Seleção de genes

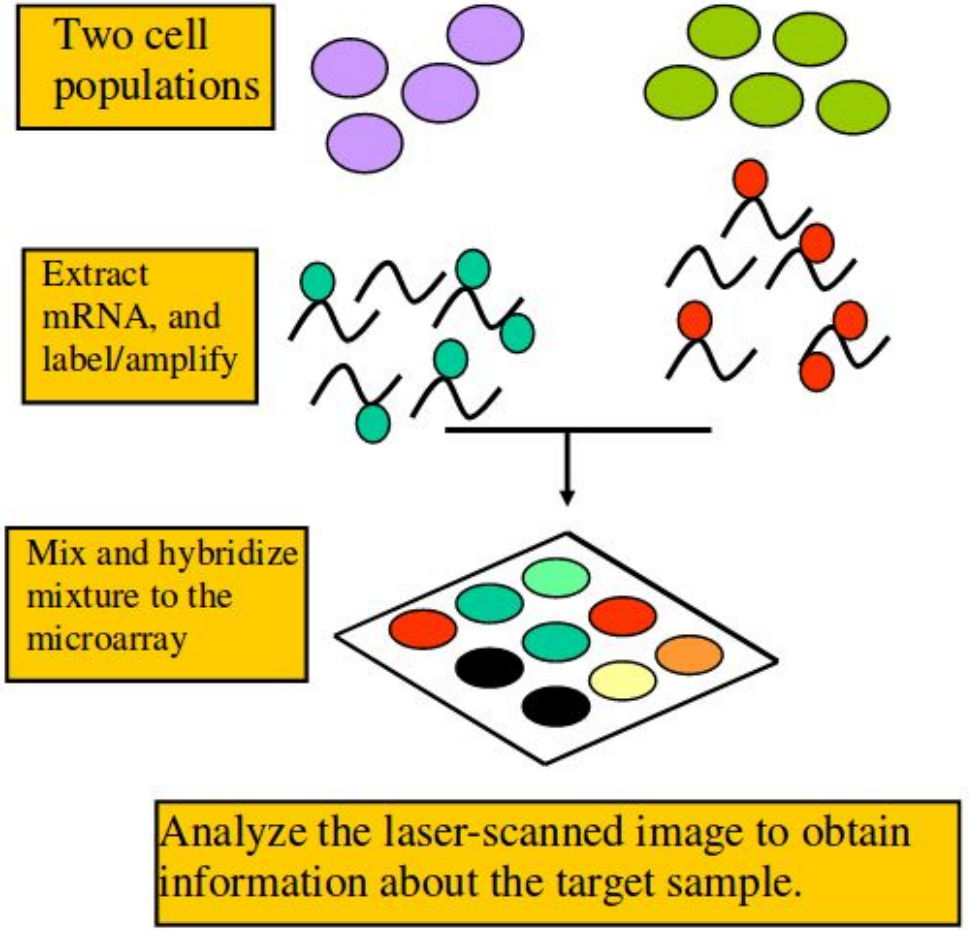
Bruno Henrique Meyer

O que será apresentado

- Expressão gênica
- Classificadores
- Seleção de atributos
- Resultados e comparações de trabalhos

Expressão gênica

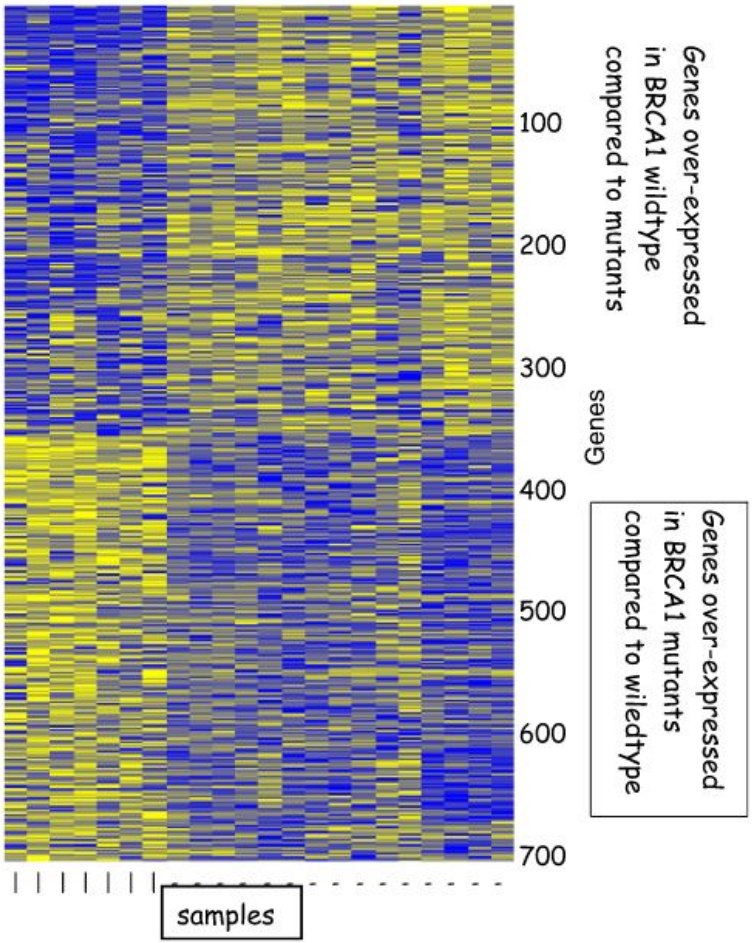
Microarray



Fonte: (NACHTOMY, SHAVIT e YAKHINI, 2007), Fig. 1.

Expressão gênica

Microarray



Fonte: (NACHTOMY, SHAVIT e YAKHINI, 2007), Fig. 2.

Expressão gênica

Microarray

- Poucas amostras
- Muitos genes

Samples-->	T1	T2	T3	T4	T5	T6	T7	N1	N2	N3	N4	*P* value
Genes	Expression level relative to non-tumor pool											
Gene 1	2.4		2	2.5	1.5	2	2.3	1.7	1	1	0.81	9.5E-06
Gene 2	2.9	3.2	1.4	1.7	2.6	3.7	2.5	1	0.9	0.7	0.7	8.0E-05
Gene 3	2	3.5	1.4	1.8	2	2.5	2.2	0.7	1	0.7	0.6	8.6E-05
Gene 4	2.2	2.2	2.7	1.3	2.3	3.7	1.7	0.7	0.9	0.9	0.9	0.00012
Gene 5	5.2	2	3.7	2	5.8	3.2	1.6	0.7	0.7	0.7	0.6	0.00014
Gene 6	6.9	15	18	5.8	12	21	2.3	0.9	1.6	0.7	1.2	0.00015
Gene 7	5.4	2.1	3	2.2	3.5	2.8	1.5	0.8	1.2	0.9	0.8	0.00022
Gene 8	3.1	2.3	1.8	1.7	2.9	1.5	1.2	0.7	0.7	1	0.7	0.00023
Gene 9	9.7	25	23	6.1	9.5	23	2.4	1.2	1.6	0.8	1.1	0.00024
Gene 10	7.6	14	13	4.7	8.2	24	2.1	0.9	1	0.8	1.1	0.00025
Gene 11	4.8	7.7	2.1	2.3	6.6	3.7	7.4	1.1	1.2	0.6	1.2	0.00028
Gene 12	3.6	5.7	3.8	3.3	4.7	6.6	1.8	1.7	0.9	1.1	1	0.00029
Gene 13	5.7	9.8	12	4.5	6	17	1.7	1	1.3	0.8	0.8	0.00031
Gene 14	1.5	2.1	1	1.1	1.2	1.2	1.4	0.6	0.8	0.6	0.8	0.00031
Gene 15	2.5	2.9	1.9	1.8	5.5	2	1.3	1	0.8	0.7	0.7	0.0004
Gene 16	2.2	1.5	1.3	1.2	1.4	1.8	1.1	0.8	0.8	0.8	0.8	0.00042
Gene 17	5.9	2	3.4	2.5	4.3	3.1	2.1	1.2	1.3	1.5	1	0.00048
Gene 18	4	1.6	2.8	1.4	2.9	2.2	1.6	0.9	0.9	1	0.8	0.00052
Gene 19	1.6	1.5	2.3	1.4	1.4	1.8	1.6	1.1	1.2	1	1.1	0.00059
Gene 20	3.9	6.7	6.6	2.3	4	11	1.5	0.8	1.1	0.8	0.8	0.00059
Gene 21	5.3	1.8	2.6	1.4	3.4	2.2	1.6	0.7	0.8	0.8	0.8	0.0006
Gene 22	4.2	1.9	1	2	4.2	4.3	7.9	16	1.1	1.3	10.9	0.00061
Gene 23	2.3	1.3	2.5	1.8	5.7	2.2	1.8	1.1	0.7	0.6	0.7	0.00066
Gene 24	2.9	1	2.9	2.2	3.8	1.9	2.3	0.9	0.8	0.9	0.8	0.00071
Gene 25	2.6	1.4	1.7	1.4	2.4	1.7	1.3	0.9	0.9	0.9	0.9	0.00079
Gene 26	5.8	2.3	3.4	2.1	5	4.7	1.7	1.2	0.9	1.2	0.6	0.0009
Gene 27	5.7	2	4	2.4	5	3.5	1.5	0.8	1.2	1.3	1.1	0.00093
Gene 28	1.6	2.8	1.7	1.7	1.5	2.8	1.9	1.2	1.1	0.8	1.1	0.00094

trends in Biotechnology

Fonte: (ZWEIGER, 1999), Fig. 3.

Classificadores

Árvore de decisão

- SVM
- Random Forest
- O objetivo **não** é necessariamente classificar

Seleção de atributos

- Seleção de genes
- Classificação
- Quais atributos escolher?
  - Como?

Gene 1	Gene 2	Gene 3	Gene 4	...	Classe
1.0	1.0	0.15	0.0	...	1
0.91	0.98	0.15	0.0	...	1
0.4	0.7	0.69	0.0	...	2
0.37	0.71	0.65	0.0	...	2
...	...	...	...	...	...

Seleção de atributos

1ª Tentativa

Testar todas combinações

Para 3 genes:

(Gene1)

(Gene2)

(Gene3)

(Gene1, Gene2)

(Gene2, Gene3)

(Gene1, Gene3)

(Gene1, Gene2, Gene3)

O(2^n)

Seleção de atributos

2ª Tentativa

Não explorar todas as possibilidades

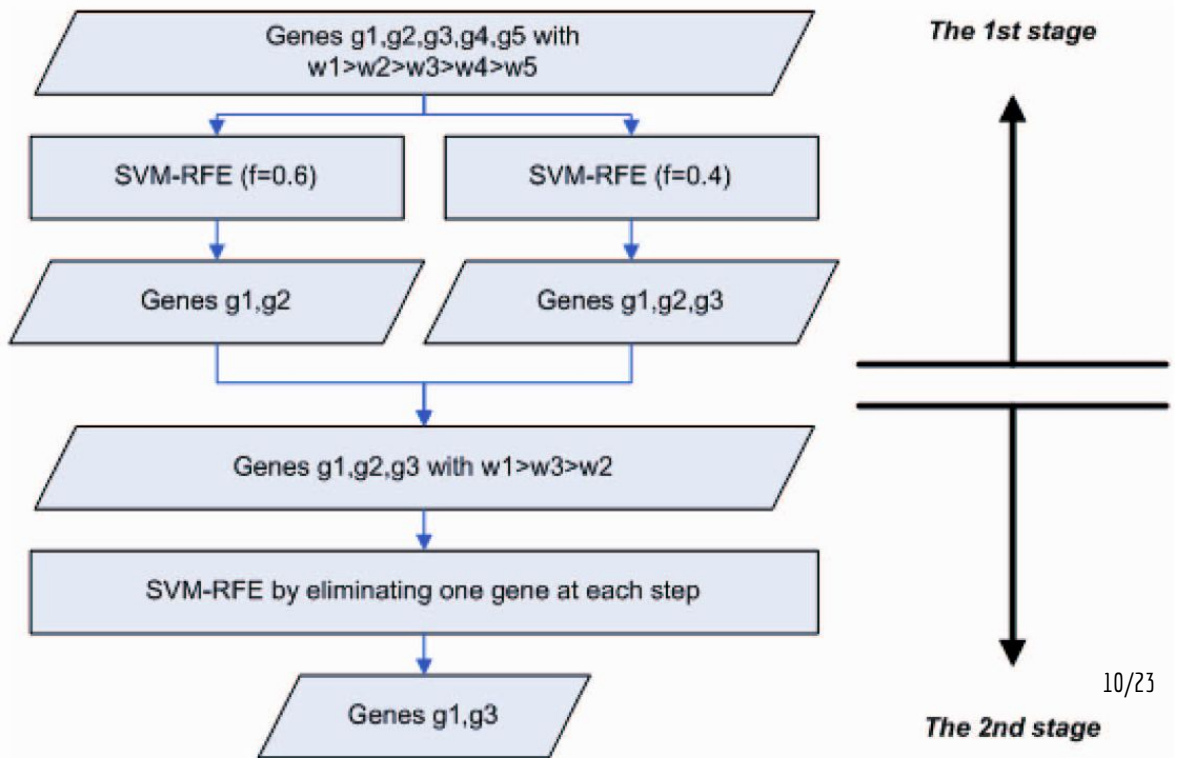
- *Filter*
  - Rápido
  - Ignora o classificador
  - Univariado ou Multivariado
- Wrapper
  - Lento
  - Classificador é uma caixa preta
  - Aleatório ou determinístico
- *Embedded*
  - Olha dentro do mecanismo do classificador

(SAEYS, INZA e LARRAÑAGA, 2007)

Seleção de atributos

SVM-RFE (SVM Recursive Feature Elimination)

1. *Treinar um SVM no conjunto de treino*
2. *Ordenar os genes de acordo com os pesos do classificador*
3. *Eliminar o gene com menor significância*
4. *Repetir o processo até que não exista mais nenhum gene no conjunto de treino*

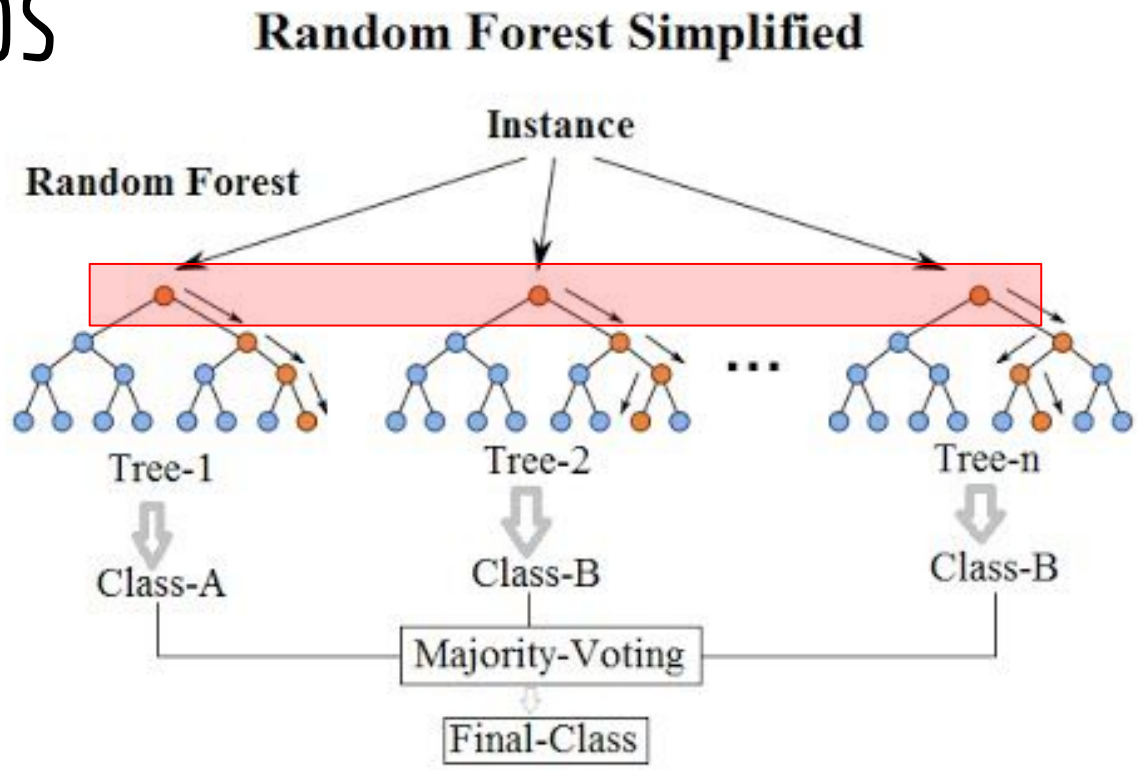


Fonte: (TANG, ZHANG e HUANG, 2007), Fig. 3.

Seleção de atributos

Random Forest

- Olhar para a raiz e nodos rasos
- Observar o desempenho da seleção de característica interna de cada árvore



Fonte: https://www.youtube.com/watch?v=loNcrMjYh64

Seleção de atributos

Qual a diferença?

- Qual a relação entre os desempenhos dos seletores?
  - Tempo
  - Eficiência
- O resultado final é semelhante?

- *Filter*
  - Rápido
  - Ignora o classificador
  - Univariado ou Multivariado
- Wrapper
  - Lento
  - Classificador é uma caixa preta
  - Aleatório ou determinístico
- *Embedded*
  - Olha dentro do mecanismo do classificador

Seleção de atributos

Metodologia

- Linguagem de programação python
  - DEAP
  - *scikit-learn*
  - Código fonte disponível em: https://github.com/BrunoMeyer/gene-selection-to-classification

Seleção de atributos

Metodologia - Métricas utilizadas

- 7 Classificadores
  - AUC (média das classes)
  - Cross-Validation com 5 *folds*
  - MLP, SVM, Random Forest, KNN, GNB, BNB, MNB
- Tempo de processamento (CPU)
  - Apenas para a seleção

Seleção de atributos

Metodologia

- 4 métodos de seleção
  - *Variance Threshold* (Filter)
  - Algoritmo Genético (Wrapper)
    - tamanho da população: 70
    - número de gerações: 2000
    - taxa de mutação: 5%
    - taxa de cruzamento: 30%
    - SVM, RDF, MNB
  - RFE (Embedded)
    - SVM, MNB
    - tamanho de passo: 1
    - tamanho total: 150
  - Random Forest (Embedded)

Seleção de atributos

Materiais

- Mesmas bases de dados que STATNIKOV, WANG e ALIFERIS
- Disponível em http://www.gems-system.org/
- 3 bases de dados utilizadas
  - *5 human brain tumor types*
    - 5921 genes; 90 instâncias; 5 classes
  - *4 malignant glioma types*
    - 10368 genes; 50 instâncias; 4 classes
  - *Prostate tumor and normal tissues*
    - 10510 genes; 102 instâncias; 2 classes

Resultados

AUC Média - *5 human brain tumor types* (5921 genes)

	Todos	Alg. Genético SVM	Alg. Genético RDF	Alg. Genético MNB	RFE SVM	RFE MNB	RF	VAR
MLP	0,826	0,835	0,850	0,879	0,969	0,931	0,890	0,826
RDF	0,850	0,836	0,877	0,870	0,943	0,898	**0,935**	0,850
SVM	**0,895**	**0,908**	0,884	0,898	**0,988**	0,901	0,895	0,892
KNN	0,791	0,794	0,794	0,794	0,819	0,738	0,791	0,791
GNB	0,739	0,745	0,759	0,750	0,789	0,838	0,859	0,739
BNN	0,519	0,512	0,511	0,521	0,576	0,559	0,453	0,519
MNB	0,893	0,903	**0,894**	**0,917**	0,986	**0,931**	0,931	**0,893**

Tabela 1: Área sobre a curva ROC obtida por meio da medida dos resultados dos classificadores (linhas) em relação aos seletores (colunas). Cada coluna tem o seu maior valor em negrito, e o maior resultado da tabela está sublinhado. Os resultados são referentes à primeira base de dados utilizada.

Resultados

AUC Média - *4 malignant glioma types* (10368 genes)

	Todos	Alg. Genético SVM	Alg. Genético RDF	Alg. Genético MNB	RFE SVM	RFE MNB	RF	VAR
MLP	0,865	0,749	0,872	0,892	**0,975**	0,826	0,952	0,865
RDF	**0,906**	0,872	**0,947**	0,910	0,959	**0,873**	**0,954**	**0,906**
SVM	0,812	0,776	0,829	0,770	0,882	0,661	0,881	0,833
KNN	0,777	0,776	0,765	0,801	0,962	0,621	0,790	0,777
GNB	0,837	0,854	0,828	0,839	0,938	0,807	0,902	0,837
BNN	0,475	0,456	0,439	0,399	0,344	0,439	0,423	0,475
MNB	0,906	**0,899**	0,898	**0,918**	0,887	0,628	0,941	0,906

Tabela 2: Área sobre a curva ROC obtida por meio da medida dos resultados dos classificadores (linhas) em relação aos seletores (colunas). Cada coluna tem o seu maior valor em negrito, e o maior resultado da tabela está sublinhado. Os resultados são referentes à segunda base de dados utilizada.

Resultados

AUC Média - *Prostate tumor and normal tissues* (10510 genes)

	Todos	Alg. Genético SVM	Alg. Genético RDF	Alg. Genético MNB	RFE SVM	RFE MNB	RF	VAR
MLP	0,916	0,914	**0,939**	**0,931**	**0,996**	0,900	0,933	0,916
RDF	**0,929**	**0,924**	0,936	0,929	0,976	**0,921**	**0,959**	**0,929**
SVM	0,474	0,357	0,569	0,577	0,994	0,300	0,899	0,563
KNN	0,805	0,756	0,825	0,805	0,952	0,775	0,813	0,805
GNB	0,616	0,623	0,639	0,646	0,969	0,740	0,890	0,616
BNN	0,602	0,630	0,672	0,627	0,502	0,595	0,531	0,602
MNB	0,729	0,734	0,730	0,742	0,988	0,890	0,888	0,729

Tabela 3: Área sobre a curva ROC obtida por meio da medida dos resultados dos classificadores (linhas) em relação aos seletores (colunas). Cada coluna tem o seu maior valor em negrito, e o maior resultado da tabela está sublinhado. Os resultados são referentes à terceira base de dados utilizada.

Resultados

Custo computacional

	Alg. Genético SVM	Alg. Genético RDF	Alg. Genético MNB	RFE SVM	RFE MNB	RF	VAR
5 human	9503,124	3856,288	3439,869	862,370	28,702	0,002	0,006
4 malignant	7189,514	3082,493	1417,126	959,327	59,896	0,002	0,007
prostate	**158074,374**	**36417,890**	**11146,599**	**2659,137**	**80,217**	**0,004**	**0,015**

Tabela 4: Tempo (segundos) de execução em CPU para cada base de dados utilizada no trabalho. Cada coluna tem o seu maior valor em negrito, e o maior resultado da tabela está sublinhado.

Resultados

Avaliação dos resultados

- SVM não terá, necessariamente, o melhor desempenho como classificador
- SVM-RFE apresentou melhores resultados
- O tempo de execução depende do número de instâncias e do número de genes
- MNB+Genético

Conclusão

Seleção de genes

- Variedade
- O classificador pode não ser relevante
- Custo computacional
- Escalabilidade

Discussão

Seleção de genes

- Variação nos parâmetros
  - RFE
  - Alg. Genético
- Regressão e Clusterização
- A interseção dos resultados é interessante?
- Métodos direcionados para a seleção
  - Existe algum que funciona bem?
- Custo
  - Se é polinomial, então é bom?

Dúvidas?

Referências

[1] NACHTOMY, Ohad; SHAVIT, Ayelet; YAKHINI, Zohar. Gene expression and the concept of the phenotype. Studies in History and Philosophy of Science Part C: **Studies in History and Philosophy of Biological and Biomedical Sciences**, v. 38, n. 1, p. 238-254, 2007.

[2] ZWEIGER, Gary. Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. **Trends in biotechnology**, v. 17, n. 11, p. 429-436, 1999.

[3] TOUW, Wouter G. et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. **Briefings in bioinformatics**, v. 14, n. 3, p. 315-326, 2012.

[4] SAEYS, Yvan; INZA, Iñaki; LARRAÑAGA, Pedro. A review of feature selection techniques in bioinformatics. **bioinformatics**, v. 23, n. 19, p. 2507-2517, 2007.

Referências

[5] TANG, Yuchun; ZHANG, Yan-Qing; HUANG, Zhen. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 4, n. 3, p. 365-381, 2007.

[6] STATNIKOV, Alexander; WANG, Lily; ALIFERIS, Constantin F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. **BMC bioinformatics**, v. 9, n. 1, p. 319, 2008.

[7] DUAN, Kai-Bo et al. Multiple SVM-RFE for gene selection in cancer classification with expression data. **IEEE transactions on nanobioscience**, v. 4, n. 3, p. 228-234, 2005.

[8] DENG, Houtao; RUNGER, George. Gene selection with guided regularized random forest. Pattern Recognition, v. 46, n. 12, p. 3483-3489, 2013.