Mineração de Regras de Classificação de Câncer utilizando Nondominated Sorting Genetic Algorithm II (NSGA-II)

Vitor L. Coelho[1], Claudomiro de S. de Sales Junior[2]

[1]Laboratório Nacional de Computação Científica (LNCC)
25651-075 – Petrópolis – RJ – Brasil

[2]Faculdade de Computação – Universidade Federal do Pará (UFPA)
66075-110 – Belém – Pará – Brasil

vitorlc@lncc.br, cssj@ufpa.br

Abstract. *This paper presents a genetic algorithm based on Nondominated Sorting Genetic Algorithm II (NSGA-II) for mining IF-THEN rules for classification of samples of gene expression of cancer cell database (NCI60) with 1000 genes and nine cancer classes. Rules are formed by the total of 30 genes and classified the database with accuracy greater than 98%.*

Resumo. *Este trabalho apresenta um algoritmo genético baseado no Nondominated Sorting Genetic Algorithm II (NSGA-II) para mineração de regras do tipo IF-THEN para classificação de amostras de células cancerígenas de uma base de expressões de 1000 genes (NCI60) e nove classes de câncer. As regras são compostas pelo total de 30 genes e classificaram a base de dados com precisão superior a 98%.*

1. Introdução

O volume crescente de dados de origem médica coletados continuamente pelo monitoramento dos parâmetros fisiológicos dos pacientes e resultantes de pesquisas é um desafio para obtenção de conhecimento, identificação de padrões e classificação. Os algoritmos genéticos (AG) são abordagens da computação evolucionária de busca e otimização global inspirada em mecanismos de seleção e genética natural [Goldberg 1989], que podem ser aplicados na tarefa de mineração de conhecimento e utilizado para obtenção de regras de classificação. O trabalho proposto empregou o algoritmo genético multiobjetivo *Nondominated Sorting Genetic Algorithm* II (NSGA-II) [Deb *et al.* 2002] para obtenção de regras do tipo IF-THEN para classificação de câncer da base de expressões gênicas NCI60 [Ross *et al*. 2000].

2. Materiais e Métodos

O minerador de regras utilizado neste trabalho foi desenvolvido na linguagem de programação JAVA. O ambiente evolutivo foi baseado no trabalho de Amaral (2007), os parâmetros genéticos foram definidos experimentalmente e os operadores de reinserção e seleção foram baseados no algoritmo multiobjetivo NSGA-II. Em seguida, um classificador na linguagem Python foi construído contendo as regras mineradas para avaliar a generalização e precisão das regras.

2.1. *Nondominated Sorting Genetic Algorithm* II (NSGA-II)

O NSGA-II é um AG multiobjetivo desenvolvido por Deb *et al.* (2000). Nesse algoritmo, a população é ordenada por não dominância e cada indivíduo é organizado

em frentes de Pareto. Cada solução recebe um *fitness* ou *rank* igual ao seu nível de não dominância (1 é o melhor nível, 2 é o segundo melhor e assim por diante), utilizando o mecanismo de *fast nondominated sorting*. Além disso, calcula-se também a diversidade de uma frente, medida pela distância de multidão ou *crowding distance* com complexidade $O(MN^2)$, onde M é o número de objetivos e N é o tamanho da população. A cada ciclo, uma nova frente é criada com as soluções retiradas do conjunto, caracterizando-se como um AG elitista.

2.2. Funções de avaliação

Neste trabalho, a avaliação das regras pelo NSGA-II constitui um problema de minimização de três funções:

$$f = 1 - Es \tag{1}$$

$$g = 1 - Se \tag{2}$$

$$m = \frac{f+g}{2} \tag{3}$$

Se (sensibilidade) e *Es* (especificidade) [Amaral 2007] são dois indicadores comumente utilizados em domínios médicos, considerando os resultados verdadeiramente positivos (*vp*), verdadeiramente negativos (*vn*), falso positivos (*fp*) e falso negativos (*fn*), definidos como:

$$Es = \frac{vn}{vn+fp} \tag{4}$$

$$Se = \frac{vp}{vp+fn} \tag{5}$$

2.3. Codificação do Indivíduo ou Cromossomo

O indivíduo ou cromossomo foram modelados como em Amaral (2007): é representado por uma lista de *n* genes, onde *n* é o número de genes da base de expressões gênicas utilizada. Ou seja, para uma base de expressões gênicas com 10 genes, cromossomo será composto por uma lista de 10 genes, como ilustrado na figura 1:

Gene_1				...	*Gene_10*			
I	*P*	*V*	*O*	...	*I*	*P*	*V*	*O*

Figura 1. Cromossomo composto por dez genes.

Cada gene do cromossomo é subdividido em quatro campos: *I* (índice), *P* (peso), *O* (operador) e *V* (valor). O campo *I* é o índice do gene na base de dados investigada. O campo *P* é uma variável limite responsável pela inserção ou exclusão de um gene do cromossomo, variando de 0 (zero) a 10 (dez). Caso um gene do cromossomo possua peso inferior ao limite pré-definido, não irá compor a regra. O campo *O* varia entre os operadores de comparação $<$ e $\geq$. O campo *V* é um valor real que varia entre o menor e maior valor de expressão gênica da base de dados investigada.

2.4. Operadores Genéticos

Na seleção dos pais para cruzamento, foi utilizado o método de torneio de *tour* de tamanho 3 (três). O novo pai é determinado pelo operador de comparação de multidão ou *crownded-comparison operator* do NSGA-II. Foi utilizado o método de cruzamento de dois pontos com probabilidade igual a 100%. O operador de mutação varia para cada campo do gene do cromossomo filho, com probabilidade igual a 10%. Para o campo *P*, decrementa-se ou incrementa-se uma unidade. Para o campo *O*, troca-se o operador "≥" por "<", e vice-versa. Para o campo *V*, sorteia-se um incremento ou decremento de 0,1.

2.5. Bases de expressões gênicas analisadas

Foram analisadas 5 (cinco) sub-bases de dados obtidas a partir base NCI60 [Ross et al 2000]. São compostas por expressões gênicas medidas por *microarray* de 61 amostras de células cancerígenas de 9 (nove) classes: leucemia, cólon, mama, próstata, pulmão, ovário, renal, sistema nervoso central e melanoma. Ooi e Tan (2003) normalizaram a base NCI60 obtendo três conjuntos de A (1000 genes), B1(13 genes) e B4(12 genes). A partir das técnicas de classificação e predição de genes de Dudoit *et al.* (2002) e Golub *et al.* (1999), foram obtidas as sub-bases com B2 e B3 (20 e 17 genes, respectivamente). A utilização da base NCI60 incrementa dificuldade para um classificador, pois contém uma pequena quantidade de amostras para um número grande de classes.

3. Resultados

As bases B1, B2, B3 e B4 (61 amostras cada) foram utilizadas para minerar as regras. Foram divididas em três partições (P1, P2 e P3) de tamanhos aproximadamente iguais. Portanto, cada partição tem aproximadamente 20 amostras. Duas partições foram utilizadas para treinamento e a terceira partição para avaliar o nível de generalização das regras obtidas em treinamento. Para cada classe, o algoritmo genético foi executado 80 vezes, por 100 gerações para uma população 400 indivíduos. Em seguida, a base de 1000 genes foi classificada pelas regras e avaliadas somente em relação a quantidade de erros encontrados na classificação:

Tabela 1. Melhores regras e respectivos erros de classificação.

Classe	Regra	Erros – Treino	Erros – Teste
1	gene_46 < -1.3 AND gene_289 < -0.9	1	0
2	gene_289 < -0.2 AND gene_839 ≥ 0.5 AND gene_881 ≥ 0.7	0	0
3	gene_50 < -2.3 AND gene_194 < -1.1 AND gene_289 ≥ -0.3	0	0
4	gene_2 < -0.2 AND gene_485 ≥ 0.7 AND gene_890 < -0.5	0	0
5	gene_11 ≥ -1.5 AND gene_97 < 0.1 AND gene_348 < -1.5 AND gene_839 < 0.0	0	0
6	gene_2 < -1.3 AND gene_379 ≥ 0.2 AND gene_637 ≥ 0.5 AND gene_890 < -0.7 AND gene_929 ≥ 0.1	0	0
7	gene_63 ≥ 0.3 AND gene_379 < 0.9 AND gene_890 < -0.6	0	0
8	gene_177 ≥ 0.2 AND gene_336 ≥ 0.7 AND gene_865 < -0.3	0	0
9	gene_456 ≥ 0.6 AND gene_485 ≥ -1.5 AND gene_525 < -0.8 AND gene_786 < -0.6	0	0
	Total de Erros / Precisão:	**1 / 98,36%**	

Os resultados foram comparados com outros trabalhos que também usuram a base de 1000 genes e partição de treinamento e teste para validação:

Tabela 1 - Comparativos de erros encontrados para base de 1000 genes.

Referência	Nº de genes	Erros - Treinamento	Erros - teste	Erros total
Dudoit et al (2002)	40	-	8	≥ 8
Deb e Reddy (2003)	12	3	2	5
Ooi e Tan (2003)	12	4	4	8
Lin et al (2006)	15	5	4	9
Amaral K1 (2007)	20	2	6	8
Amaral K2 (2007)	22	1	3	4
Este trabalho	30	1	0	0

4. Conclusões e Perspectivas

A aplicação do NSGA-II para mineração de regras de classificação gerou regras do tipo IF-THEN com 98,36% precisão utilizando 30 genes de uma base formada por 1000 genes. Além disso, foram obtidas regras com sensibilidade e especificidade igual a 100% em treinamento e teste para sete das nove classes. A geração de regras deste tipo permite a obtenção de conhecimentos que relacionem a expressão de cada gene com um tipo determinado de tecido celular. Como trabalho futuro, sugere-se a utilização de bases de expressão gênicas medidas pela técnica de RNA-Seq, por ser mais precisa quanto aos níveis de transcrição que o *microarray* [Zhong *et al.* 2009]. Outra abordagem seria minerar bases de miRNAs [Breving e Esquela-Kerscher 2010] para predição de genes de miRNA e classificação de características.

Referências

Amaral, L. R. do. (2007) "Mineração de regras para classificação de oncogenes medidos por Microarray utilizando algoritmos genéticos". 123f. Dissertação de mestrado do curso de Ciências da Computação. Universidade Federal de Uberlândia, Minas Gerais, Brasil.

Deb, K. et al. (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. In *IEEE Transactions on Evolutionary Computation*, vol. 6, n. 2, p.182-197.

Deb, K.; Reddy, A. R. (2003) Classification of two and multi-class cancer data reliably using multi-objective evolutionary algorithms. In: *KanGAL Report*.

Dudoit, S. *et al.*. (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. In *Journal of the American Statistical Association*, 97(457).

Goldberg, D. E. (1989) Genetic Algorithms in Search, Optimization e Machine Learning, Addison-Wesley, 1ª edição.

Golub, T. R. (1989) Molecular classification of cancer: class discovery and class prediction. In *Science*, 288.

Lin, T-C; *et al.*(2006) Pattern classification in DNA microarray data of multiple tumor types. *Patterns Recognition*, vol. 39, p. 2426-2438.

Ross, D. T. *et al.*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. In *Nature Genetics*, vol. 24, p. 227-235.

Ooi, C. H.; Tan, P. (2003) Genetic algorithms applied to multi-class prediction for the analisys of gene expression data. In *Bioinformatics*, p. 37-44.