Resumo do trabalho: Métodos de aprendizado de máquina supervisionado para classificação aplicados a dados de microarranjo de DNA.
Aluna: Ana Carolina Alves Oliveira.
Orientadora: Samara Flamini Kiihl.

1 Introdução

Ao longo dos últimos anos, a geração de uma grande quantidade de dados genômicos e proteômicos tem resultado em um acúmulo de dados biológicos que necessitam ser interpretados (Raza, 2010). As áreas de bioinformática e bioestatística vêm desenvolvendo algoritmos capazes de processar e analisar um volume cada vez maior de dados biológicos, muitos dos quais, provenientes de chips de microarranjo.

Os dados de microarranjo são oriundos de técnicas que consistem na análise do nível de luz fluorescente captada por meio de imagens de chips de microarranjo, sendo possível quantificar e qualificar expressões de genes de determinado organismo. Os algoritmos de aprendizado de máquina, muito utilizados nesse contexto, surgiram da necessidade de automação das análises de dados. Quando aplicados a dados de microarranjo, podem potencializar novos desenvolvimentos e contribuir para o diagnóstico e tratamento de diversas doenças.

O presente trabalho consiste no estudo e aplicação de algoritmos de aprendizado de máquina supervisionado para classificação, construindo classificadores que sejam capazes de auxiliar no processo de identificação de pacientes com determinada doença, nesse caso, o Transtorno do Espectro Autista (TEA). Será feita a comparação entre acurácias de classificação de fenótipos em duas metodologias: k-vizinhos mais próximos e análise de discriminante linear.

2 Materiais e Métodos

Foi utilizado o conjunto de dados de Kuwano *et al* (2011), disponível em domínio público na plataforma *Gene Expression Omnibus* (GEO), sob o número de série GSE26415. Os dados consistem de 84 amostras de sangue venoso de indivíduos distribuídos nos seguintes grupos: adultos com TEA (21), mães saudáveis que tiveram filhos com TEA (21) e grupo controle pareado com mesma idade e sexo de ambos grupos anteriores (42). Além disso, há expressões de 19195 genes registradas. No presente trabalho operou-se sob os dados já pré-processados, informações adicionais podem ser verificadas em Kuwano *et al* (2011).

Todo desenvolvimento computacional foi realizado no *software* RStudio, a análise é reprodutível e está disponível no *GitHub*, através do link: *github.com /anacarolina-estat/Iniciacao-Cientifica-CNPq*. O *download* dos dados foi feito utilizando as

bibliotecas *Biobase* e *GEOquery*. Os algoritmos de aprendizado de máquina foram aplicados dispondo, principalmente, do pacote caret. Ele possui ferramentas de divisão dos dados, pré-processamento, ajuste, entre outras. Foram escolhidos os métodos K-Vizinhos Mais Próximos e Análise de Discriminante Linear, ambos por serem clássicos e conhecidos e pela simples compreensão quando se está iniciando os estudos em aprendizado de máquina.

Para treinar o algoritmo os dados foram divididos aleatoriamente de forma que 75% correspondem a dados de treinamento e 25% de teste. A variável dependente que se tem interesse em fazer previsões é o grupo/classe em que o indivíduo pertence: adulto com TEA, mães saudáveis que tiveram filhos com TEA ou grupo controle.

3 Resultados

3.1 K-Vizinhos Mais Próximos

Usando o algoritmo K-Vizinhos Mais Próximos nos dados de treinamento, com validação cruzada utilizando 5 dobras, foram obtidos os resultados presentes na Tabela 1. Nela são exibidos as métricas acurácia (proporção de casos previstos corretamente) e coeficiente Kappa (proporção de observações concordadas pelo valor real e predito) para diferentes valores de k, encontrados via validação cruzada. A partir dos resultados, o algoritmo seleciona automaticamente o melhor valor para k, que nesse caso corresponde a k = 3.

Tabela 1: Métricas do algoritmo KNN nos dados de treinamento, com validação cruzada com 5 dobras.

k	Acurácia	Kappa
1	0.57	0.34
2	0.59	0.36
3	0.67	0.46
4	0.60	0.36
5	0.57	0.32

De posse do modelo treinado é possível fazer previsões com os dados de teste e montar a matriz de confusão, que retorna os valores reais e preditos pelo algoritmos (Tabela 2).

Tabela 2: Matriz de confusão do algoritmo KNN aplicado aos dados de teste.

Predição	Referência		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Adultos com TEA	2	1	3
Mães saudáveis com filhos com TEA	3	0	2
Grupo controle	2	2	6

Ainda é possível explorar um pouco mais os valores previstos pelo modelo. Para isso são observadas algumas métricas da matriz de confusão, além da acurácia: a sensibilidade e especificidade. A primeira mensura a proporção de casos positivos que foram identificados corretamente, enquanto a segunda, a proporção de casos negativos também identificados corretamente. Esses resultados estão apresentados nas Tabelas 3.

Tabela 3: Métricas do algoritmo KNN aplicado aos dados de teste.

Métrica	Classe		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Sensibilidade	0.29	0	0.55
Especificidade	0.71	0.72	0.60
Acurácia	0.50	0.36	0.57

3.2 Análise de Discriminante Linear

Para treinar o algoritmo LDA foram usados os dados processados deixando os normalizados e considerando as 10 primeiras componentes principais, a fim de reduzir o número de preditores e tornar o modelo mais simples. Outros valores de componentes foram testados mas não melhoraram o desempenho algoritmo. As métricas de precisão (acurácia) e coeficiente Kappa são 0.57 e 0.31, respectivamente. A comparação entre valores reais e previsões encontradas a partir da aplicação do modelo nos dados de teste é mostrada na Tabela 4. Além disso, na Tabela 5 é possível visualizar métricas importantes da matriz de confusão.

Tabela 4: Matriz de confusão do algoritmo LDA aplicado aos dados de teste.

Predição	Referência		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Adultos com TEA	3	0	2
Mães saudáveis com filhos com TEA	2	1	2
Grupo controle	2	2	7

Tabela 5: Métricas algoritmo LDA aplicado aos dados de teste.

Métrica	Classe		
	Adultos com TEA	Mães saudáveis com filhos com TEA	Grupo controle
Sensibilidade	0.43	0.33	0.64
Especificidade	0.86	0.78	0.60
Acurácia	0.64	0.56	0.62

4 Discussões

O presente trabalho foi desenvolvido com o intuito de aplicar métodos de aprendizado de máquina supervisionado para classificação em um conjunto de dados de microarranjo de DNA. Foram escolhidos dados de expressão gênica relacionados ao autismo e os algoritmos k-vizinhos mais próximos e análise de discriminante linear.

Foi levantada a relevância das técnicas de microarranjos, uma vez que, ela permite coletar dados extremamente importantes para detecção de doenças. Aliado ao aprendizado de máquina se torna possível otimizar a precisão dos diagnósticos e o tempo no qual são realizados. Infelizmente ambos os algoritmos executados não tiveram desempenhos tão satisfatórios quanto se esperava, não sendo possível classificar tão precisamente os indivíduos no seus respectivos grupos. Embora o KNN tenha revelado uma melhor acurácia e Kappa quando treinado, nota-se que ao ser aplicado ao conjunto de dados de teste, não foi o que obteve as melhores métricas, especificamente quando se trata da classificação de indivíduos na classe “mães saudáveis que tiveram filhos com TEA”.

Uma possibilidade para melhoria no desempenho do modelo e obtenção de uma melhor discriminação dos grupos é realizar um filtro de gene. Esse processo consiste em analisar e selecionar genes que sejam mais expressivos entre os demais e posteriormente aplicar os algoritmos de aprendizado de máquina, levando em consideração apenas os que foram selecionados.