Artigo submetido em 24/1/12; aceito para publicação em 11/4/12

Aplicação de Métodos Computacionais de Mineração de Dados na Classificação e Seleção de Oncogenes Medidos por *Microarray*

Oncogenes Classification Measured by Microarray using Data Mining Computational Methods

Aplicación de Métodos Computacionales de Minería de Datos en la Clasificación y Selección de Oncogenes Medidos por *Microarray*

Fabrício Alves Rodrigues[1]; Laurence Rodrigues do Amaral[2]

Resumo

Introdução: Nas últimas décadas, o câncer ganhou uma dimensão maior, convertendo-se em um evidente problema de saúde pública mundial. A Organização Mundial da Saúde estimou que, no ano 2030, podem-se esperar 27 milhões de casos incidentes de câncer e 17 milhões de mortes por câncer. Frente a esse cenário alarmante, a mineração de dados traz métodos e ferramentas capazes de auxiliar na construção de conhecimentos mais incisivos sobre o câncer. **Objetivo:** Este trabalho tem por objetivo aplicar cinco métodos tradicionais da mineração de dados à base de dados NCI60, construída com dados oriundos de experimentos de *microarray*, com níveis de expressão de 1.000 genes agrupados em nove classes de câncer. **Método:** Foram utilizados neste trabalho os métodos J48, Random Forest, PART, IBK e Naive Bayes, pertencentes ao ambiente Weka, bem tradicionais na mineração de dados. Devido ao baixo número de registros para determinadas classes, utilizou-se, na validação dos resultados obtidos pelos classificadores, o 3-*fold cross validation*. **Resultados:** O classificador que obteve a melhor precisão foi o IBK, enquanto os classificadores J48 e PART conseguiram diminuir o conjunto de genes drasticamente, construindo conhecimento de alto nível na forma de árvores ou regras. **Conclusão:** Os resultados obtidos neste trabalho podem ser utilizados como ferramentas que visam a auxiliar no enfrentamento do câncer, podendo ser utilizadas na classificação de novos casos ou para se conhecer, cada vez mais, as relações gene/gene e gene/câncer.

Palavras-chave: Biologia Computacional; Expressão Gênica; Mineração de Dados; Oncologia; Bases de Dados como Assunto

[1] Bacharel em Ciência da Computação. Campus Jataí. Universidade Federal de Goiás. Jataí (GO), Brasil. *E-mail:* fabrício1989@gmail.com.
[2] Professor de 3º grau. Mestre em Ciência da Computação. Faculdade de Computação (FACOM). Universidade Federal de Uberlândia (UFU). Patos de Minas (MG), Brasil. *E-mail:* lramaral@yahoo.com.br.
Endereço para correspondência: Laurence Rodrigues do Amaral. FACOM/UFU. Campus Patos de Minas, sala 207. Av. Getúlio Vargas, 230 - Centro. Patos de Minas (MG), Brasil. CEP: 38700-103.

INTRODUÇÃO

Nas últimas décadas, o câncer ganhou uma dimensão maior, convertendo-se em um evidente problema de saúde pública mundial. A Organização Mundial da Saúde (OMS) estimou que, no ano 2030, podem-se esperar 27 milhões de casos incidentes de câncer, 17 milhões de mortes por câncer e 75 milhões de pessoas vivas, anualmente, com câncer[1].

No Brasil, as estimativas para o ano de 2012, que também são válidas para o ano de 2013, apontam a ocorrência de aproximadamente 518.510 novos casos de câncer, incluindo os casos de pele não melanoma, reforçando a magnitude do problema do câncer no país[1].

A prevenção e o controle do câncer precisam adquirir o mesmo foco e a mesma atenção que a área de serviços assistenciais, pois, quando o número de novos casos aumentar de forma rápida, não haverá recursos suficientes para dar conta das necessidades de diagnóstico, tratamento e acompanhamento. Então mais e mais pessoas terão câncer e correrão o risco de morrer prematuramente por causa da doença. As consequências poderão ser devastadoras nos aspectos social e econômico. O câncer pode se tornar um grande obstáculo para o desenvolvimento socioeconômico de países emergentes como o Brasil[1].

Frente a esse cenário alarmante, a mineração de dados (MD) traz métodos e ferramentas capazes de auxiliar na construção de conhecimentos mais incisivos sobre o câncer, que podem, seguramente, complementar os conhecimentos já existentes sobre esse mal.

A MD, ou também chamada de *data mining*, é o processo de descoberta automática de informações úteis em grandes depósitos de dados. As técnicas de MD são organizadas para agir sobre grandes aglomerados de dados com o intuito de descobrir padrões ou regras úteis que poderiam, de alguma forma, permanecer ignorados[2]. Em outras palavras, a MD refere-se às atividades que analisam os dados, descobrem problemas e oportunidades ocultos em seus relacionamentos, constroem modelos computacionais com base nessas descobertas e, então, utilizam esses modelos para prever determinado comportamento, exigindo a mínima intervenção do usuário final[3]. Os processos de MD focam na aplicação de técnicas estatísticas e de inteligência artificial para a análise interativa dos dados, visando à identificação de padrões de comportamento, tendências ou predição[4].

Ao aplicar métodos computacionais, como os oriundos da MD, a dados biológicos ou médicos, constroem-se aplicações na área de conhecimento conhecida como Bioinformática ou Biologia Computacional. Ela abrange a aquisição, o processamento, o armazenamento, a distribuição, a análise e a interpretação da informação biológica ou médica. Por meio da combinação de procedimentos e técnicas advindos da Matemática, Estatística e Ciência da Computação, são elaborados métodos e ferramentas que auxiliam na compreensão do significado biológico representado nos dados genômicos[5].

Essas aplicações bioinformáticas estão ganhando espaço no meio biológico e médico, principalmente na manipulação e análise de dados oriundos de experimentos de *microarray*. Em alguns trabalhos, conforme citado a seguir, foram aplicados métodos computacionais, tais como: redes neurais artificiais, regras *fuzzy, Support Vector Machines* (SVM), entre outros, na classificação de amostras oriundas de *microarray*.

Foi apresentado por Schaefer et al.[6] um classificador baseado em regras *fuzzy* o qual foi aplicado com sucesso em dados de expressão gênica. O classificador é formado por um conjunto de regras do tipo IF-THEN e foi aplicado em quatro *datasets* relacionados aos cânceres de cólon e ovário, além de leucemia e linfoma, obtendo taxas de classificação médias de 81,85%.

Foi proposto por Wutao Chen et al.[7] um classificador utilizando várias redes neurais artificiais. O classificador proposto foi aplicado em *datasets* contendo níveis de expressão gênica oriundos de experimentos de *microarray* de leucemia. Os resultados obtidos foram animadores, mostrando que, devido à complexidade dos dados oriundos de *microarray*, uma simples rede neural artificial não consegue prover boas taxas de classificação. O conjunto de redes neurais obteve taxas de classificação médias de 93,27%.

Uma característica marcante dos dados oriundos de experimentos de *microarray* é a sua alta dimensionalidade. Dessa forma, este trabalho utiliza um método chamado *locality preserving projections* (LPP), a fim de reduzir a dimensionalidade desses *datasets* e prover a seleção de características que sejam realmente importantes na tarefa de classificação. Esse método foi aplicado em sete *datasets*, são eles: ALL-AML-3, ALL-AML-4, GCM, TR41, PROSTATE, DLBCL e BRAIN1, obtendo taxas de classificação médias de 90,98% para todos os sete *datasets*[8].

Foi proposto por Vecchiola et al.[9] um novo ambiente de aprendizado de máquina que utiliza aprendizado coevolutivo, baseado em *feature set partitioning* na classificação de dados oriundos de expressão gênica. Foram utilizados dois *datasets*, BRCA (*breast cancer gene profiles*) e o *Prostate*, obtendo taxas de classificação médias de 98% e 70%, respectivamente.

No trabalho proposto por Su Liangliang et al.[10], foi apresentado um classificador que utiliza matrizes de adjacências espectrais na obtenção de autovetores que são utilizados como parâmetros de entrada em SVM e também para o K-NN (*k-nearest neighbor*). Esse ambiente híbrido,

formado pela SVM e pelo KNN, foi aplicado em *datasets* com dados relacionados a câncer de cólon e próstata, e obteve taxas de classificação médias de 94,35% e 89,7%, respectivamente.

Foi apresentado por Ghorai et al.[11] um classificador chamado *nonparallel plane proximal classifier* (NPPC), que foi aplicado na classificação de dados oriundos de experimentos de *microarray*. O método foi aplicado em sete *datasets* relacionados ao câncer. São eles: leucemia (ALL-AML), câncer de cólon, língua, mama, fígado e próstata, além de leucemia, obtendo taxas de classificação médias de 89,81%.

Seguindo a mesma direção desses trabalhos, este trabalho tem por objetivo aplicar cinco métodos tradicionais da MD à base de dados NCI60 construída com dados oriundos de experimentos de *microarray*, com níveis de expressão de 1.000 genes agrupados em nove classes de câncer.

MÉTODO

Serão apresentados aqui informações sobre os *microarrays*; descrição da base de dados NCI60; e os métodos computacionais da MD aplicados neste trabalho.

O *microarray* de DNA é uma metodologia utilizada para comparar a expressão de um grande número de genes simultaneamente. Essa técnica emprega arranjos (*arrays*) que contêm uma grande quantidade de genes distribuídos por um braço robótico de forma ordenada sobre placas de vidro. As sondas podem ser conjuntos de DNA complementar de fita simples (cDNAs) gerados a partir de células ou tecido em duas situações diferentes. Os resultados são produzidos sob a forma de diferentes intensidades de fluorescência que são captadas por microscopia a laser em função dos diferentes níveis de expressão de cada gene[12].

A imagem dos pontos fluorescentes é processada por meio de métodos computacionais com o objetivo de calcular a intensidade obtida para cada *spot*. A tecnologia de *microarrays* não fornece apenas informações sobre a função de genes anônimos, mas também constitui uma ferramenta indispensável para estudos globais de expressão gênica[12].

O *microarray* é uma tecnologia que nos fornece a possibilidade de criar conjuntos de dados de informação molecular para representar muitos sistemas de interesse biológico ou clínico. Esses perfis de expressão gênica são usados como conteúdo em grande escala para que possam ser analisados[13]. Um exemplo desses perfis gênicos é a base NCI60[14]. Essa base de dados faz parte do NCI60 *Cancer Microarray Project*, advindo da colaboração entre o laboratório Brown/Bolstein, do grupo John Weinstiens do *Laboratory of Molecular Pharmacology e do Laboratory of Developmental Therapeutics*, ambos pertencentes ao *National Cancer Institute*, nos Estados Unidos da América (EUA).

Para a construção dessa base, foram utilizados *microarrays* de cDNA para buscar expressões gênicas de aproximadamente 8.000 genes distintos. Esses genes são oriundos de 61 linhagens celulares e foram classificados em nove classes de câncer: (C_1) mama, (C_2) sistema nervoso central, (C_3) cólon, (C_4) leucemia, (C_5) melanoma, (C_6) pulmão, (C_7) ovário, (C_8) renal e (C_9) células reprodutivas. Os índices C_n, onde *n* varia de [1..9] referem-se ao código utilizado para representar cada classe na base de dados. O número de ocorrências de cada classe é dado a seguir: mama (7), sistema nervoso central (6), cólon (7), leucemia (6), melanoma (8), pulmão (9), ovário (6), renal (8) e células reprodutivas (4), totalizando 61 amostras[15].

No trabalho de Ooi e Tan[16] foi realizado um pré-processamento, no qual foram excluídos genes que estavam em *spots* inválidos, de controle e vazios, totalizando 6.176 genes. Finalmente, partindo dos 6.176 genes pré-processados, Ooi e Tan chegaram a um *dataset* reduzido contendo 1.000 genes, os quais apresentaram os maiores valores de desvio-padrão na base NCI60. Esses genes foram indexados de 1 a 1.000.

A Tabela 1 apresenta uma visão geral da base NCI60, composta pela expressão de 1.000 genes (colunas), medida para 61 amostras de células (linhas), sendo que cada amostra é classificada em uma das nove classes de câncer citadas anteriormente (última coluna). Os dados de expressão gênica são valores do tipo ponto flutuante que podem assumir valores negativos ou positivos, sendo obtidos através das intensidades dos pontos fluorescentes obtidos no *microarray*.

A base de dados NCI60 possui uma característica que a faz desafiadora para os métodos tradicionais da MD. Essa característica está relacionada à sua alta dimensionalidade; isto é, ao elevado número de atributos (1.000) que a mesma possui. Além disso, ela possui um número muito baixo de registros (61). Segundo Xu et al.[17], é muito difícil propor regras ou critérios na determinação de um conjunto de genes que sejam discriminantes no diagnóstico de doenças, especialmente quando as bases de dados estudadas possuem um elevado número de classes, como a base NCI60.

Os métodos computacionais utilizados neste trabalho fazem parte do ambiente Weka que permite manipular conjuntos de dados de diversos domínios desde que estejam no formato *.ARFF. Esse ambiente possui vários métodos computacionais da MD e foi desenvolvido na *University of Waikato* em Waikato na Nova Zelândia.

Foram utilizados neste trabalho os métodos J48, Random Forest, PART, IBK e Naive Bayes, bem tradicionais na MD.

Tabela 1. Visão geral da base NCI60 reduzida utilizada nos experimentos de Ooi e Tan[16]

Amostra	Expressão Gene 1	Expressão Gene 2	...	Expressão Gene 1000	Classificação
1			...		
2			...		
3			...		
...	...	...	...	...	...
60			...		
61			...		

O classificador J48 é uma implementação escrita em Java do algoritmo C4.5, presente na ferramenta de MD Weka. O C4.5 é um algoritmo utilizado na geração de árvores de decisão e foi proposto por J. R. Quinlan, no livro "C4.5: *Programs for Machine Learning*" em 1993. Esse algoritmo é uma extensão do algoritmo ID3 e suas árvores de decisão geradas são utilizadas em tarefas de classificação.

O método Random Forest é um classificador do tipo comitê ou "ensemble" e é constituído de várias árvores de decisão. Cada uma dessas árvores de decisão dá um voto que indica sua decisão sobre a classe à qual pertencerá um determinado objeto. O objeto então pertencerá à classe que obtiver o maior número de votos. O algoritmo de Random Forest foi desenvolvido por L. Breiman em 2001 e foi publicado no periódico *Machine Learning*.

O algoritmo PART produz um conjunto de regras do tipo IF-THEN a partir de uma árvore de decisão construída através do J48 (algoritmo C4.5). Ele foi proposto por Eibe Frank e Ian H. Witten em 1998 no trabalho: "Generating Accurate Rule Sets Without Global Optimization".

O algoritmo IBK é uma versão do algoritmo de clusterização k-NN (*k-nearest neighbor*) utilizado em tarefas de clusterização. Esse método representa cada instância como um ponto de dado em um espaço *d*-dimensional, onde d é o número de atributos. Dada uma nova instância, calcula-se a sua proximidade com o resto dos pontos de dados no conjunto de treinamento, usando medidas de proximidade, tais como a distância euclidiana. Os *k* vizinhos mais próximos de uma instância z são classificados como sendo da mesma classe de *z*.

Os classificadores Bayesianos são fundamentados no teorema de Bayes e utilizam fundamentos de probabilidade condicional na tarefa de classificação. Dessa forma, eles calculam a probabilidade de uma instância pertencer a cada uma das classes pré-determinadas, pertencendo essa instância à classe que obteve a maior probabilidade após a aplicação do teorema de Bayes.

Devido ao baixo número de registros para determinadas classes, foi utilizado, na validação dos resultados obtidos pelos classificadores, o 3-*fold cross validation*. Essa validação divide a base de dados em três partes, em que duas partes serão utilizadas para treinamento e uma para teste. A precisão final do classificador é dada pela média das avaliações obtidas em cada uma das composições.

Para cada um dos métodos pertencentes ao Weka, a saída é dada na forma de uma matriz de confusão e, a partir dessa matriz, são calculados os valores de sensibilidade e especificidade. Esses valores são frequentemente utilizados em domínios médicos e foram extraídos do trabalho proposto por Lopes et al.[18]. As equações 1 e 2 ilustram como são calculados esses dois valores: (1) Sensibilidade = VerdadeiroPositivo ÷ (VerdadeiroPositivo + FalsoNegativo); (2) Especificidade = VerdadeiroNegativo ÷ (VerdadeiroNegativo + FalsoPositivo). De posse desses dois valores, a avaliação final de um classificador é dada pela sua multiplicação. A Equação 3 ilustra esse cálculo: (3) Score = Sensibilidade * Especificidade.

RESULTADOS E DISCUSSÃO

A Tabela 2 ilustra as precisões de classificação obtidas para os métodos J48, Random Forest, PART, IBK e Naive Bayes, analisando cada uma das nove classes de câncer separadamente.

Os valores mostrados na Tabela 2 foram calculados utilizando as equações 1, 2 e 3, e correspondem ao score obtido por cada método. O *score* é calculado através da multiplicação dos valores obtidos de sensibilidade e especificidade. Quanto maior o valor de sensibilidade e especificidade, maior o *score*. É importante salientar que para um determinado método obter um alto valor de *score*, é necessário que os valores encontrados para sensibilidade e especificidade sejam altos. Se apenas um deles for baixo, obrigatoriamente o score também será baixo.

Analisando classe a classe, pode-se perceber que o classificador que melhor classificou as amostras foi o IBK,

Tabela 2. Valores de *score* obtidos pelos cinco métodos analisados em cada uma das nove classes de câncer analisadas, calculados utilizando a equação 3

Métodos	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	Média método
J48	0,59	0,64	0,85	0,77	0,67	0,49	0,38	0,45	0,62	0,60
Random Forest	0,51	0,71	0,78	0,82	0,82	0,64	0,4	0,5	0,62	0,64
PART	0,55	0,74	0,83	0,79	0,70	0,45	0,48	0,35	0,49	0,59
IBK	0,63	0,62	0,98	0,82	0,94	0,54	0,37	0,56	0,72	0,68
Naive Bayes	0,59	0,5	0,79	0,74	0,81	0,49	0,5	0,75	0,62	0,64
Média simples	**0,574**	**0,642**	**0,846**	**0,788**	**0,788**	**0,522**	**0,426**	**0,522**	**0,614**	
Desvio-padrão	**0,045**	**0,093**	**0,080**	**0,034**	**0,107**	**0,073**	**0,059**	**0,148**	**0,035**	

obtendo os melhores resultados nas classes C_1 (mama), C_3 (cólon), C_4 (leucemia), C_5 (melanoma) e C_9 (células reprodutivas). Além disso, foi o segundo melhor para as classes C_6 (pulmão) e C_8 (renal). Isto é, não ficou em primeiro ou segundo lugar somente para as classes C_2 (sistema nervoso central) e C_7 (ovário).

Ao analisar a precisão média obtida nas nove classes, o resultado obtido na análise anterior se repete, tendo como melhor método de classificação o IBK, seguido pelo Random Forest e pelo Naive Bayes.

Ainda com relação à Tabela 2, percebe-se que algumas classes de câncer são mais facilmente classificadas, obtendo, assim, melhores valores de sensibilidade e especificidade, e consequentemente maiores valores de *score*. Analisando a média obtida por todos os classificadores para cada uma das classes, percebe-se que a classe C_3 (cólon) é a mais fácil de ser classificada, com *score* de 0,846 ou 84,6% de classificações corretas. As segundas melhores foram as classes C_4 (leucemia) e C_5 (melanoma) com *score* de 0,788 ou 78,8%. O pior valor de *score* foi obtido para a classe C_7 (ovário), obtendo valor médio de 0,426 ou 42,6%.

Outro ponto a ser elucidado refere-se aos baixos valores de desvio-padrão encontrados por classe, mostrando baixas discrepâncias entre os valores obtidos para os métodos analisados para cada uma das classes analisadas.

O conhecimento gerado por um classificador é tão importante quanto a sua precisão, sendo de suma importância para o especialista, pois auxilia no entendimento de quais características são importantes para a resolução do problema em si. É importante não só escolher um classificador que tenha altos valores de classificação, mas também um que possa mostrar quais são os genes que estão relacionados com determinadas classes de câncer. Por exemplo, é tão (ou até mais) importante que um classificador informe que os genes X, Y e Z estejam relacionados ao câncer K, do que somente altas taxas de classificação.

De todos os métodos analisados nesta pesquisa, somente o J48 e o PART geram conhecimento de alto nível como saída. Todos os demais são classificadores do tipo "caixa-preta" ou *black-box*, onde é fornecida como *input* a amostra, e o classificador fornece como *output* a classe à qual aquela amostra pertence.

A Tabela 3 traz as regras geradas pelo PART, com uma regra do tipo IF-THEN para cada uma das nove classes de câncer. Já a Tabela 4 traz importantes informações sobre os 21 genes selecionados pelos classificadores J48 e PART., além de informações sobre o número de acesso desses genes no GenBank e também sua anotação. De posse do número de acesso do GenBank, várias informações podem ser obtidas com relação a um determinado gene, utilizando o portal do GenBank dentro do *National Center for Biotechnology Information* (NCBI).

A árvore de decisão gerada pelo J48 possui tamanho 25, possuindo 13 níveis e é composta por 12 genes distintos. Foram fornecidos, para o método J48, os níveis de expressão de 1.000 genes. Esse método mostrou que para classificar uma nova amostra de expressão gênica, em uma das nove classes de câncer analisadas, precisam-se de apenas 12 genes. Isto é, 988 genes ou 98,8% dos genes analisados não são importantes para a classificação em uma das nove classes de câncer.

Na construção da árvore de decisão, o J48 utiliza os atributos mais significativos, selecionando-os e utilizando conceitos de entropia, ganho de informação, entre outros. A árvore de decisão gerada correlaciona os atributos selecionados utilizando os operadores matemáticos ≤ (menor ou igual) e > (maior). Dessa forma se, por exemplo, o nível de expressão do Gene320, que possui número de acesso no GenBank igual a [5':W90268,3':W90593], for menor ou igual a -0,917228; e o nível de expressão do Gene7, que possui número de acesso no GenBank igual a [5':AA055721,3':AA055664], for menor ou igual a 0,286123, então a amostra é classificada como sendo classe C_4 (leucemia). Se o nível de expressão do Gene320 é menor ou igual a -0,917228 e o nível de expressão do Gene7 for maior do que 0,286123, então a amostra é classificada como sendo classe C_9 (células reprodutivas).

Essa análise pode ser expandida para todos os demais ramos da árvore de decisão.

Pode-se perceber que as regras geradas pelo PART são pequenas, com tamanho médio de 2,22 genes/regra. As menores regras foram geradas para as classes C7 (ovário) e C8 (renal) com apenas um gene. Isso quer dizer que, por exemplo, se ao medir o nível de expressão do gene de código 21, que possui o número de acesso ao GenBank igual a [5':AA046218, 3':AA046260], e o mesmo for menor ou igual a -2,53883, o indivíduo possui valor de sensibilidade igual a 0,773 (ou 77,3%) e valor de especificidade igual a 0,622 (ou 62,2%) para câncer de ovário, obtendo valor de *score* igual a 0,48.

Tabela 3. Regras do tipo IF-THEN geradas pelo método PART

if (Gene18 > -3.04138) and (Gene502 <= 0.418436) and (Gene47 <= -2.53169) and (Gene39 <= -3.54862) then C1 (mama)	if (Gene18 > -3.04138) and (Gene901 > 0.283578) and (Gene15 > -1.83587) then C2 (sistema nervoso central)
if (Gene747 <= -0.744382) and (Gene6 > -1.11935) then C3 (cólon)	if (Gene320 <= -0.917228) and (Gene7 <= 0.286123) then C4 (leucemia)
if (Gene18 > -3.04138) and (Gene333 <= -1.85647) then C5 (melanoma)	if (Gene316 > -0.934852) and (Gene330 > -2.03856) and (Gene242 <= -0.22917) then C6 (pulmão)
if (Gene21 <= -2.53883) then C7 (ovário)	if (Gene284 <= 0.265265) then C8 (renal)
if (Gene709 <= -0.828556) and (Gene3 <= -0.533707) then C9 (células reprodutivas)	

Tabela 4. Número de acesso do GenBank e anotações dos genes selecionados pelos métodos J48 e PART

Índice	Número de acesso no GenBank	Anotação
2	[5':AA055858,3':AA055808]	Member of the GA733 family
3	[5':AA057287,3':AA058732]	Highly similar to a region of Yes-associated protein
6	[5':W31089, 3':N98525]	Homolog of Drosophila Tid56 tumor suppressor protein
7	[5':AA055721,3':AA055664]	Inhibitor of G1-specific CDK-cyclin
15	[5':H51958,3':H52087]	Caldesmon 1; actomyosin regulatory protein
18	[5':AA047106,3':AA047243]	Caveolin 1; tumor suppressor and structural
21	[5':AA046218, 3':AA046260]	Proteoglycan 1, secretory granule
39	[5':H18563, 3':H18456]	ESTs
47	[5':AA040884,3':AA040885]	Calcyclin; interacts with target proteins
50	[5':N94809, 3':N63511]	Dihydropyrimidine dehydrogenase
59	[5':R35963, 3':R49477]	Neuromedin B receptor
242	[5':W16630,3':N78729]	Galectin 1; beta-galactoside-binding lectin
284	[5':H29665,3':H29581]	ESTs, Highly similar to STATHMIN
316	[5':H14669, 3':H14579]	KIAA1232 protein
320	[5':W90268,3':W90593]	Ubiquitin-like 4
330	[5':W38991,3':N93208]	Solute carrier family 38, member 1
333	[5':W38991, 3':N93208]	Solute carrier family 38, member 1
502	[5':AA024655, 3':AA025275]	DAPK1 Death-associated protein kinase 1
709	[5':N71211, 3':N22009]	Homolog of mouse quaking QKI
747	[5':W00805, 3':N69996]	Serine (or cysteine) proteinase inhibitor
901	[5':AA039403, 3':AA039283]	ESTs, Weakly similar to 2206426A globin245

Os resultados obtidos para essas duas classes (C_7 e C_8) apresentam uma informação muito importante, elucidando que a expressão de 999 genes (isto é, 99,9% dos genes) não é importante na determinação dos cânceres de ovário e renal. A maior regra foi obtida para a classe C_1 (mama) com quatro genes.

As regras geradas pelo PART, para as nove classes de câncer, são formadas por 18 genes distintos, tendo o gene 18, que possui número de acesso no GenBank igual a [5':AA047106,3':AA047243], aparecido em três regras, C_1 (mama), C_2 (sistema nervoso central) e C_5 (melanoma).

Informações a respeito do conjunto de genes gerado por um classificador é de suma importância para o especialista, pois auxilia no entendimento de quantos e quais genes estão ligados a determinadas classes de câncer. Quanto menor o conjunto de genes melhor é o classificador. A Tabela 5 traz um comparativo entre os métodos J48 e PART contrapondo-os com outros oito classificadores publicados na literatura, que foram aplicados à base de dados NCI60.

O método J48, juntamente com os métodos propostos por Deb e Reddy[20] e Ooi e Tan[16], obteve os melhores resultados, sendo compostos todos os três classificadores por 12 genes. O classificador PART ficou em segundo lugar, somente atrás dos três classificadores supracitados, sendo composto por 18 genes. Dessa forma, os métodos J48 e PART, além de fornecerem boas taxas de precisão, são compostos por um baixo número de genes, bem abaixo do número de genes formados pela maioria dos classificadores presentes na literatura.

O problema do câncer no Brasil ganha relevância e, com isso, tem conquistado espaço nas agendas políticas e técnicas de todas as esferas de governo. O conhecimento sobre essa doença permite estabelecer prioridades e alocar recursos de forma direcionada para a modificação positiva desse cenário na população brasileira[1].

Para o enfrentamento do câncer, são necessárias ações que incluam: educação em saúde em todos os níveis da sociedade; promoção e prevenção orientadas a indivíduos e grupos; geração de opinião pública; apoio e estímulo à formulação de leis que permitam monitorar a ocorrência de casos[1].

Para que essas ações sejam bem-sucedidas, será necessário ter como base as propostas em informações oportunas e de qualidade (consolidadas, atualizadas e representativas) e análises epidemiológicas a partir dos sistemas de informação e vigilância disponíveis[1].

Dessa forma, a MD e a Bioinformática têm papel importantíssimo no enfretamento do câncer, pois podem acelerar as pesquisas por novos medicamentos e tratamentos, além de serem importantes aliadas no diagnóstico precoce do câncer.

Os arquivos de saída gerados pelo Weka para os cinco métodos analisados podem ser baixados no link: https://sites.google.com/site/laurenceamaral/research/rbcinca.

Tabela 5. Comparativo entre os métodos J48 e PART e outros métodos publicados na literatura, levando em consideração o número de genes que compõem o classificador

Referência	Número de genes
Dudoit et al.[19]	30
Deb e Reddy[20]	12
Ooi e Tan[16]	12
Amaral[15]	20
Liu et al.[21]	40
Jirapech-Umpai e Aitken[22]	30
Xu et al.[17]	79
Sihua Peng et al.[23]	24
J48	12
PART	18
Melhor resultado	**12**
Pior resultado	**79**

CONCLUSÃO

Neste trabalho, utilizaram-se cinco métodos de MD pertencentes ao ambiente Weka na classificação da base de dados NCI60 construída com os níveis de expressão de 1.000 genes hierarquizados em nove classes de câncer. Esses métodos foram avaliados com relação à sua precisão de classificação e também ao conhecimento gerado.

Os resultados obtidos neste trabalho podem ser utilizados como ferramentas que visam a auxiliar no enfrentamento do câncer, podendo ser utilizadas na classificação de novos casos.

Através da medição dos níveis de expressão de um determinado indivíduo, pode-se fazer uma triagem mais eficiente e, consequentemente, o tratamento desse tipo de câncer poderá ser iniciado mais precocemente. Ao iniciar o tratamento precocemente, aumentam-se as chances de cura desse mal.

Além disso, ao se conhecer mais profundamente as relações gene/gene e gene/câncer, novos tratamentos e medicamentos poderão ser desenvolvidos, focando em determinada classe de câncer e podendo diminuir o enorme *gap* existente entre a descoberta de uma nova droga e seu uso nos hospitais.

CONTRIBUIÇÕES

Fabrício Alves Rodrigues contribuiu no planejamento do projeto de pesquisa, na análise e interpretação dos

dados e na redação e revisão crítica; Laurence Rodrigues do Amaral contribuiu na concepção e planejamento do projeto de pesquisa, na análise e interpretação dos dados e na redação e revisão crítica.

Declaração de Conflito de Interesses: Nada a Declarar.

REFERÊNCIAS

1. Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2012: incidência de câncer no Brasil. Rio de Janeiro: INCA; 2011. 118 p.
2. Pang-Ning T, Steinbach M, Kumar V. Introdução ao data mining (mineração de dados). Rio de Janeiro: Ed. Ciência Moderna; 2009. 978 p.
3. Rob P, Coronel C. Sistemas de banco de dados: projeto, implementação e administração. São Paulo: Cengage Learning; c2011. 744 p.
4. Pinheiro CAR. Inteligência analítica: mineração de dados e descoberta de conhecimento. Rio de Janeiro: Ed. Ciência Moderna; 2008. 397 p.
5. Borém A, Del Giúdice MP, Sedyiama T, editores. Melhoramento genômico. Viçosa: UFV; 2003. 224 p.
6. Schaefer G, Nakashima T, Yokota Y, Ishibuchi H. Fuzzy Classification of gene expression data. IEEE International Fuzzy Systems Conference (FUZZ-IEEE); 2007 Jul 23-26; London, GB.
7. Wutao Chen, Huijuan Lu, Mingyi Wang, Cheng Fang. Gene Expression data classification using artificial neural network ensembles based on samples filtering. Proceedings of the 2009 International Conference on Artificial Intelligence and Computational Intelligence; 2009; Shangai, CN. Washington: IEEE Computer Society; c2009. Vol. 1, p. 626-8.
8. Houqin Bian, Chung R. Gene expression data classification using locality preserving projections. IEEE 11th International Conference on Bioinformatics and Bioengineering (BIBE); 2011 Oct 24-26; Taichung, TW.
9. Vecchiola C, Abedini M, Kirley M, Xingchen Chu, Buyya R. Gene Expression Classification with a novel coevolutionary based learning classifier system on Public Clouds. 2010 Sixth IEEE International Conference on e–Science Workshops; 2010 Dec 7-10; Brisbane, AU. Washington: IEEE Computer Society; c2010. p. 92-7.
10. Su Liangliang, Wang Nian, Tang Jun, Chen Le, Wang Ruiping. The Classification of Gene Expression Profile Based on the Adjacency Matrix Spectral Decomposition. International Conference on Computer Application and System Modeling; 2010 Oct 22-24; Taiyuan, CN.
11. Ghorai S, Mukherjee A, Sengupta S, Dutta PK. Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(3):659-71.
12. Carneiro NP, Carneiro AA. A era genômica: desvendando o código genético. Lavras: UFLA/FAEPE; 2002. 74 p. (Textos acadêmicos).
13. Piatetsky-Shapiro G, Tamayo P. Microarray data mining: facing the challenges. SIGKDD Explor. 2003;5(2):1-5.
14. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. 2000;24(3):227-35.
15. Amaral LR. Mineração de regras para classificação de oncogenes medidos por microarray utilizando algoritmos genéticos [dissertação]. Uberlândia: Universidade Federal de Uberlândia, Programa de Pós-graduação em Ciência da Computação; 2007.
16. Ooi CH, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics. 2003;19(1):37-44.
17. Xu R, Anagnostopoulos GC, Wunsch DC 2nd. Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. IEEE/ACM Trans Comput Biol Bioinform. 2007;4(1):65-77.
18. Lopes HS, Coutinho MS, Lima WC. An evolutionary approach to simulate cognitive feedback learning in medical domain. Proceedings of the 2000 Congress on Evolutionary Computation; 2000 Jul 16-19; La Jolla, US.
19. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc. 2002;97(457):77-87.
20. Deb K, Reddy AR. Classification of two and multi-class cancer data reliably using multi-objective evolutionary algorithms. KanGAL Report 2003. 10 p.
21. Liu JJ, Cutler G, Li W, Pan Z, Peng S, Hoey T, et al. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics. 2005;21(11):2691-7.
22. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinformatics. 2005;6:148.
23. Sihua Peng, Xiaoping Liu, Jiyang Yu, Zhizhen Wan, Xiaoning Peng. A new implementation of recursive feature elimination algorithm for gene selection from microarray data. Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering; 2009 Mar 31-Apr 2; Los Angeles, US. Washington: IEEE Computer Society; c2009. Vol. 3, p. 665-9.

Abstract

Introduction: In recent decades, cancer has been given a great dimension, becoming an evident world public health problem. The World Health Organization estimates that, in 2030, 27 million cancer cases and 17 million cancer deaths can be expected. Faced with this alarming scenario, Data Mining brings methods and tools to help building more meaningful knowledge about cancer. **Objective:** This paper aims to apply five traditional Data Mining methods on the NCI60 dataset. The database was created with data from microarray experiments, with levels of expression of 1,000 genes grouped into nine cancer classes. **Method:** The methods used on this paper are: J48, Random Forest, PART, IBK and Naive Bayes, which belong to the Weka environment, very traditional in Data Mining. Due to the low number of records for some cancer classes, the validation of the results obtained by the classifiers used the 3-fold cross validation. **Results:** The classifier which obtained the highest accuracy was IBK, while J48 and PART classifiers drastically reduced the set of genes, building high level knowledge as trees or rules. **Conclusion:** The results of this study can be used as tools aiming at assisting cancer cure research and may be used in the classification of new cases or to further improve understanding of gene/gene and gene/cancer relations.

Key words: Computational Biology; Gene Expression; Data Mining; Medical Oncology; Databases as Topic

Resumen

Introducción: En las últimas décadas, el cáncer se ha ganado una nueva dimensión, se hace evidente en un problema de salud pública en todo el mundo. La Organización Mundial de la Salud estima que en 2030, se puede esperar 27 millones de casos incidentes de cáncer. Ante este escenario alarmante, la minería de datos trae métodos y herramientas capaces de auxiliar en la construcción del conocimiento más relevante sobre el cáncer. **Objetivo:** Por lo tanto, este estudio tiene como objetivo aplicar los métodos tradicionales de minería de datos, aplicadas a la base de datos NCI60 construida con datos provenientes de experimentos de microarray niveles de expresión de 1.000 genes agrupados en nueve clases de cáncer. **Método:** Fue utilizado en este trabajo los métodos J48, Random forest, PART, IBK y Naive Bayes, que pertenecen al Weka. Todos los métodos son muy tradicionales en minería de datos. Debido al bajo número de registros para ciertas clases, que se utiliza en la validación de los resultados obtenidos por los clasificadores, la validación 3-*fold cross validation*. **Resultados:** El clasificador con la mayor precisión ha sido el IBK, mientras que el clasificadores J48 y PART lograron disminuir el conjunto de genes drásticamente, construyendo conocimiento de alto nivel en la manera de árboles o reglas. **Conclusión:** Los resultados de este trabajo se pueden utilizar como herramientas diseñadas para ayudar a hacer frente al cáncer, y puede ser utilizado en la clasificación de casos nuevos o profundizar en su comprensión de las relaciones gene/gene y genes/cáncer.

Palabras clave: Biología Computacional; Expresión Génica; Minería de Datos; Oncología Médica; Bases de Datos como Asunto