Modelos Estatísticos Aplicados à Biologia Computacional e Medicina de Precisão

Palavras-chave: BIOESTATÍSTICA, GENÔMICA, APRENDIZADO ESTATÍSTICO DE MÁQUINA

Autores/as:
Mariângela Lima Rodrigues [Universidade Estadual de Campinas]
Prof. Dr. Benilton de Sá Carvalho (Orientador) [Universidade Estadual de Campinas]
Co-autores/as:
Bruna Caroline Dias Silva [Universidade Estadual de Campinas]

1. O Projeto de Pesquisa

O processo de tomada de decisão permeia todos os âmbitos da vida e o seu desenvolvimento, e este processo ocorre baseado no estudo de evidências. Neste contexto, fica clara a responsabilidade da Estatística como contribuidora fundamental do estudo dos dados e das informações neles contida, de modo a fornecer material suficiente para que seja possível decidir pela melhor escolha com maior confiabilidade e precisão.

No âmbito da Biologia Computacional e Medicina de Precisão é de suma importância que sejam desenvolvidas novas técnicas de análise de dados, a fim de que seja possível extrair informações de forma mais eficiente a partir de bases de dados de alta dimensão. Com isso em mente, o enfoque deste trabalho foram os problemas que cercam a Biologia Computacional, Estatística e Medicina de Precisão.

2. Introdução à Biologia Molecular e o problema da Classificação de Dados

Os seres vivos são constituídos por um conjunto de células as quais são responsáveis por formar todo o organismo e seu funcionamento. Dentro dessas células está contido todo o conteúdo genético que identifica cada ser vivo, o genoma. Parte da carga genética está no que chamamos de ácido desoxirribonucleico, o DNA, que trata de uma fita de dupla hélice formada por bases nitrogenadas, açúcar e fosfato, e nele as diferentes combinações das bases nitrogenadas - a Adenina, Timina, Citosina e Guanina - geram todas as características biológicas do ser vivo.

Quando falamos sobre a reprodução de seres vivos podemos mencionar o processo de hereditariedade de informações genéticas. Neste contexto, acerca do organismo humano, surgiu o interesse de estudar os impactos da transmissão genética aos descendentes, uma vez que o conhecimento a respeito da genética do paciente pode fornecer diretrizes para a conclusão a respeito do diagnóstico, prover ferramentas para tratamentos preventivos de doenças para as quais observou-se predisposição, entre outros benefícios. Sendo assim, evidencia-se a importância do sequenciamento do genoma humano para a determinação de características normais e patológicas que persistem no desenvolvimento da vida, e a Biologia Molecular e Medicina de Precisão se dedicam em partes a este fim, sendo uma das técnicas mais utilizadas para o sequenciamento do genoma a Análise de Dados de Microarranjos.

Microarranjos são placas de vidro ou acrílico que são dispersas em forma matricial e nelas estão contidas moléculas de DNA sintético, denominadas sondas. No processo de experimento com microarranjos é realizada a fragmentação e amplificação de amostras de DNA, depois é feita a rotulagem e hibridização da amostra,

e por fim a quantificação da molécula-alvo em cada sonda, que é feita através da análise de fluorescência (Bolstad, B. M. (2004)).

Uma das aplicações no estudo de dados de microarranjos é a identificação e classificação de genótipos, e no contexto de classificação de dados ressalta-se que este processo trata da obtenção de uma resposta para uma pergunta de interesse, que em termos gerais pode ser resumida a: "a qual grupo uma dada observação pertence?".

Neste trabalho o objetivo principal foi o estudo e aplicação de modelos de mistura finita e técnicas de aprendizado de máquina - em suas três vertentes principais (apresentadas em maiores detalhes em Sousa, H. M. d. R. (2019) e De Souto, M. C. P., et. al (2003)) -, no processo de identificação e classificação de genótipos.

3. Metodologia e Análise de dados

A fim de exemplificar o processo de identificação e classificação de genótipos foram utilizados dados, disponibilizados pelo projeto internacional HapMap, de microarranjos de 1.000 SNPs diferentes, sendo que para cada SNP foram observados 1.130 indivíduos diferentes e coletadas as informações com respeito a intensidade da sonda para cada indivíduo para cada alelo, sendo dois alelos no total. Além disso, os dados trazem a informação, para uma grande maioria das observações, com respeito ao genótipo observado para cada indivíduo, para o seu respectivo SNP.

A análise dos dados foi realizada sob três óticas distintas: Algoritmo EM, Algoritmo CRLMM e Árvore de Inferência Condicional. Para a aplicação das técnicas mencionadas foi calculado o valor da log-razão das intensidades (denotada por M) e a log intensidade média (S) para cada intensidade observada dos alelos A e B. A equação abaixo descreve a obtenção dessas estatísticas.

$$M = log_2(\frac{I_A}{I_B})S = \frac{1}{2}log_2(I_A) + log_2(I_B), \tag{1}$$

3.1 Algoritmo EM

O Algoritmo EM trata de uma técnica de aprendizado não-supervisionado que resumidamente consiste em um método iterativo utilizado para calcular os estimadores de máxima verossimilhança no contexto de variáveis latentes, onde sua primeira etapa (E) consiste em calcular o valor esperado da log-verossimilhança utilizando a distribuição condicional obtida através dos valores dos parâmetros já estimados. Em seguida, é realizada a etapa de maximização (M), que visa otimizar a função encontrada no passo (E), ou seja, nesta etapa os parâmetros são atualizados. Neste trabalho sua aplicação se deu a partir da função `mvnormalmixEM` disponível no pacote `mixtools` do *software* R Studio 4.0.3.

Na Figura 1, é possível observar a distribuição e classificação do genótipo dos indivíduos para o SNP_A-1899111 junto a resposta do algoritmo EM. Além disso a Figura 1 também apresenta esse resultado para o SNP_A-2139184. Note que o desempenho do algoritmo foi bastante satisfatório, para o primeiro SNP, e para as observações para as quais não se conhece o genótipo, estas foram classificadas em alguns dos três grupos existentes, e a classificação ocorreu de acordo com o esperado. Já para o SNP_A-2139184, nota-se que devido à divisão entre os grupos não ser clara, o algoritmo teve um desempenho inferior na classificação das observações, mas esse desempenho não foi insatisfatório.

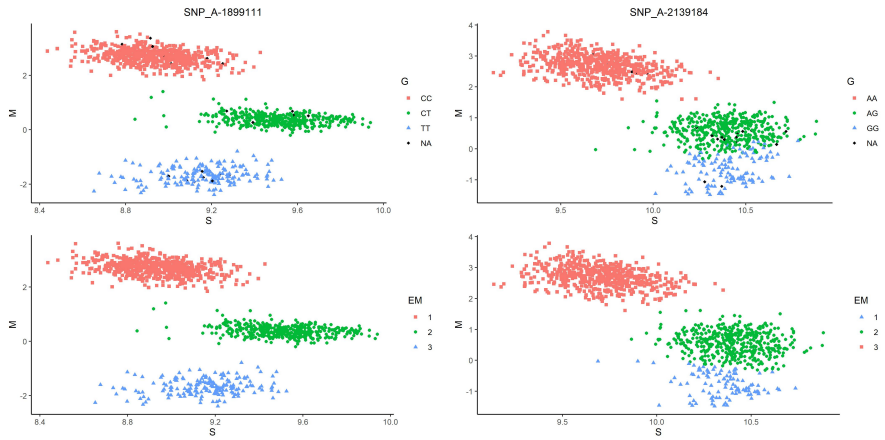

Figura 1: Classificação dos genótipos utilizando o algoritmo EM.

3.2 Algoritmo CRLMM

O Modelo Linear Robusto Corrigido com Distância de Mahalanobis (CRLMM) é uma técnica alternativa no processo de identificação e classificação dos genótipos, pois trata de um algoritmo de aprendizado de máquina semi-supervisionado que permite o uso da informação contida nos dados disponibilizados pelo projeto HapMap garantindo que o modelo seja previamente treinado tornando-se capaz de classificar dados de novas amostras com maior eficiência. É interessante ressaltar que ao contrário de outros métodos de genotipagem o CRLMM modela a log-razão de intensidade, ao invés do par de intensidades, e faz uso da distância de Mahalanobis para decidir a qual grupo será alocada uma observação. Este fato é demasiadamente importante, uma vez que a log-razão se mostrou mais eficiente em distinguir os grupos e, por isto, confere maior robustez ao modelo de classificação.

A aplicação do CRLMM se deu a partir do *software* R Studio 4.0.3 através da função `crlmm` disponível no pacote `crlmm` da plataforma Bioconductor. Como na apresentação do algoritmo EM, foram selecionados alguns SNPs para fins de análise do desempenho do modelo.

A Figura 2 apresenta o SNP_A-1899111 e o SNP_A-2139184, nela é possível observar que o algoritmo teve bom desempenho, revelando acurácia de 100% e 96,66% para cada um dos SNPs, respectivamente.

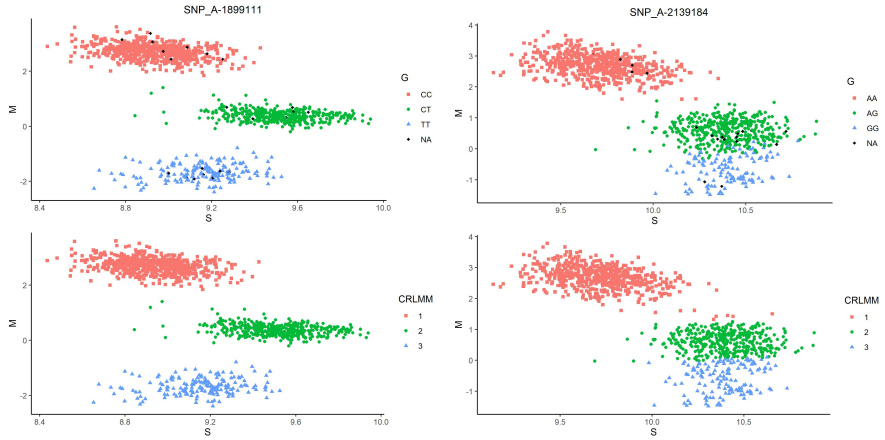

Figura 2: Classificação dos genótipos utilizando o algoritmo CRLMM.

3.3 Árvore de Inferência Condicional

A árvore de decisão é um método de aprendizado supervisionado utilizado tanto em problemas de classificação como de regressão. Sua premissa básica é subdividir o problema em subproblemas mais simples até que se possa encontrar uma solução. Uma das principais vantagens na utilização de árvores de decisão é a simplicidade de sua interpretação, uma vez que sua representação gráfica apresenta, de forma clara, o caminho para determinar a classe de uma observação. A árvore de inferência condicional é um tipo de árvore de decisão na qual o critério para particionamento dos grupos se dá por meio de testes estatísticos que avaliam a associação entre a variável resposta e as covariáveis, optando por aquela que possui a maior a relação com a variável de interesse para iniciar o particionamento, estabelecendo dessa forma uma hierarquia, de modo a obter grupos com observações semelhantes.

Para a aplicação da metodologia referida foi utilizada a função `ctree()` do pacote `partykit` disponível no *software* R Studio 4.0.3. Inicialmente o conjunto de dados foi separado de maneira aleatória entre treino (utilizando 70% dos dados) e teste (utilizando 30% dos dados). Parte do desempenho da metodologia é apresentado nas Figuras 3 e 4 que trazem o resultado da classificação para os SNPs A-1899111 e A-2139184, respectivamente. Note que para o SNP_A-1899111, duas regras de classificação foram suficientes para distinguir os três grupos de forma clara. Já para o SNP_A-2139184, o número de condições resultantes foi um pouco maior, mas ainda sim o ajuste obteve um bom desempenho, resultando em uma acurácia de 96,4%.

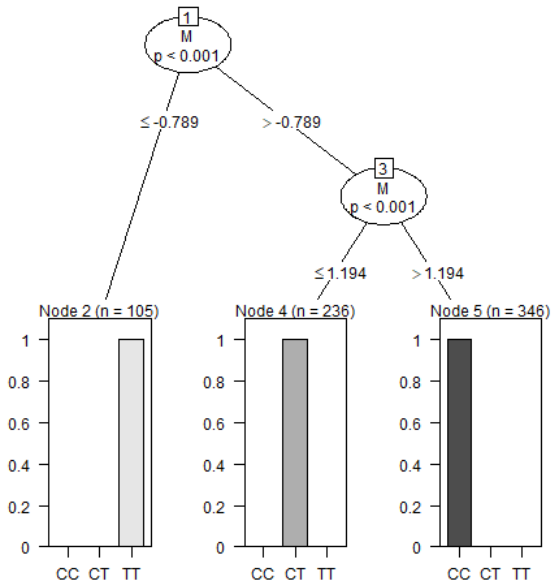

Figura 3: Classificação dos genótipos utilizando Árvore de Inferência Condicional SNP A-1899111.

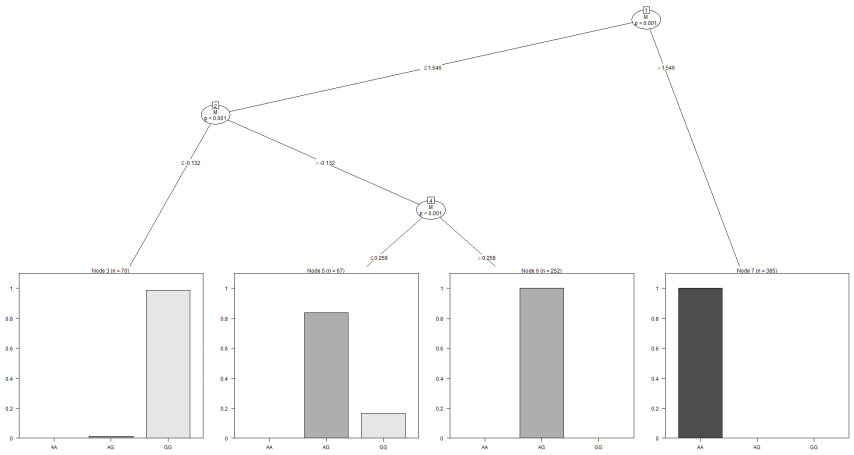

Figura 4: Classificação dos genótipos utilizando Árvore de Inferência Condicional SNP A-2139184.

4. Conclusões

Apesar de observarmos um bom desempenho para as três metodologias apresentadas é fundamental ressaltar que existem diferenças significativas entre elas, desde o tipo de aprendizado de máquina até a técnica estatística utilizada. Dito isto, denota-se que em se tratando dos algoritmos EM e CRLMM, ambos são eficientes, mas é preferível o uso do CRLMM uma vez que este permite o uso de informações prévias - como o conjunto de dados padrão ouro do HapMap - para treinamento do algoritmo, a fim de que seja possível obter melhor desempenho, com maior nível de confiabilidade, na identificação e classificação de genótipos de novas observações.

Já com relação a árvore de inferência condicional, ressalta-se que esta é uma técnica de aprendizado supervisionado bastante enviesada e que tem poder preditivo pequeno comparada às demais técnicas apresentadas. Entretanto, com o auxílio de outras técnicas estatísticas, como por exemplo a Regressão Linear, é possível fazer uso de seus resultados de modo a tornar a técnica eficiente para classificação de genótipos.

5. Bibliografia

Bolstad, Benjamin Milo. Low-level analysis of high-density oligonucleotide array data: background normalization and summarization. University of California, Berkeley, 2004.

De Souto, M. C. P., et al. “Técnicas de aprendizado de máquina para problemas de biologia molecular.” Sociedade Brasileira de Computaçao 1.2 (2003).

Sousa, Heidi Mara do Rosário. “Estudo de modelos de classificação com aplicação a dados genômicos.” (2019).