resultados da Tabela 2, verifica-se que a redução obtida pelos AGMO ficou em torno de 50% da base original, enquanto que o PBIL ficou em torno de 89%.

Tabela 1. Tempo de classificação

	NSGAII		NSGAIII		NSGA-DO		PBIL		Original	
	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio
coil2000	347.9	32.0	348.5	31.5	369.2	58.0	29.6	8.7	1218.9	137.2
magic	176.6	26.4	171.9	26.2	176.6	21.1	36.1	7.1	507.9	109.7
marketing	102.0	8.8	102.6	9.7	106.7	11.4	12.4	6.2	307.8	31.9
optdigits	115.1	7.6	119.8	11.6	117.7	11.2	15.6	6.9	306.2	7.7
penbased	91.6	12.6	100.5	18.7	95.8	8.8	15.7	0.5	217.1	16.4
ring	180.7	16.4	174.5	11.6	178.0	14.3	23.2	7.8	492.2	39.2
Satimage	110.5	8.0	114.1	13.5	108.8	7.5	15.8	0.4	273.5	7.9
Texture	105.2	8.0	106.3	9.4	105.2	8.1	14.2	4.7	242.1	7.9
Twonorm	84.9	7.7	86.4	8.8	86.5	7.8	10.8	7.1	229.8	12.1

Tabela 2. Taxas de redução alcançadas pelos algoritmos nas bases grandes.

	NSGAII		NSGAIII		NSGA-DO		PBIL	
	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio
coil2000	51.7%	0.4%	51.8%	0.4%	51.6%	0.4%	88.1%	0.3%
magic	50.8%	0.7%	50.8%	0.6%	50.7%	0.6%	85.5%	0.2%
marketing	51.2%	0.7%	51.1%	0.8%	50.9%	0.5%	87.5%	0.2%
optdigits	51.8%	0.4%	51.9%	0.5%	52.0%	0.6%	91.2%	0.2%
penbased	51.4%	0.4%	51.5%	0.4%	51.5%	0.3%	88.9%	0.2%
ring	51.2%	1.2%	51.7%	1.4%	51.5%	1.2%	86.7%	0.2%
satimage	51.3%	0.8%	51.1%	1.1%	51.8%	0.9%	89.5%	0.1%
texture	51.6%	0.6%	51.7%	0.8%	51.8%	0.7%	90.7%	0.3%
twonorm	51.6%	0.6%	51.6%	0.6%	51.7%	0.6%	89.6%	0.2%

Tabela 3. Acurácia das bases de dados reduzidas comparas com a base original

	NSGAII		NSGAIII		NSGA-DO		PBIL		Original	
	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio
coil2000	0.94	0.00	0.94	0.00	0.94	0.00	0.94	0.00	0.94	0.00
magic	0.85	0.01	0.85	0.01	0.85	0.01	0.84	0.01	0.85	0.01
marketing	0.30	0.02	0.30	0.02	0.30	0.02	0.36	0.02	0.31	0.02
optdigits	0.88	0.01	0.88	0.01	0.88	0.01	0.80	0.01	0.91	0.01
penbased	0.95	0.01	0.95	0.01	0.95	0.01	0.91	0.01	0.96	0.01
ring	0.89	0.01	0.89	0.01	0.89	0.01	0.85	0.02	0.91	0.01
satimage	0.85	0.02	0.85	0.01	0.85	0.01	0.84	0.01	0.86	0.01
texture	0.91	0.01	0.91	0.01	0.90	0.01	0.85	0.02	0.93	0.01
twonorm	0.84	0.01	0.84	0.01	0.84	0.01	0.82	0.01	0.85	0.01

Por fim é necessário analisar se a redução das bases impactou no desempenho do classificador. Na Tabela 3 é possível verificar que a redução dos dados não impactou na classificação, pois os resultados são próximos ou iguais quando comparados com a base de dados de tamanho original. De acordo com o método estatístico de Friedman, não

houveram diferenças estatísticas na acurácia, quando comparados as bases de dados reduzidas pelos AGMO e PBIL com a original.

CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo analisar o impacto da classificação de bases de dados reduzidas por AGMO. Os resultados mostram que todas as bases de dados foram reduzidas em aproximadamente 50% e apesar da redução o desempenho dos classificadores não foi prejudicado de acordo com o método de Friedman. Deste modo, a seleção de instâncias, utilizando AGMO torna-se uma alternativa para obter um subconjunto reduzido de dados com o mesmo desempenho do conjunto original.

Como trabalhos futuros, as bases de dados reduzidas neste trabalho serão utilizadas para a construção de Sistemas Baseados em Regras Fuzzy.

REFERÊNCIAS

FRIEDMAN, Milton (1940). A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, v. 11, n. 1, p. 86-92.

WILCOXON, Frank (1945) Individual comparisons by ranking methods. Biometrics bulletin, v. 1, n. 6, p. 80-83.

ISHIBUCHI, H., NOZAKI, K., YAKAMOTO, N. e TANAKA, H. (1995). Selecting Fuzzy If-Then Rules for Classification Problems using Genetic Algorithms. IEEE Transactions on Fuzzy Systems, vol.3, n.3, pp.260-270.

KOHAVI, Ron et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial intelligence, p. 1137-1145.

PEDRYCZ, W. (1996). Fuzzy modelling: Paradigms and practice, Kluwer Academic Publishers.

CORDÓN, O., HERRERA, F. e VILLAR, P. (2001). Generating the Knowledge Base of a Fuzzy Rule-Based System by the Genetic Learning of the Data Base. IEEE Transactions on Fuzzy Systems, vol.9, n.4, pp.667-674.

DEB, K., PRATAP, A., AGARWAL, S., MEYARIVAN, T. (2002). A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, vol.6, n.2, pp.182-197.

LEEPHAKPREEDA, T. (2011). Fuzzy logic based PWM control and neural controlled-variable estimation of pneumatic artificial muscle actuators. Expert Systems with Applications, vol.38, n.6, pp.7837-7850.

JUAN A. J. N., DURILLO, J. (2011). jMetal: A Java framework for multiobjective optimization, Advances in Engineering Software, p. 760–771.

FAZZOLARI, M., GIGLIO, B., ALCALÁ, R., MARCELLONI, F. HERRERA, F. (2013). A study on the application of instance selection techniques in genetic fuzzy rule-based classification systems: Accuracy-complexity trade-off. Knowledge-Based Systems, vol.54, pp.32-41.

DEB, K., JAIN, H. (2014) An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems with Box Constraints. IEEE Transactions on Evolutionary Computation, vol.18, no.4.

PIMENTA, A. H. M. e CAMARGO, H. A. (2015). NSGA-DO: Non-Dominated Sorting Genetic Algorithm Distance Oriented. IEEE International Conference on Fuzzy Systems, pp.1-8.