Seleção de características para classificação de dados de expressão gênica

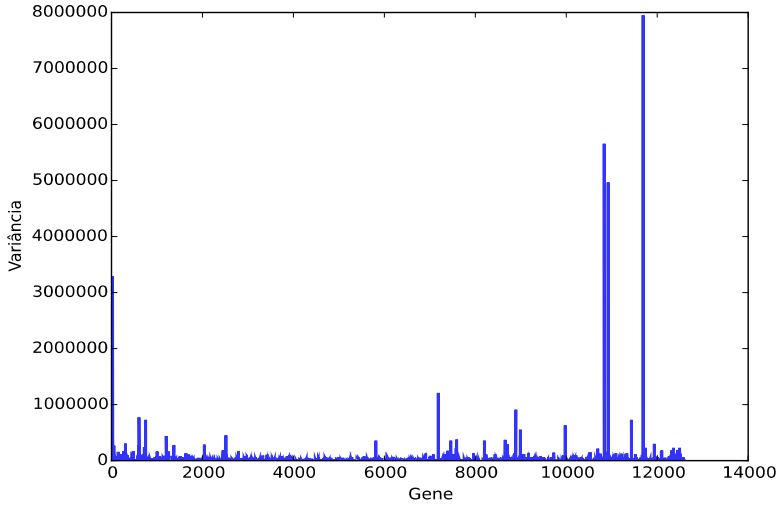
Pedro Sousa Lacerda
<pedro.lacerda@ufabc.edu.br>
UFABC

Introdução

A classificação molecular do câncer pela análise discriminatória de dados de expressão gênica vindos de experimentos de microarrays de DNA é caracterizada pela medição de muitas variáveis *p* (genes) em poucas amostras *N*, situação desfavorável ($N < p$) para a aplicação da maioria das técnicas de classificação por aprendizagem. Torna-se necessária a redução da dimensionalidade do espaço gênico *p*-dimensional para um espaço menor (Nguyen & Rocke, 2002), pois, além de diminuir o ruído nos dados e aumentar eficiência e eficácia dos algorítmos, a redução de dimensionalidade, facilita a identificação dos genes relacionados com o étimo do problema a ser classificado GUYON&ELISSEEF (Ding & Peng, 2003). Em muitos estudos, mesmo após a redução, continua sendo imprescindível a consideração de muitos genes (50~2000) para discriminar entre duas ou mais classes (Ai-Jun & Xin-Yuan, 2010).

A base de dados utilizada possui amostras dos níveis de expressão gênica de pacientes com câncer de próstata (52) e de indivíduos saudáveis (50) (Golub et al., 1999). A quantidade de genes mensurados (126000) é muito maior que a quantidade de amostras (102), confirmando a necessidade de redução de dimensionalidade para aplicação das técnicas de classificação. Observamos no Gráfico 1 a grande quantidade de atributos (genes) de baixa variância.

Gráfico 1: Grande parte dos genes apresentam pouca variância, um indicador de baixo poder de explicação



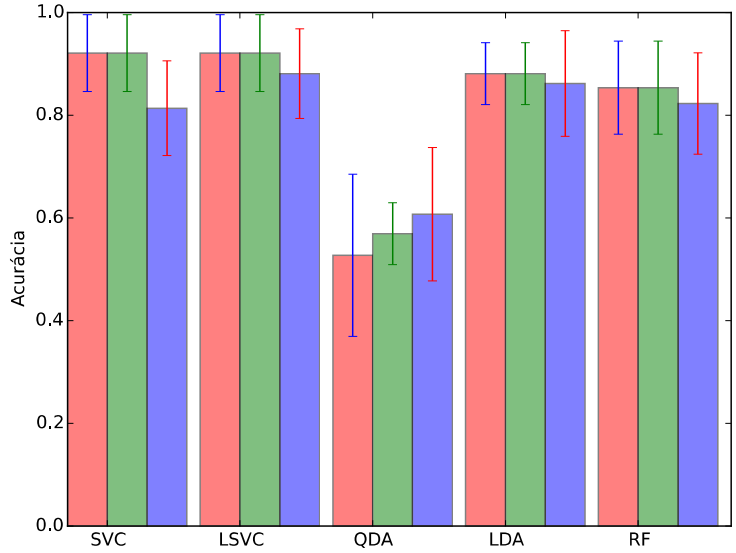
Neste estudo utilizamos a biblioteca de aprendizagem de máquina Sckit-Learn (Pedregosa et al.,

2011) para comparar o desempenho de algorítmos classificadores utilizando diferentes técnicas de redução de dimensionalidade. Mesmo sendo um problema de classificação binário, foram utilizados somente algorítmos capazes de lidar com múltipas classes devido ao desejo de avançar os estudos para discriminação de diferentes subtipos de câncer.

Métodos e Resultados

Para termos uma linha de base para comparação do impacto das técnicas de redução de dimensionalidade nos algorítmos classificadores, fizemos uma classificação inicial ingênua, utilizando todos os genes. Apesar de amostras de microarray obtidas por um mesmo protocolo serem naturalmente padronizadas, estando todas as variáveis numa mesma escala, também comparamos a acurácia entre os dados não tratados, dados repadronizados pelo escalonamento z-score, e dados normalizados. Curiosamente o uso de dados repadronizados implicou melhor acurácia na maioria dos classificadores. Daqui para frente, todas as análises ocorrem sobre dados repadronizados. Devido ao desbalanceamento da base de dados, o peso de cada classe foi considerado nos classificadores sensíveis a esta informação.

Gráfico 2: Acurácia de diferentes classificadores usando dados não tratados (vermelho), padronizados com z-score (verde) e normalizado (azul).

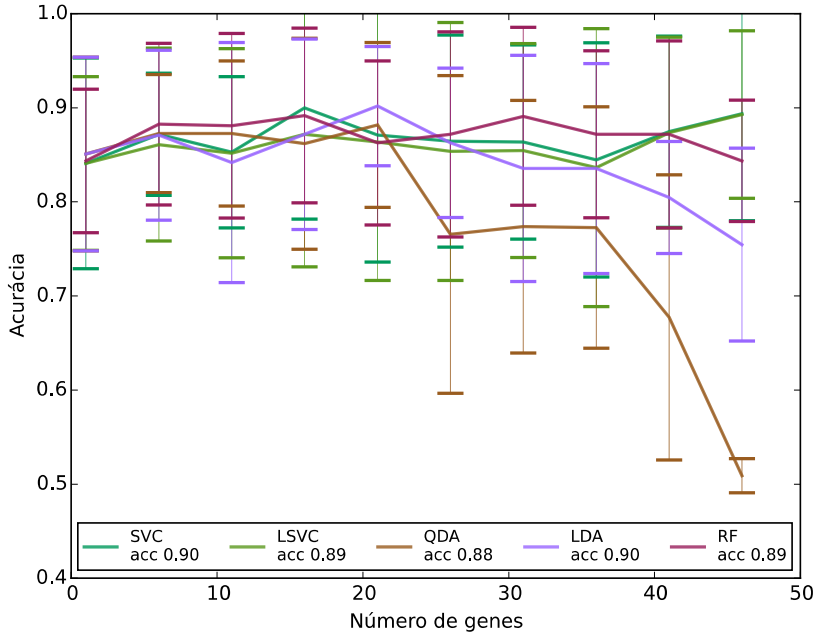


Florestas de Decisão Aleatória

Florestas de decisão aleatória (Random Forests, RF) apresentam diversas características que as tornam úteis para classificação e seleção de características em dados de microarray, incluindo a possibilidade de sere utilizada em casos $N < p$ (Díaz-Uriarte & Alvarez de Andrés, 2006). Funcionam pela construção de um conjunto árvores de decisão aleatória e o resultado da classificação é dado pela moda dos resultados das árvores. A importância de cada característica numa floresta é dada pela média da importância das características nas árvores. Utilizamos estas importâncias para selecionar genes relevantes, reduzindo a dimensionalidade dos dados e tornando-os mais adequados para alimentar classificadores que se comportam melhor quando $p < N$. No Gráfico 2 notamos que todos os classificadores apresentaram desempenho similar, exceto o LSVC,

cuja acurácia, considerando o erro, foi inferior em quase todas as etapas.

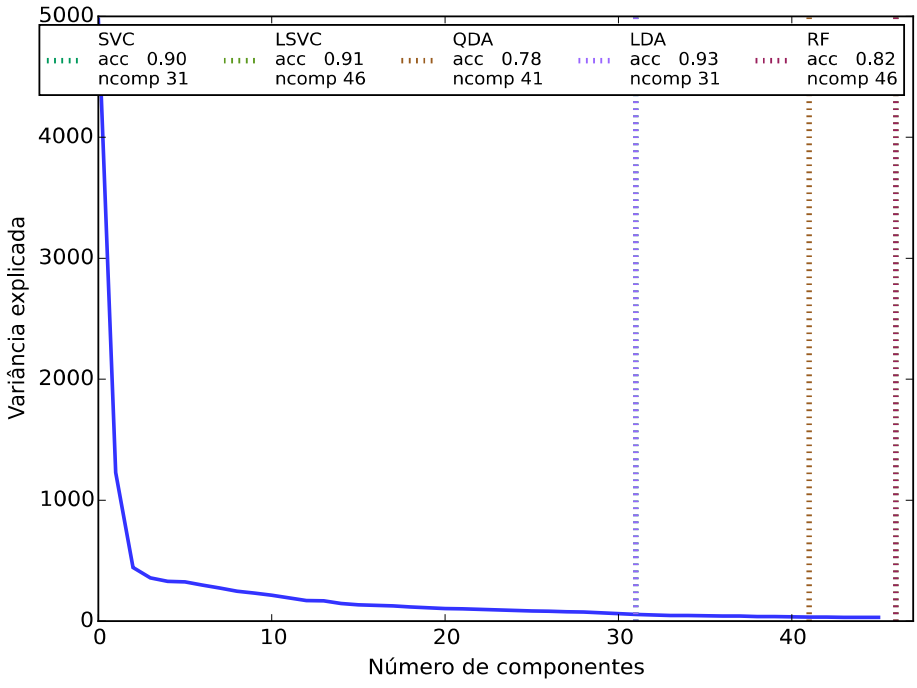
Gráfico 3: Comparação da acurácia utilizando seleção incremental para frente considerando a importância dada pela floresta.



Análise de Componentes Principais

Outra técnica para redução de dimensionalidade analisada foi a Análise de Componentes Principais (Principal Component Analysis, PCA), popular em dados de expressão gênica, na qual é feita uma busca pela combinação linear entre os atributos que maximizam a variância ao longo das amostras. A PCA projeta variáveis correlacionadas em novas variáveis não correlacionadas chamadas de componentes, de modo a remover redundâncias lineares. É dito que os componentes principais explicam, com alguma perda, a variância dos dados originais. Tem como desvantagem ignorar

Gráfico 4: O uso dos componentes principais não apresentou melhorias significativas no desempenho , inclusive piorando a acurácia de alguns classificadores.



as classes reais das amostras. (Elizondo, Passow, Birkenhead, & Huemer, 2008; Nguyen & Rocke, 2002)Observamos no Gráfico 4 o impacto do uso de até $p < N$ componentes, e a baixa significância na melhoria do desempenho classificatório. A piora da acurácia da RF é justificada pela perda de genes que, apesar de pouco importantes, contribuíam nas decisões. QDA continua um mistério.

Análise de Variância

No Gráfico 5 utilizamos a Análise de Variância (ANOVA) com o método F, que escolhe os atributos de modo a aumentar a variância intergrupos e diminuir a variância intragrupos, facilitando a discriminação. Apesar da ANOVA ser sensível a conjuntos de dados desbalanceados, como os que utilizamos, mostra-se bastante robusta a tal situação quado combinada com o método F. Outras condições que também invalidam a técnica não foram avaliadas.(Dell_Inc, 1995) Foi observado um aumento significativo da acurácia em todos os classificadores em relação à seleção pela Árvore Aleatória de Decisão.

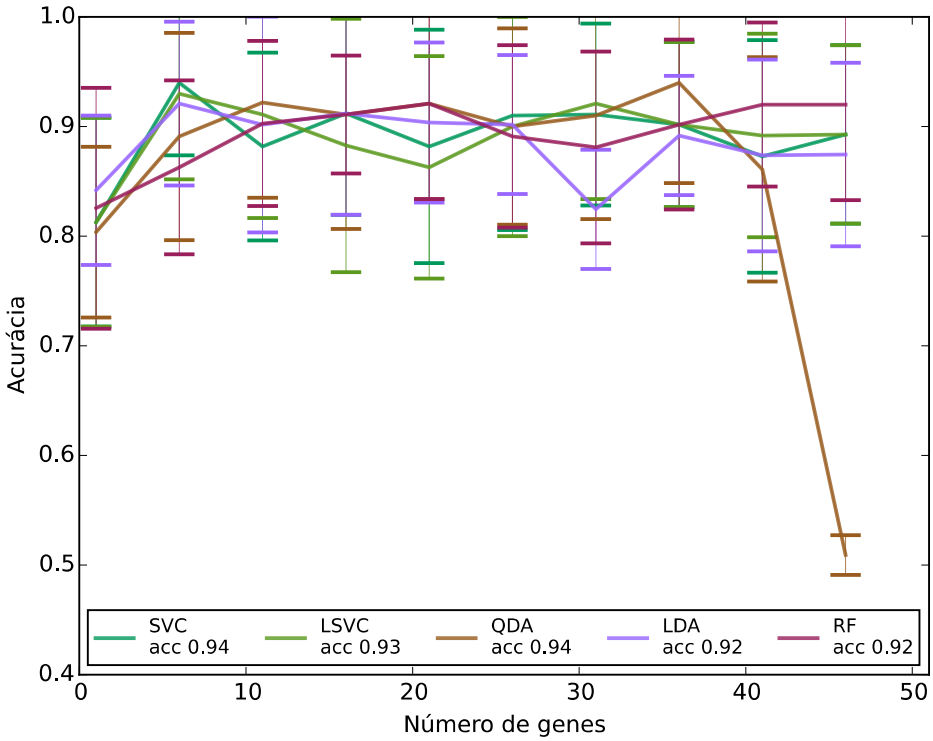
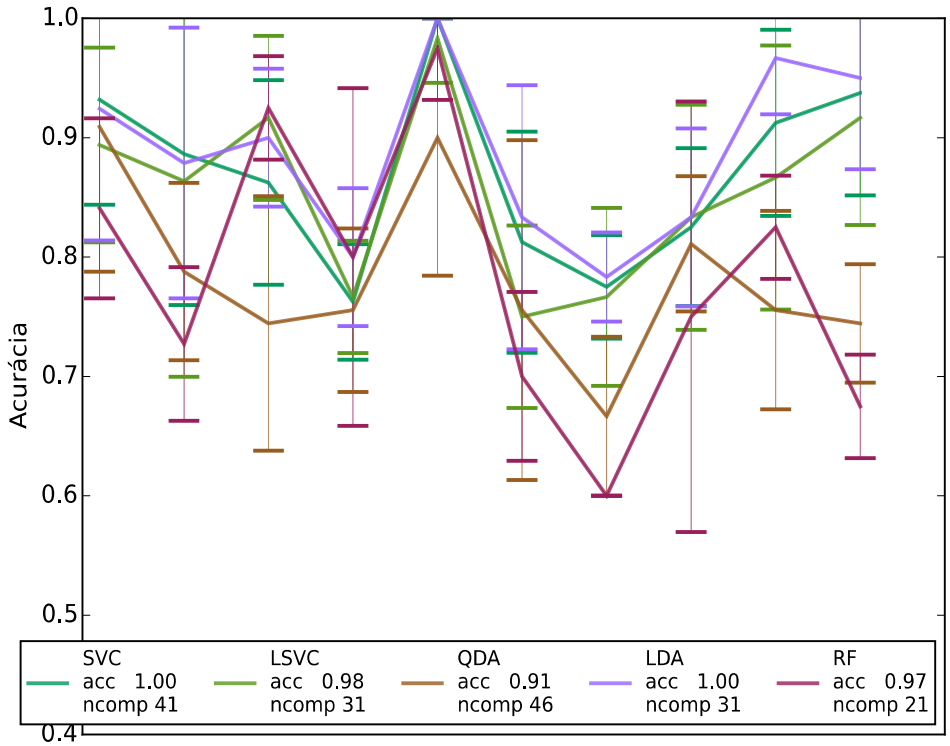


Gráfico 5: O uso da ANOVA causou uma melhoria significativa na predição das classes

Análise de Variância seguida por Análise de Componentes Principais

Também experimentamos a combinação das duas técnicas anteriores para redução de dados de expressão gênica. A eliminação de variáveis pela ANOVA acentua a variância intergrupos facilitando a construção de componentes pela PCA. Foi feita uma busca em grade pela combinação de até $q < N$ melhores atributos segundo a ANOVA, e até $r < q$ melhores componentes segundo a PCA, isto é, duas reduções sucessivas. Podemos verificar no Gráfico 6 um pico próximo a 100% na acurácia da maioria dos classificadores, mas infelizmente não conseguimos identificar os genes cuja combinação linear produz tal efeito por não estarmos operando sobre o espaço de atributos reais, mas de componentes.

Gráfico 6: Neste espaço de busca observamos um pico no desempenho classificatório com baixa margem de erro.



Conclusões

A redução de dimensionalidade melhorou significativamente a acurácia de todos os algorítimos de classificação. Curiosamente SVC (com kernel linear) e LSVC, que são duas implementações do mesmo algorítimo, apresentaram desempenhos diferentes em todos as análises, indicando a indissociabilidade do resultado e sua implementação. Efeito semelhante é observado com o LDA, que pode ser considerado um caso específico de QDA porque equações lineares são um caso específico de equações quadráticas.

A base de dados utilizada neste artigo já foi exaustivamente estudada na literatura, inclusive por diversos aritgos citados. Uma perspectiva futura é utilizar recentes dados de tumores de ovário (junho/2015, código de acesso GSE66957) ainda não publicados na literatura. Outro afazer é identificar os genes responsáveis pelo pico no Gráfico 5, na esperança de auxiliar a identificar as origens moleculares do câncer em questão. O uso de uma matriz de confusão também se faz necessário para identificar os casos onde modelo classificatório falha.

Não foi realizada uma busca pelos hiperparâmetros dos métodos SVC e LSVC, sugerindo ser possível obter melhores resultados nas análises onde não apresentaram desempenho satisfatório. Outras abordagens interessantes para redução da dimensionalidade são o ranqueamento das variáveis por um método baseado em estatística T e construções de componentes pelo método de Mínimos Quadrados Parciais (Partial Least Squares, PLS) (Nguyen & Rocke, 2002).

Um novo e promissor algorítimo baseado no PLS e relevante para classificação de dados de expressão gênica é descrito por Chung & Keles (2010), mas não foi possível utilzá-lo devido à ainda não ter sido implementado na biblioteca de aprendizagem de máquina utilizada.

Apesar de todos as análises terem sido validadas em 10 pastas, também faz-se necessário de separar

uma parte do conjunto de dados para avaliação externa da qualidade dos modelos gerados.

Referências

Ai-Jun, Y., & Xin-Yuan, S. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, *26*(2), 215–222. http://doi.org/10.1093/bioinformatics/btp638

Chung, D., & Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, *9*(1).

Dell_Inc. (1995). *ANOVA / MANOVA - Statistics Textbook*. Retrieved from http://documents.software.dell.com/Statistics/Textbook/ANOVA-MANOVA

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 1–13. http://doi.org/10.1186/1471-2105-7-3

Ding, C., & Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE* (pp. 523–528). http://doi.org/10.1109/CSB.2003.1227396

Elizondo, D., Passow, B., Birkenhead, R., & Huemer, A. (2008). Dimensionality Reduction and Microarray Data. In A. Gorban, B. Kégl, D. Wunsch, & A. Zinovyev (Eds.), *Principal Manifolds for Data Visualization and Dimension Reduction* (Vol. 58, pp. 293–308). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-73750-6_13

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … others. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537.

Nguyen, D. V., & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, *18*(1), 39–50. http://doi.org/10.1093/bioinformatics/18.1.39

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.