Mineração de Genes Ligados ao Câncer utilizando Ferramentas de Aprendizado de Máquina em Bases de Dados Biológicas

Fabrício Alves Rodrigues[1], **André Bevilaqua**[1],
Franciny Medeiros Barreto[1], **Reslley Gabriel Oliveira Silva**[1],
Thamer Horbylon Nascimento[1], **Laurence Rodrigues do Amaral**[1]

[1]Curso de Ciência da Computação - Universidade Federal de Goiás/Jataí (UFG)
BR364 Km 192, Setor Industrial, Jataí - GO - Brasil

{fabricio1989,laurence.amaral}@gmail.com

Abstract. *Bioinformatics originated in the 1960s, when computers emerged as important tools in molecular biology. The amount of information has reached significant proportions developing the necessity to store and organize these informations. Because of the need imposed by the advances in biological area, the databases of the referred area have become an essential part of the biological literature. Thus, the Machine Learning methods J48, Random Forest, Part, IBK and Naive Bayes were applied to NCI60 dataset. This dataset is one of the most complex to be mined, due to it's high dimensionality. IBK and Naive Bayes were the ones responsible for the best results obtained, while Part method was the one responsible for the worst result obtained.*

Resumo. *A Bioinformática têm suas origens na década de 1960, quando os computadores emergiram como ferramentas importantes na Biologia Molecular. Com o avanço da Bioinformática, a quantidade de informações atingiu grandes proporções, necessitando organização e armazenamento. Devido a este fato, os bancos de dados biológicos se tornaram uma parte vital na literatura biológica. Desta forma, aplicamos os métodos de aprendizado de máquina J48, Random Florest, Part, IBK e Naive Bayes ao dataset NCI60, um dos datasets mais complexos de serem minerados, devido à sua alta dimensionalidade. O IBK e o Naive Bayes foram os métodos que obtiveram as melhores avaliações enquanto que o PART foi responsável pelos piores índices de classificação.*

Palavras-chave: Bioinformática, Aprendizado de Máquina, Genes Ligados ao Câncer, Bases de Dados Biológicas

1. Introdução

A construção de um banco de dados é o processo de armazenar os dados em alguma mídia apropriada controlada pelo SGBD (Sistema Gerenciador de Banco de Dados). A manipulação inclui algumas funções, como pesquisas a fim de recuperar um dado específico, atualização para refletir as mudanças no minimundo (aspectos do mundo real) e gerar os relatórios dos dados. O compartilhamento permite aos múltiplos usuários e programas acessar, de forma concorrente, o banco de dados [Elmasri and Navate 2005]. Os SGBDs foram criados devido à precariedade dos sistemas de processamento de arquivos,

onde o armazenamento de informações utilizava vários arquivos do Sistema Operacional, precisando assim de diferentes aplicações a fim de efetuar modificações necessárias nos arquivos apropriados [Silberschatz et al. 2006].

A Mineração de dados é o processo de identificar, em grandes bases de dados, informações que às vezes podem ser ignoradas pelos usuários ou até mesmo por outras ferramentas tradicionais de busca de dados. Alguns métodos pertencentes à Mineração de dados podem ser utilizados como ferramenta de predição, isto é, eles possuem a capacidade de prever certas informações, baseando-se em padrões presentes na base [Tan et al. 2009].

Não se pode confundir os métodos da Mineração de dados com os métodos de busca de registros individuais de uma base de dados ou de busca de uma página na internet, por exemplo. Esse tipo de busca é chamada de recuperação de dados e a Mineração de dados tem sido usada para aprimorar essas ferramentas de recuperação [Tan et al. 2009].

A Mineração de dados também é classificada como KDD (*Knowledge Discovery in Database*). O KDD identifica os padrões e constrói conhecimento utilizando informações relevantes. Os métodos de KDD têm como objetivo gerar regras de associação, descrevendo padrões de relacionamento entre itens de um database [Dantas et al. 2008].

Quando utilizamos técnicas tradicionais de análise de dados encontramos algumas dificuldades práticas. As principais dificuldades estão relacionadas à: escalabilidade, alta dimensionalidade, dados complexos e heterogêneos e distribuição de dados. Esses são os maiores problemas encontrados em base de dados, e que motivam a utilização da mineração de dados nos dias de hoje [Tan et al. 2009].

A Bioinformática e a Biologia Computacional têm suas origens na década de 1960, quando os computadores emergiram como ferramentas importantes na Biologia Molecular. Este surgimento teria sido motivado por três fatores principais [Catanho and Miranda 2007]:

- pelo crescente número de sequências proteicas disponíveis, que representavam, ao mesmo tempo, uma fonte de dados e um conjunto de problemáticas importantes, porém intratáveis sem o auxílio de um computador;
- pela descoberta de que as macromoléculas carregam muita informação e se tornaram parte fundamental do modelo conceitual da Biologia Molecular;
- pela disponibilidade de computadores mais velozes nas universidades e centros de pesquisa.

Com o avanço cada vez mais rápido da Bioinformática, a quantidade de informações descobertas atingiu proporções consideráveis, levando à necessidade de que estas informações fossem organizadas e armazenadas para que pudessem ser utilizadas como base para outros estudos e consultas. A Bioinformática, que lida com dados de sequenciamento de DNA (sigla em inglês para ácido desoxirribonucleico), RNA (sigla em inglês para ácido ribonucleico) e proteínas, trabalha frequentemente com grande volume de informação, o que torna necessário o uso de técnicas que facilitem a manipulação e interpretação destes dados [Lorena and de Carvalho 2003]. Pela necessidade, a comunidade de Bioinformática passou a utilizar os bancos de dados biológicos (BDB) para o armazenamento de tais informações. Os BDBs podem ser compostos por informações

como sequências de DNA, genomas inteiros, estruturas tridimensionais de proteínas, dentre vários outros tipos de informações [Gibas and Jambeck 2001].

Neste trabalho aplicaremos os métodos J48, Random Forest, PART, Naive Bayes e IBK pertencentes ao ambiente Weka [Homes et al. 1994] à base de dados NCI60 [Ross et al. 2000] (descrita seção 2), e analisaremos cada método, pontuando quais foram os melhores e os piores métodos. É de suma importância ressaltar que a base de dados NCI60 é uma base biológica atípica, dotada de alta dimensionalidade, com baixo número de registros, com dados complexos e heterogêneos e com distribuições totalmente diferentes de aplicações tradicionais.

2. *Dataset* NCI60

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 um grande número de genes distribuídos por um braço robótico de forma ordenada (*spots*) sobre placas de vidro. As sondas podem ser conjuntos de cDNAs (DNA complementar de fita simples) gerados a partir de células ou tecidos em duas situações diferentes, que se deseja comparar. Os resultados são produzidos sob forma de diferentes intensidades de fluorescência que são captadas por microscopia a laser em função dos diferentes níveis de expressão de cada gene [Carneiro and Carneiro 2002].

A imagem dos pontos fluorescentes é processada por meio de métodos computacionais com o objetivo de calcular a intensidade obtida para cada mRNA (RNA mensageiro). 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, com grande aplicabilidade nos estudos de biologia molecular e fisiologia vegetal [Carneiro and Carneiro 2002].

Como exemplo do resultado obtido por essa técnica, podemos citar a base NCI60 [Ross et al. 2000] utilizada em nosso trabalho. Essa base de dados faz parte do NCI60 *Cancer Microarray Project*, projeto este, advindo da colaboração entre o laboratório *Brown/Bolstein* do grupo *John Weinstien's do Laboratory of Molecular Pharmacology* e do *Laboratory of Developmental Therapeutics*, ambos pertencentes ao *National Cancer Institute*, nos EUA.

Para a construção desta base, foram utilizados *microarrays* de cDNA na busca de expressões gênicas de aproximadamente 8.000 genes distintos. Estes genes, oriundos de 61 linhagens celulares, foram classificados em 9 (nove) classes de câncer: (1) mama, (2) sistema nervoso central, (3) cólon, (4) leucemia, (5) melanoma, (6) pulmão, (7) ovário, (8) renal e (9) células reprodutivas. Os números entre parênteses 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 [Amaral 2007].

No trabalho de Ooi e Tan [Ooi and Tan 2003] 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

Tabela 1. Visão geral da base NCI60 reduzida e utilizada nos experimentos de Ooi e Tan [Ooi and Tan 2003]

Amostra	Expressão Gene 1	Expressão Gene 2	Expressão Gene 3	...	Expressão Gene 999	Expressão Gene 1000	Classificação
1				...			
2				...			
3				...			
...	...	...	...	...	...	...	...
60				...			
61				...			

maiores valores de desvio padrão na base NCI60. Estes 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 e positivos, sendo obtidos através das intensidades dos pontos fluorescentes obtidos no *microarray*.

Escolhemos o *dataset* NCI60 nesta análise devido ao seu elevado grau de interesse dentro da área de Bioinformática. Segundo Xu e colaboradores [Xu et al. 2007], é muito difícil propor regras ou critérios na determinação de um conjunto de genes que seja discriminantes no diagnóstico de doenças, especialmente quando as bases de dados estudadas possuem um elevado número de classes, tais como a complexa NCI60 [Ross et al. 2000]. A base NCI60 é considerada um desafio para os algoritmos de classificação por suas características peculiares: um número relativamente alto de classes (9) para um número relativamente baixo número de amostras (61), resultando em número baixo de amostras por classe, variando de 4 a 9 amostras por classe.

Além disso, até 2030 o câncer deve matar por ano, 13,2 milhões de pessoas em todo o mundo, números estes apresentados pela Agência Internacional para Pesquisa sobre Câncer das Nações Unidas (Iarc). Em 2008, o número de mortes por câncer chegou a 7,6 milhões. A pesquisa constatou ainda que serão diagnosticados 21,4 milhões de casos por ano nas próximas duas décadas. Desta forma, torna-se importantíssimo o estudo e desenvolvimento de bancos de dados e ferramentas de análise e mineração de dados voltados para esta doença, fornecendo assim, ferramentas que possam auxiliar no prognóstico do câncer.

3. Métodos de Aprendizado de Máquina Selecionados

Neste trabalho nós utilizamos os métodos J48, Random Florest, Part, IBK e Naive Bayes, todos pertencentes ao Weka [Homes et al. 1994], ferramenta esta extensamente utilizada na área de Aprendizado de Máquina e Mineração de Dados. Escolhemos estes métodos por serem os mais populares e gerarem, como saída, conhecimento de alto nível na forma de árvores ou regras.

Utilizando o Weka, nós submetemos o *dataset* NCI60 aos métodos supracitados e os resultados obtidos serão apresentados na seção 4. Para validar os resultados obtidos, utilizamos o 3-*fold cross validation* e a avaliação final do classificador é dada pela

multiplicação de termos comumente utilizados em domínios médicos chamados sensibilidade e especificidade. No cálculo destes dois termos utilizamos a matriz de confusão dada como saída do Weka. Os detalhes deste cálculo é apresentado na subseção 3.1.

3.1. Função de Avaliação - FA

Neste trabalho a FA avalia a qualidade do resultado de cada método. A FA aqui aplicada pode ser encontrada em [Lopes et al. 2000]. Para o perfeito entendimento da FA aqui aplicada, alguns conceitos precisam ser reforçados. Quando utilizamos uma determinada solução na classificação de um exemplo, quatro diferentes tipos de resultados podem ser observados, dependendo da classe predita pela solução e a verdadeira classe do exemplo. São eles:

- *True Positive* (tp) - A solução prediz que o exemplo pertence a uma determinada classe e o mesmo pertence;
- *False Positive* (fp) - A solução prediz que o exemplo pertence a uma determinada classe mas o mesmo não pertence;
- *True Negative* (tn) - A solução prediz que o exemplo não pertence a uma determinada classe e o mesmo não pertence;
- *False Negative* (fn) - A solução prediz que o exemplo não pertence a uma determinada classe mas o mesmo pertence;

A FA utiliza dois indicadores comumente utilizados em domínios médicos, chamados de sensibilidade (*Se*) e especificidade (*Sp*). Se e Sp são definidos abaixo:

$$Se = \frac{tp}{(tp + fn)} \tag{1}$$

$$Sp = \frac{tn}{(tn + fp)} \tag{2}$$

Finalmente, a FA utilizada é definida como o produto destes dois indicadores, *Se* e *Sp*, como segue abaixo:

$$Aptidao = Se * Sp \tag{3}$$

O objetivo do trabalho é maximizar ao mesmo tempo *Se* e *Sp* e consequentemente *Aptidao*, utilizando para isso, as seguintes equações (1), (2) e (3). Em cada execução, o nosso AG trabalha com um problema de classificação de duas classes, isto é, quando o AG está procurando por regras de uma dada classe, todas as outras classes são agrupadas em uma única classe.

4. Resultados obtidos

A Tabela 2 ilustra os resultados obtidos para os métodos J48, Random Florest, PART, IBK e Naive Bayes, analisando cada uma das 9 classes de câncer separadamente.

Para a classe C_1, o melhor método foi o IBK (63%, resultado este obtido na função de avaliação) e o pior foi o Random Forest (51%). Com relação à classe C_2, o melhor foi o PART (74%) e o pior foi o Naive Bayes (50%). Para a classe C_3, o melhor foi o IBK (98%) e o pior foi também o Naive Bayes com 79% de avaliação.

Com relação à classe C_4, os métodos que obtiveram os melhores resultados foram o Random Forest e o IBK, ambos com 82%. O pior método para a classe C_4 foi o Naive

Tabela 2. Resultados obtidos pelos 5 métodos nas 9 classes de câncer

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	0,7	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**	

Bayes. Para a classe C_5, o melhor e o pior método foi o IBK e o J48, com 94% e 67%, respectivamente. Para a classe C_6 o melhor método foi o Random Forest com 64% e o pior foi o PART com 45%.

Para a classe C_7, o melhor método encontrado foi o Naive Bayes (50%) e o pior foi o IBK, com 37%. Para a classe C_8, com 75% de avaliação o melhor método encontrado foi o Naive Bayes e o pior foi o PART, com 35%. Para a classe C_9, a melhor avaliação, com 72%, foi o IBK e o pior foi o PART com 49% de avaliação.

Confrontando os melhores e os piores resultados obtidos por classe, a maior diferença foi encontrada para a classe C_8 (com 40%, 0,75 contra 0,35), seguida pela classe C_5, com 27% de diferença. As diferenças pertentes às classes C_2, C_9, C_3, C_6, C_7, C_1 e C_4 foram de 24%, 23%, 20%, 19%, 13%, 12% e 8%, respectivamente.

Tomando como parâmetro os valores médios obtidos pelos 5 métodos em cada uma das 9 classes, podemos perceber que a classificação de algumas classes são tarefas mais fáceis que em outras, isto é, conseguimos valores médios mais altos em algumas classes e em outras classes não. Os melhores valores médios foram obtidos para a classe C_3, seguido pelas classes C_4 e C_5. Os piores resultados médios foram obtidos para as classes C_7 seguido pelos valores obtidos para as classes C_6 e C_8.

Tentando apontar um método que obtivesse boas classificações em todas as classes, calculamos a média aritmética simples em todas as classes para cada um dos 5 métodos analisados. Os resultados desta análise podem ser observados na Figura 1.

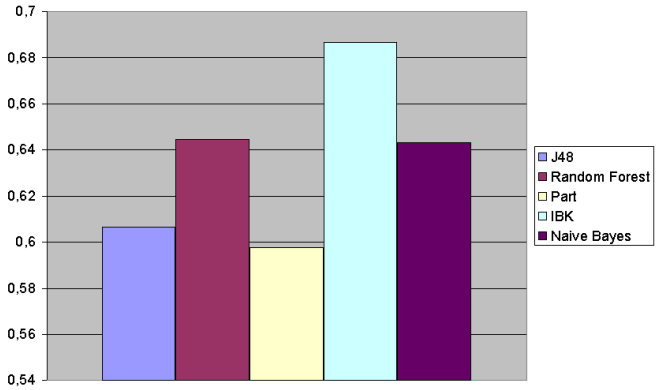

Figura 1. Médias obtidas para todas as 9 classes de câncer

O método que obteve a melhor média em todas as 9 classes foi o IBK, com média de 0,686, sendo 4,22% melhor que o segundo colocado (Random Forest), 4,33% melhor

Tabela 3. Melhores e Piores encontrados para cada uma das 9 classes de câncer

Classe	Pior Método	Melhor Método
C_1	Random Forest	IBK
C_2	PART	Naive Bayes
C_3	Naive Bayes	IBK
C_4	Naive Bayes	IBK e Random Forest
C_5	J48	IBK
C_6	PART	Random Forest
C_7	IBK	Naive Bayes
C_8	PART	Naive Bayes
C_9	PART	IBK

que o terceiro (Naive Bayes), 8% melhor que o J48 e 8,88% melhor que o PART. A Tabela 3 ilustra os melhores e os piores métodos para cada classe.

O método que obteve os melhores resultados em cada classe foi o IBK, obtendo os melhores resultados em 5 das 9 classes (55,56% das classes). Outro método que obteve bons resultados por classe foi o Naive Bayes, obtendo os melhores resultados em 3 classes (33,34% das classes). Desta forma, estes dois métodos obtiveram os melhores resultados em 88,9% das classes, isto é, 8 em 9 possíveis.

Por outro lado, o método PART obteve os piores resultados, sendo ó pior em 4 classes, isto é, ele foi o pior em 44,45% das classes. Curiosamente, o Naive Bayes (que obteve os melhores resultados em 3 classes), obteve os piores resultados em duas classes. Desta forma, podemos concluir que este método é bem específico a algumas das 9 classes de câncer do NCI60.

5. Conclusões e Trabalhos Futuros

Como descrito na seção 2, o câncer é uma doença responsável por milhões de mortes anualmente. Quando mais cedo ocorrer o diagnóstico da doença, maior as chances de cura. Desta forma, a construção de conhecimento, utilizando métodos de aprendizado de máquina e mineração de dados, voltado para a área médica está conquistando um espaço cada vez maior. Ferramentas computacionais estão sendo usadas no auxílio do prognóstico de doenças, aumentando a eficácia e diminuindo o número de mortes.

Baseado neste fato, os métodos apresentados acima conseguiram valores expressivos na classificação de novos casos de câncer, selecionando genes que podem estar relacionados a uma das 9 classes de câncer descritas. Estes genes são chamados de oncogênese. Desta forma, como trabalhos futuros, pretendemos analisar cada um dos resultados gerados por cada método, analisando não somente taxas de classificação, mas também o conhecimento gerado por estes métodos. Interoperabilidade e compreensibilidade dos resultados gerados são muito importantes em um classificador, pois os resultados gerados serão analisados por especialistas da área. Quanto maior o nível deste conhecimento, melhor o entendimento do mesmo.

Referências

Amaral, L. R. (2007). Mineração de regras para classificação de oncogenes medidos por microarray utilizando algoritmos genéticos. Dissertação de mestrado, Pós-graduação em Ciência da Computação, Universidade Federal de Uberlândia.

Carneiro, N. P. and Carneiro, A. A. (2002). *A Era Genômica - Desvendando o Código Genético*. UFLA.

Catanho, M. and Miranda, A. B. D. (2007). Comparando genomas: bancos de dados e ferramentas computacionais para a análise comparativa de genomas procarióticos. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*.

Dantas, E. R. G., Junior, J. C. A. P., Lima, D. G., and Azevedo, R. R. (2008). O uso da descoberta de conhecimento em base de dados para apoiar a tomada de decisões. *V Simposio de Excelência em Gestão e Tecnologia - SEGeT*, 01:50–60.

Elmasri, R. and Navate, S. B. (2005). *Sistemas de banco de dados*. Pearson Addison Wesley, São Paulo.

Gibas, C. and Jambeck, P. (2001). *Desenvolvendo Bioinformática: ferramentas de software para aplicações em biologia*. Campus, Rio de Janeiro.

Homes, G., Donkin, A., and Witten, I. H. (1994). Weka: A machine learning workbench. *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*.

Lopes, H. S., Coutinho, M. S., and Lima, W. C. (2000). An evolutionary approach to simulate cognitive feedback learning in medical domain. In *Congress on Evolutionary Computation - (CEC-2000)*. La Jolla, CA, USA.

Lorena, A. C. and de Carvalho, A. C. P. L. F. (2003). Utilização de técnicas inteligentes em bioinformática. Technical report, Universidade de São Paulo - ICMC.

Ooi, C. H. and Tan, P. (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1):37–44.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*.

Silberschatz, A., Korth, H. F., and Sudarshan, S. (2006). *Sistema de banco de dados*. Elsevier, Rio de Janeiro.

Tan, P.-N., Steinbach, M., and Kumar, V. (2009). *Introdução ao Datamining*. Ciência Moderna, Rio de Janeiro.

Xu, R., Anagnostopoulos, G. C., and II, D. C. W. (2007). Multiclass cancer classification using semisupervised ellipsoid artmap and particle swarm optimization with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(1).