Uso de Critérios Multiobjetivo Baseados em Enxames na Escolha dos Melhores Métodos para Seleção de Atributos em Microarranjos Gênicos

Rodolfo Garcia[1], Júlio Cesar Nievola[1]

[1]Programa de Pós-Graduação em Informática (PPGIa) – Pontifícia Universidade Católica do Paraná (PUCPR)
CEP 80.215-901 – Curitiba – PR – Brasil

{rodolfobbgarcia,nievola}@ppgia.pucpr.br

Abstract. *Microarray technique is responsible for extracting information about expression of large numbers of genes. Knowing that only few genes participate in the activation or inhibition of each characteristic, this paper aims to evaluate methods to select relevant features in genes datasets obtained by microarray technique. The criteria used to measure the quality of each selection method characterize a multiobjective problem and the proposed swarm technique called MOPSO evaluates the methods and defines satisfactorily the best solutions, facilitating the genes expression datasets analysis.*

Resumo. *A técnica de microarranjo é responsável pela extração de informação sobre a expressão de grande quantidade de genes. Sabendo-se que poucos são os genes que fazem parte da ativação ou inibição de uma característica, este trabalho objetiva avaliar métodos para selecionar os atributos mais relevantes em bases de dados obtidas pela técnica de microarranjo. Os critérios utilizados para medir a qualidade de cada método de seleção caracterizam um problema multiobjetivo e uma proposta técnica baseada em enxames, o MOPSO, avalia os métodos e define satisfatoriamente as melhores soluções, facilitando a análise de bases de dados contendo expressões gênicas.*

1. Introdução

O aumento da quantidade de informações resultantes de pesquisas relacionadas ao Projeto Genoma Humano faz com que análises manuais feitas por cientistas demorem até anos para serem completadas. Para agilizar esse trabalho, necessita-se da alta velocidade computacional, além de sua capacidade de analisar simultaneamente vários genes [Dy 2008].

A análise das expressões gênicas é de extrema importância para diagnosticar doenças e identificar o estado de determinados genes em ciclos específicos [NIH 2001]. A técnica de microarranjo é responsável pela extração de informação sobre a expressão de grande quantidade de genes em relação a uma determinada característica. Essa técnica tem sido utilizada para buscar os possíveis genes envolvidos em doenças como o câncer, Alzheimer, Parkinson e diabetes [D'haesleer 2005].

Apesar de obter informações simultaneamente de milhares de genes, sabe-se que poucos são aqueles que participam da ativação ou inibição de uma característica, o que

torna necessário o uso de métodos que possam selecionar os genes mais relevantes e obter resultados mais compreensíveis [Handl e Knowles 2007]. Este trabalho objetiva avaliar métodos de seleção de atributos, o C-FOCUS e o *Relief*-F, em bases de dados de expressões gênicas obtidas pela técnica do microarranjo. Para medir a qualidade de cada seleção são usados critérios como a quantidade de genes selecionados e o índice Jaccard, resultante do agrupamento das instâncias de uma base pelo algoritmo K-*means*.

Otimizar vários critérios caracteriza um problema multiobjetivo. Neste trabalho, a avaliação dos métodos de seleção aqui utilizados e a definição dos métodos ótimos para cada base de dados são feitas através de uma técnica multiobjetivo baseada em enxames, o *Multi-Objective Particle Swarm Optimization*. Um método de seleção é tido como ótimo se gerar uma base de dados composta pelos atributos mais relevantes, o que torna sua análise mais simples, já que seus atributos representam da melhor forma possível cada classe existente.

A seção 2 expõe a importância da seleção de atributos na análise de bases de dados gênicos e os algoritmos usados neste trabalho. Na seção 3 é mostrada a etapa do agrupamento, sua influência sobre a seleção de atributos, o método K-*means* e o índice Jaccard. A seção 4 apresenta a técnica *Multi-Objective Particle Swarm Optimization* baseado na frente de Pareto e seu funcionamento. Na seção 5 é apresentada detalhadamente a proposta deste trabalho. Em seguida, na seção 6, são descritos os experimentos, juntamente com os resultados e, na seção 7, as últimas considerações.

2. Seleção de Atributos

Muitas aplicações do mundo real, como análise de expressões gênicas obtidas pelos microarranjos, apresentam grande quantidade de atributos [Dash et al. 2002]. O problema é que, apesar do microarranjo obter informações de milhares de genes simultaneamente, sabe-se que poucos são aqueles que fazem parte da ativação ou inibição de uma característica.

Para os genes, ou atributos, que não melhoram os resultados das análises são atribuídos o nome de redundantes ou irrelevantes e devem ser removidos [Kohavi e John 1997]. A redução da quantidade de genes pode garantir resultados mais compreensíveis, obtidos em menos tempo e com altas taxas de acerto, já que são usados apenas atributos relevantes [Dy 2008].

Os algoritmos utilizados neste trabalho fazem parte do método de filtro e podem ser utilizados por qualquer outro algoritmo pertencente às demais etapas da mineração de dados, como o agrupamento. Segundo [Yu e Liu 2004], esses métodos são úteis em bases de grande dimensionalidade, como no problema de seleção de genes.

O primeiro algoritmo, o C-FOCUS, é um extensão do FOCUS para ser usado em atributos nominais e discretos [Azofra, Sanchez e Peña 2003]. Seu objetivo é eliminar inconsistências e redundâncias retornando o menor subconjunto ótimo, chamado *Min-Features*, formado apenas por atributos relevantes [Kohavi e John 1997].

Outro método usado é o *Relief*-F, uma extensão do *Relief* que envolve mais de duas classes, o qual escolhe uma quantidade de atributos especificada pelo usuário. Quanto maior o poder de um atributo em distinguir instâncias de classes diferentes, maior é a probabilidade dele ser selecionado [Kononenko e Sikonja 2008].

3. Agrupamento

A etapa de agrupamento objetiva organizar um conjunto de instâncias de forma que aquelas com comportamentos similares pertençam ao mesmo grupo [Xu e Wunsch 2005]. Um bom agrupamento é sinônimo de que os atributos das bases de dados são relevantes e suas instâncias definem bem as classes a que pertencem. Logo, essa etapa pode servir como avaliador dos métodos de seleção de atributos. Nas bases de dados de expressão gênica, instâncias com comportamentos similares podem representar a mesma característica.

O algoritmo usado neste trabalho é o K-*means,* que é baseado em centróide, bastante popular e simples. A partição resultante é formada pela quantidade de grupos especificada pelo usuário e cada objeto pertence a um único grupo [Berkhin 2002].

Neste trabalho, a qualidade do agrupamento é calculada pelo índice Jaccard, que compara o agrupamento obtido (A_o) com um agrupamento de referência (A_r). Segundo [Faceli, Carvalho e Souto 2005], duas instâncias são ditas:

- SS, se pertencem ao mesmo grupo em A_o e ao mesmo grupo em A_r;

- SD, se pertencem ao mesmo grupo em A_o e a grupos diferentes em A_r;

- DS, se pertencem a grupos diferentes em A_o e ao mesmo grupo em A_r;

Em que a_1, a_2 e a_3 são, respectivamente, a quantidade de pares SS, SD e DS. O índice Jaccard, calculado pela Equação 1, pertence ao intervalo [-1, 1] e o valor máximo representa agrupamentos idênticos [Faceli, Carvalho e Souto 2005].

$$jaccard = \frac{a_1}{(a_1 + a_2 + a_3)} \tag{1}$$

4. Técnicas Multiobjetivo

O uso de vários critérios facilita na avaliação dos métodos de seleção, pois nem sempre o melhor método em um único critério é o melhor método possível. Um problema de otimização que trata simultaneamente de vários critérios, ou funções objetivo, é chamado de multiobjetivo. Em [Suresh et al. 2009] encontra-se a formulação matemática para um problema de otimização multiobjetivo composto por m objetivos, cada um com n variáveis de decisão, que são as soluções para o problema (Equação 2).

$$Otimizar\ \vec{Y} = \vec{f}(\vec{Y}) = \left(f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)\right) \tag{2}$$

No problema multiobjetivo, uma variável de decisão pode não ser a melhor solução para todos os objetivos, sendo possível eleger mais de uma variável como melhor solução. Essa é a finalidade da técnica de enxames *Multi-Objective Particle Swarm Optimization* (MOPSO) baseada na frente de Pareto.

O MOPSO é um algoritmo evolucionário que se baseia na simulação de vôo que uma população de pássaros faz ao buscar alimentos [Kennedy e Eberhart 1995]. Além de ser simples, outro fator que o tornou popular foi sua eficiência em varias aplicações, produzindo bons resultados com baixo custo computacional [Chuang et al. 2008].

Cada membro da população, ou enxame, é chamada de partícula e representa uma possível solução que se movimenta pelo espaço de busca. Sua posição $\vec{x}_i(t)$,

(Equação 3), em um determinado tempo t é definida pela sua posição no tempo $t-1$ acrescida pelo valor do operador $\vec{v}_i(t)$, que simula a velocidade.

$$\vec{x}_i(t) = \vec{x}_i(t-1) + \vec{v}_i(t) \quad (3)$$

As partículas que estiverem situadas em regiões promissoras são tidas como as melhores soluções e são chamadas de líderes, podendo ser mais de uma porque se trata de um problema multiobjetivo. Uma partícula é eleita líder se, primeiramente for definida para um problema de maximização ou minimização, e satisfizer a chamada regra de dominância de Pareto [Coello, Lamont e Vldhuizen 2007].

Dadas duas soluções $\vec{x}$ e $\vec{y}$, pertencentes ao espaço de soluções π, dizemos que $\vec{x}$ domina $\vec{y}$, ou $\vec{x} \prec \vec{y}$, em um problema de minimização com n critérios, se:

- Existe pelo menos um critério $i = 1,\ldots,n$ em que x_i é estritamente melhor que y_i, ou seja $x_i \prec y_i$;

- x_i não é pior que y_i em nenhum critério i, ou seja $x_i \leq y_i$;

Se não existir solução $\vec{x}' \in \pi$, onde $f(\vec{x}') \prec f(\vec{x})$, $\vec{x}$ é Pareto Ótimo e eleito um dos líder do enxame pelo MOPSO [Suresh et al. 2009]. Por sua vez, determinando o conjunto de todos os vetores do Pareto Ótimo, que geralmente consiste em um problema NP-Completo, é possível gerar a frente de Pareto [Carvalho e Pozo 2009].

O conjunto de líderes do enxame são armazenados em um repositório no qual seu conteúdo é retornado como resultado ao término da execução do algoritmo [Sierra e Coello 2006]. Dentro deste conjunto, a melhor solução deve ser selecionada pelo *decision maker*, que pode ser o próprio usuário [Coello, Lamont e Vldhuizen 2007].

Para cada partícula, um líder é escolhido para servir-lhe de guia no espaço de busca. A escolha desse guia pode ser feita pelo método sigma ($\vec{\sigma}$), apresentado em [Mostaghim e Teich 2003] que é muito eficaz em problemas multiobjetivo baseados em técnicas de enxames. O guia só é eleito se fizer parte da mesma vizinhança da partícula. Por isso, é extremamente importante estabelecer como as partículas estão interligadas, ou seja, estabelecer uma topologia [Sierra e Coello 2006].

O vetor $\vec{\sigma}$, calculado pela Equação 4, realiza a combinação dos objetivos e é composto por $\binom{m}{2}$ elementos, em que m é a dimensão do espaço objetivo. O líder cujo vetor sigma é mais próximo que o vetor sigma da partícula, e se eles pertencerem a mesma vizinhança, será eleito líder para essa partícula [Mostaghim e Teich 2003].

$$\vec{\sigma} = \begin{pmatrix} {f_1}^2 - {f_2}^2 \\ {f_1}^2 - {f_3}^2 \\ \cdots \\ {f_{m-1}}^2 - {f_m}^2 \end{pmatrix} \Big/ \left({f_1}^2 + {f_2}^2 + {f_3}^2 + \cdots + {f_{m-1}}^2 + {f_m}^2\right) \quad (4)$$

Para um problema de dois objetivos f_1, f_2, o vetor sigma terá apenas um elemento, calculado pela Equação 5.

$$\sigma = \frac{f_1(x)^2 - f_2(x)^2}{f_1(x)^2 + f_2(x)^2} \quad (5)$$

O cálculo da velocidade de uma partícula, na Equação 6, depende de sua melhor posição, $\vec{x}_{pbest_i}$, e da melhor posição do seu líder, $\vec{x}_{leader}$ [Carvalho e Pozo 2009]. Seu resultado é influenciado pelo valor da inércia W, que mede o impacto da velocidade anterior sobre a velocidade atual, das constantes positivas de aprendizagem cognitiva C_1

e de aprendizagem social C_2, que são as influências que uma partícula tem, respectivamente, sobre sua própria velocidade e sobre a velocidade da vizinhança, e pelos valores aleatórios $r_1, r_2 \in [0,1]$ [Sierra e Coello 2006].

$$\vec{v}_i(t) = W\vec{v}_i(t-1) + C_1 r_1 \left(\vec{x}_{pbest_i} - \vec{x}_i(t-1)\right) + C_2 r_2 (\vec{x}_{leader} - \vec{x}_i(t-1)) \quad (6)$$

5. Proposta

Este trabalho tem como meta avaliar, por meio do MOPSO, métodos de seleção de atributos em bases de dados obtidas pela técnica de microarranjo. Como resultado, o MOPSO seleciona os métodos de seleção ótimos e identifica a quantidade dos atributos mais relevantes para uma característica, afim de facilitar na análise da base de dados obtidas pelo microarranjo, que são formadas por milhares de atributos.

As bases de dados usadas neste trabalho, apresentadas na Tabela 1, são formadas por instâncias com expressões gênicas de pessoas saudáveis e de portadoras de doenças relacionadas ao câncer. Algumas dessas bases foram usadas em [Borges 2006] e estão disponíveis no formato "arff", juntamente com suas documentações descrevendo as instâncias e revelando a qual classe cada uma pertence, na página da Kent Ridge[1].

Tabela 1: Descrição das bases

Nome	Número de atributos	Número de instâncias	Número de grupos*
DLBCL-Stanford	4026	47	2
DLBCL-Tumor	7129	77	2
DLBCL-NIH	7399	80	2
Leukemia-ALL/AML	7129	58	2
Leukemia-MLL	12582	57	3

* Conhecimento *a priori*

Essas bases foram submetidas ao processo de seleção de atributos por dois métodos, C-FOCUS e o *Relief*-F, este último reduzindo cada base a 10%, 25%, 50% e 75% da quantidade de seus atributos originais. Ao final desse processo existirão, para cada base original, cinco novas bases reduzidas.

Cada base resultada pela seleção de atributos, assim como as bases originais, gera uma solução composta por dois critérios: o número de atributos e o valor do índice Jaccard. Esse índice foi calculado pela comparação entre os grupos construídos pelo método K-*means* e o agrupamento real, já que há conhecimento, *a priori*, da composição das bases de expressão gênica. O conhecimento prévio também foi usado para determinar a quantidade de grupos que deviam ser gerados pelo K-*means*.

O uso de vários critérios caracteriza um problema multiobjetivo. Este trabalho faz uso do funcionamento evolucionário do método MOPSO baseado na frente de Pareto para movimentar as soluções, ou partículas, seguindo seus guias, escolhidos pelo método sigma usando dois critérios, para regiões promissoras. Os líderes eleitos pela regra de dominância de Pareto ao final da execução, quando não houver alteração no

[1] http://datam.i2r.a-star.edu.sg/datasets/krbd/, acessado em 07 de Abril de 2011.

repositório, serão considerados soluções ótimas, revelando os melhores métodos de seleção de atributos para esses critérios.

Os objetivos utilizados neste trabalho visam a minimização da quantidade de atributos e a maximização do índice Jaccard. As regiões promissoras, por sua vez, são compostas pelas partículas que otimizam esses objetivos.

No cálculo da velocidade de cada partícula, $r_1, r_2 \in [0,1]$ foram escolhidos aleatoriamente a cada iteração. O valor da inércia foi fixada em 1 para facilitar a exploração global [Sierra e Coello 2006]. As constantes de aprendizagem também foram fixadas e o valor 2 foi escolhido a partir de experiência adquirida por [Eberhart e Shi 2001].

Por se tratar de poucas soluções, todas as partículas fazem parte da mesma vizinhança, formando um grafo completamente conectado. Segundo [Sierra e Coello 2006], essa topologia converge para o resultado final mais rapidamente.

Todos os experimentos aqui foram realizados em uma máquina Intel ® Core™ I7 com 1.73GHz e 6GB de memória RAM. O MOPSO baseado na frente de Pareto foi desenvolvido na linguagem Java, assim como os métodos de seleção de atributos, o K-*means* e o índice Jaccard, que utilizaram a biblioteca do *software* Weka[2].

6. Experimentos e Resultados

Os experimentos aqui realizados visam cumprir a meta já citada na proposta. Além das bases originais, são avaliadas as bases reduzidas pelos métodos de seleção C-FOCUS e *Relief*-F, totalizando um conjunto de seis bases para cada característica.

Cada partícula gerada pelas bases de dados caracteriza a posição no espaço de buscas e o conjunto dessas partículas forma o enxame para uma determinada característica. Os enxames estão apresentados nas tabelas a seguir contendo a identificação do método de seleção, o número de atributos selecionados por esse método e o valor do índice Jaccard.

As partículas são submetidas a otimização pelo MOPSO baseado na frente de Pareto. Os gráficos abaixo, cujos eixos representam os objetivos, mostram o posicionamento final das partículas, representadas pelos círculos brancos, e dos líderes eleitos pela regra de dominância de Pareto, representados pelos círculos pretos.

A identificação dos métodos de seleção usados para cada solução obtida está numerada, tanto nas tabelas quanto nas figuras, da seguinte forma:

1- Solução obtida pelo método C-FOCUS;

2- Solução obtida pelo método *Relief*-F com 10% do tamanho original;

3- Solução obtida pelo método *Relief*-F com 25% do tamanho original;

4- Solução obtida pelo método *Relief*-F com 50% do tamanho original;

5- Solução obtida pelo método *Relief*-F com 75% do tamanho original;

6- Solução obtida pela base original;

[2] http://www.cs.waikato.ac.nz/ml/weka/, acessado em 07 de Abril de 2011.

Na Tabela 2 estão presentes as soluções geradas pelas bases DLBCL-Stanford. Como previsto, o C-FOCUS selecionou o menor conjunto de atributos. Observam-se baixos valores para o índice Jaccard, em que a solução 2 se mostrou melhor.

Tabela 2: Soluções das bases DLBCL-Stanford

Número do método	Número de atributos	Índice Jaccard
1	3	0.33
2	402	0.56
3	1006	0.39
4	2013	0.33
5	3020	0.33
6	4027	0.34

Ao final da execução do MOPSO, o repositório dos líderes eleitos pela regra de dominância de Pareto retornou as soluções 1 e 2, justamente as soluções que obtiveram, respectivamente, a menor quantidade de atributos e o maior índice Jaccard (Figura 1).

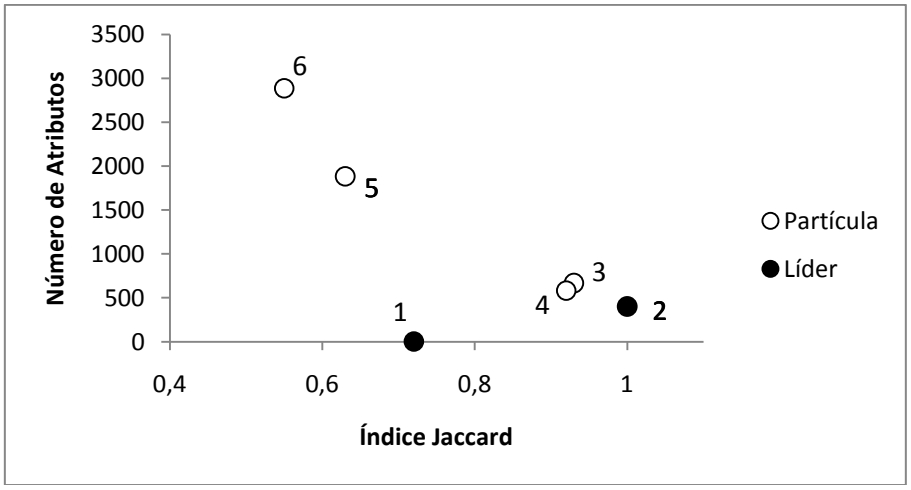

Figura 1. Posicionamento final das soluções das bases DLBCL-Stanford

As soluções geradas pelas bases DLBCL-Tumor são apresentadas na Tabela 3, onde pode ser visto que a solução 1 obteve a menor quantidade de atributos e a solução 5 conseguiu o maior valor Jaccard.

Tabela 3: Soluções das bases DLBCL- Tumor

Número do método	Número de atributos	Índice Jaccard
1	4	0.37
2	713	0.39
3	1782	0.4
4	3565	0.4
5	5347	0.46
6	7130	0.44

Na Figura 2, o repositório de líderes retornado pelo MOPSO para as bases DLBCL-Tumor é formado somente pela solução 1, que otimizou os dois critérios simultaneamente e dominou todas as outras soluções.

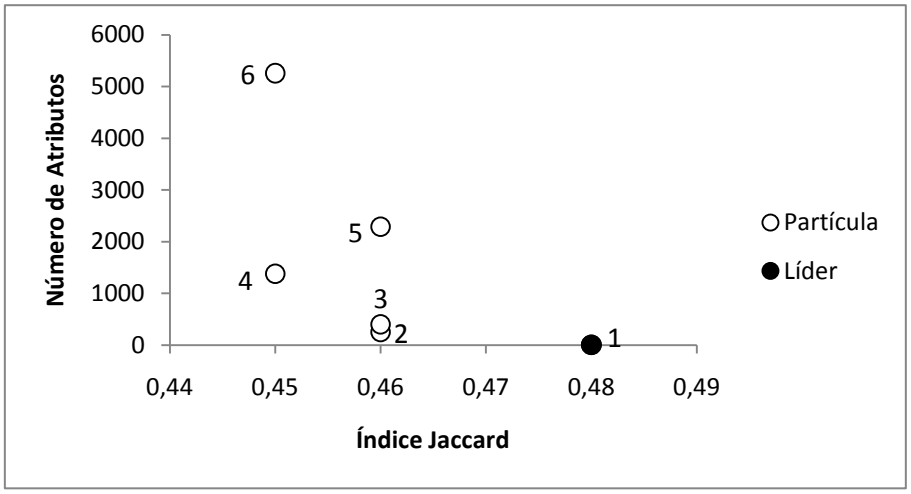

Figura 2. Posicionamento final das soluções das bases DLBCL-Tumor

Na Tabela 4, entre as soluções geradas pelas bases DLBCL-NIH, pode ser visto que a solução 1 otimizava todos os critérios, sendo o único líder do enxame. Depois de processados pelo MOPSO, a solução 2 também foi deslocada para a região promissora, tornando-se líder (Figura 3). Diferente dos demais experimentos até aqui vistos, a base de dados original dominou as solução 4 e 5, não sendo a pior solução.

Tabela 4: Soluções das bases DLBCL- NIH

Número do método	Número de atributos	Índice Jaccard
1	5	0.37
2	740	0.34
3	1850	0.353
4	3700	0.359
5	5550	0.36
6	7400	0.36

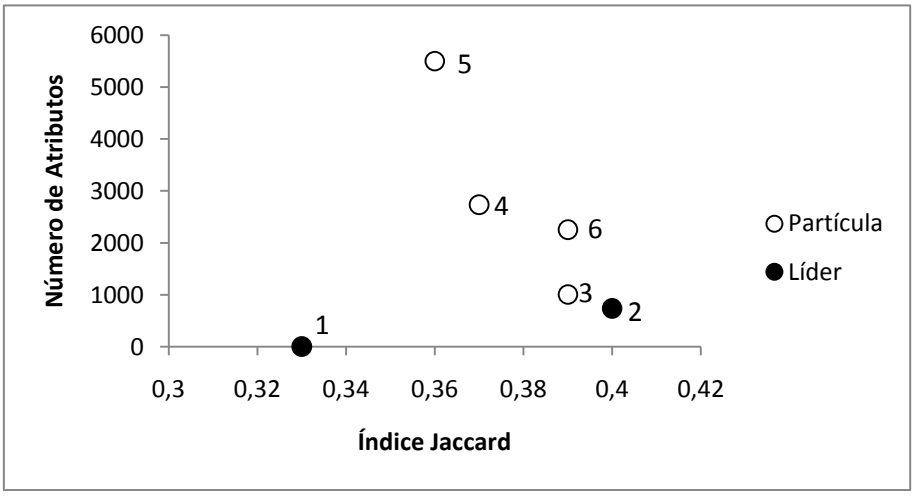

Figura 3. Posicionamento final das soluções das bases DLBCL-NIH

Os resultados das bases relacionadas a Leukemia-ALL/AML estão na Tabela 5. Apesar da solução 1 ter apenas um atributo em sua base, o valor obtido pelo índice Jaccard foi o pior entre as bases dessa característica. O melhor valor desse índice, ainda distante do valor máximo, foi obtida pela solução 5. Esse fato mostra que o agrupamento realizado foi ruim.

Tabela 5: Soluções das bases Leukemia-ALL/AML

Número do método	Número de atributos	Índice Jaccard
1	1	0.32
2	713	0.44
3	1782	0.44
4	3565	0.45
5	5347	0.47
6	7130	0.45

Na Figura 4 estão os posicionamentos das partículas após a execução do MOPSO. A solução 2, apesar de ter obtido o mesmo valor Jaccard que a partícula 3, domina-a, além das partículas 4,5 e 6, por ser composta por menos atributos. A outra solução tida como ótima é a de número 1, que se deslocou para a região de menor quantidade de atributos.

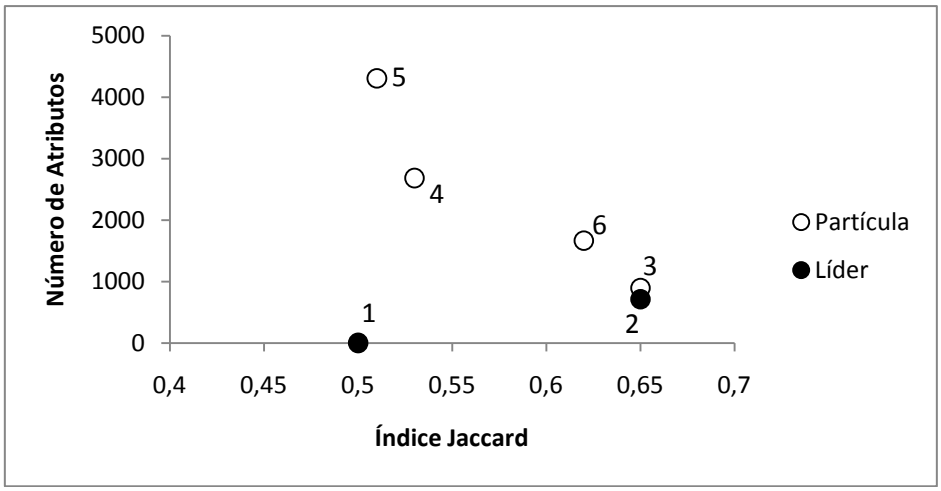

Figura 4. Posicionamento final das soluções das bases Leukemia-ALL/AML

No experimento realizado pelas bases Leukemia-MLL, a partícula 1 da Tabela 6 obteve os menores valores para o número de atributos e para o índice Jaccard. As soluções que obtiveram o melhor valor Jaccard foram as de números 5 e 6, esse último representando a base de dados original. Entretanto, como pode ser visto na Figura 5, as soluções que o MOPSO migrou para a região promissora foram as de números 1,2 e 3.

Tabela 6: Soluções das bases Leukemia-MLL

Número do método	Número de atributos	Índice Jaccard
1	3	0.21
2	1258	0.34
3	3145	0.36

4	6291	0.37
5	9437	0.53
6	12583	0.53

Nesse caso, mostra-se mais importante a necessidade de um *decision maker* para fazer a escolha pela melhor solução.

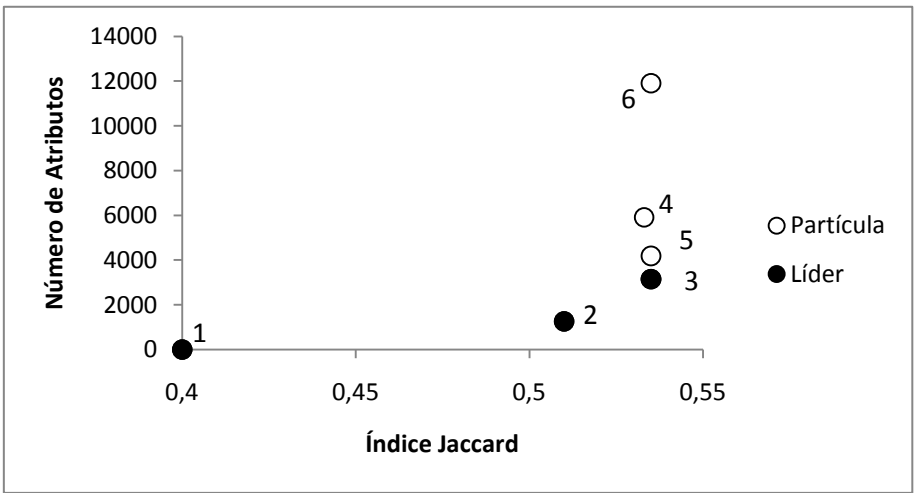

Figura 5. Posicionamento final das soluções das bases Leukemia-MLL

7. Conclusões

Este trabalho avaliou, por meio do MOPSO, métodos de seleção de atributos em bases de expressões gênicas obtidas pelo microarranjo. A primeira observação é que a redução de dimensionalidade de uma base melhora os resultados, já que nenhuma solução obtida pela base original foi classificada como ótima pelos experimentos realizados. Além disso, pelo fato de selecionar o menor subconjunto de atributos relevantes, o C-FOCUS sempre foi incluído como solução ótima, mostrando que poucos são os genes que realmente fazem parte do processo de ativação ou inibição de uma característica.

O método MOPSO baseado na frente de Pareto teve um desempenho satisfatório por ter selecionado coerentemente as soluções. Seu funcionamento evolucionário movimentou as soluções geradas pelas menores bases para as regiões mais promissoras, destacando-as das demais soluções do enxame.

Pelo fato de terem sido avaliadas soluções com poucos critérios, é possível analisá-las individualmente e comprovar a qualidade do balanceamento multiobjetivo, já que nos experimentos das Tabelas 3, 5 e 6, as soluções com o pior valor Jaccard foram selecionadas como líderes ao final do experimento por terem quantidade de atributos muito inferior das demais soluções e assim, migraram para a região promissora.

A regra de dominância de Pareto resultou vários líderes. Esse fato mostra a importância de um *decision maker* para decidir qual das soluções é a melhor. Um usuário que tenha conhecimento profundo das bases utilizadas poderá também validar o conjunto de atributos que foram selecionados, se eles realmente são os mais relevantes para cada característica abordada. Na falta do *decision maker*, o uso da regra de dominância de Pareto pode se tornar indevida pelo fato de gerar vários líderes. Nesse

caso, é mais interessante utilizar outro método multiobjetivo, que retorna apenas uma solução ótima.

A utilização do índice Jaccard mostrou que a etapa do agrupamento pode servir como avaliador dos métodos de seleção de atributos. Porém, o maior valor obtido, na solução 2 da Tabela 2, cujo valor é 0.56, mostra que o K-*means* não é a melhor escolha para agrupar dados de expressões gênicas. Além disso, caso não haja um conhecimento *a priori* das bases de dados, o K-*means* pode ser substituído pelo ISODATA, que tem o mesmo funcionamento e realiza um auto-balanceamento da quantidade de grupos [Xu e Wunsch 2005]. O mesmo ocorre com o índice Jaccard, que pode ser substituído por outro índice que não necessite de informações além das contidas na base de dados.

O objetivo principal deste trabalho foi mostrar que um método de otimização multiobjetivo pode avaliar métodos de seleção de atributos, escolher os melhores para uma característica específica, afim de tornar a análise de bases de dados de expressões gênicas obtidas pela técnica de microarranjo mais fácil e rápida.

Como proposta de trabalhos futuros para obtenção de resultados melhores e mais precisos, será estudado o método lexicográfico de otimização multiobjetivo, que estabelece uma sequência de preferência de critérios para escolher a melhor e única solução. Um método de agrupamento baseado em densidade, o DBSCAN, poderá ser usado para comparar os resultados obtidos pelo K-*means* neste trabalho.

Agradecimentos

Os autores deste trabalho gostariam de agradecer ao CNPq pelo suporte financeiro (projeto Seleção de Atributos Usando Critérios Multiobjetivo Baseados em Enxames para a Seleção de Genes em Microarranjos, referência 555264/2009-2).

Referências Bibliográficas

Azofra, A.A., Sanchez, J.M.B e Peña J.L.C. (2003) "C-FOCUS: A continuous Extension of FOCUS", Advances in Soft Computing - Engineering, Design and Manufacturing, p. 225-232.

Berkhin, P. (2002) "Survey of clustering data mining techniques", Technical report, Accrue Software, EUA.

Borges, H.B. (2006) "Redução de Dimensionalidade em Bases de Dados de Expressão Gênica", Dissertação de Mestrado, PPGIa-PUCPR.

Carvalho, A.B. e Pozo, A.T.R. (2009) "Otimização por Nuvem de Partículas Multiobjetivo na Aprendizagem Indutiva de Regras: Extensões e Aplicações", Dissertação de mestrado, Universidade Federal do Paraná.

Chuang, L.Y., Chang, H.W., Tu, C.J. e Yang, C.H. (2008), Computational Biology and Chemistry, Elsevier, vol. 32, p. 29-38.

Coello, C.A., Lamont, G.B. e Veldhuizen, D.A.V. (2007), Evolutionary Algorithms for Solving Multi-Objective Problems, segunda edição, Springer-Verlag.

Dash, M., Choi, K., Scheuermann, P. e Liu, H. (2002) "Feature Selection for Clustering – A Filter Solution", Em: 2nd International Conference on Data Mining, p. 115-122.

Dy, J. (2008) "Unsupervised Feature Selection", Computational Methods of Feature Selection, Editado por Huan Liu e Hiroshi Motoda, Chapman & Hall/CRC, p. 19-39.

D'Haeseleer, P. (2005) "How Does Gene Expression Clustering Work?", Nature Biotechnology, vol. 23, no. 12, p. 1499-1501.

Eberhart, R.C. e Shi, Y. (2001) "Particle Swarm Optimization: Developments, Applications and resources", Em: IEEE International Conference Evolutionary Computation, vol. 1, p. 81-86.

Faceli, K., Caravalho, A. e Souto, M. (2005) "Validação de Algoritmos de Agrupamento", Relatórios Técnicos do ICMC.

Handl, J. e Knowles, J. (2007) "An Evolutionary Approach to Multiobjective Clustering", Em: IEEE/ACM Transactions on Evolutionary Computation, vol. 11, no. 1, p. 56-76.

Kennedy, J. e Eberhart, R.C. (1995) "Particle Swarm Optimization", Em: IEEE International Conference on Neural Networks, p. 1942-1948, IEEE Press.

Kohavi, R. e John, G.H. (1997) "Wrappers for Feature Subset Selection", Artificial Intelligence, p. 273-324.

Kononenko, I. e Sikonja, M.R. (2008) "Non-Myopic Feature Quality Evaluation with (R)ReliefF", Computational Methods of Feature Selection, Editado por Huan Liu e Hiroshi Motoda, Chapman & Hall/CRC, p. 169-191.

Mostaghim, S. e Teich, J. (2003) "Strategies for finding good local guides in multi-objective particle swarm optimization (MOPSO)", Em: IEEE Swarm Intelligence Symposium, p. 26-33.

NIH- National Institutes of Health (2001) "Genetic Basic", http://www.nigms.nih.gov, acessado em 07 de Abril de 2011.

Sierra, M.R. e Coello, C.A.C. (2006) "Multi-objective Particle Swarm Optimizers: A Survey of The State-of-the-art", Em: International Journal of computational Intelligence Research, vol. 2.

Suresh, K., Kundu, D., Ghosh, S. e Das S. (2009) "Data Clustering Using Multi-objective Differential Evolution Algorithm", Fundamenta Informaticae, vol. 21, IOS Press, p. 1001-1024.

Xu, R. e Wunsch, D. (2005) "Survey of Clustering Algorithms", Em: IEEE Transactions on Neural Networks, vol. 16, no. 3.

Yu, L. e Liu, H. (2004) "Redundancy Based Feature Selection for Microarray Data", Em: 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.