Seleção de características em expressão gênica

Bruno Henrique Meyer
Universidade Federal do Paraná
Informática Biomédica
Email: bhm15@inf.ufpr.br

Keywords—Bioinformática, Seleção de genes

I. Introdução

Classificação é um problema da computação que é explorado atualmente em diversas aplicações e situações. Pode-se utilizar ferramentas computacionais para a predição de câncer, estimar genes relacionados a determinadas doenças, entre outros. Para que essas aplicações possuam um bom resultado, é necessário um conjunto relativamente grande de base de dados.

Na área da bioinformática, a ciência dos "omics"é uma fonte considerável de geração de dados, assim como pode ser visto em *genomics*, *proteomics*, entre outras [10]. A análise desses dados pode muitas vezes gerar informações e conhecimentos ou confirmar teorias e sensos comuns (ainda não provados ou confirmados estatisticamente) como a predição da relevância da ontologia de genes e interações de componentes químicos na predição da expectativa de vida [1].

O uso de algoritmos de aprendizado de máquina supervisionados podem servir de apoio a suposições devido aos seus embasamentos estatísticos e ao fato de que os mesmos possam provir um modelo interpretável para a identificação das principais características em um problema de classificação. Entretanto, o subconjunto das expressões escolhidas que é usado na classificação causa grande influência no resultado dos classificadores. Métodos de seleção de características em problemas de classificação de expressão gênica são amplamente utilizados para melhorar os resultados e evitar *overfitting* (especifidade da solução para apenas os dados utilizados) em alguns casos [9].

Como ilustra a figura 1 a seguir, a expressão de genes em um estudo pode ser representada em um formato de matriz. Ao analisar essas matrizes, podemos observar correlações entre expressões gênicas devido ao processo natural de casos onde genes fazem parte de uma mesma via metabólica e/ou fazem parte de um mecanismo de regulação ou expressão gênica. Há casos onde se considera até 200 genes em um único estudo [9], o que pode acabar influenciando na etapa de classificação e assim, fazendo necessário o processo de seleção de genes.

Quais são as diferenças entre as abordagens para seleção de genes para a classificação de expressão gênica? Este trabalho tem por objetivo mensurar a eficiência e custo computacional de diferentes algoritmos presentes na literatura para a resolução do problema de seleção de genes.

II. Tipos de técnicas para seleção de genes

Pode-se separar os métodos de seleção de características dentro do estudo de expressão gênica em três classes [6]:

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

Figura 1: Example of a matrix of gene-expression results [11, Figure 3]

.

- ***Filter***: São consideradas técnicas de *filtering* aquelas que são independentes do classificador utilizado no estudo. Sua principal ideia é analisar apenas os dados de cada expressão de gene e criar uma valoração que representará o desempenho de tal característica na discriminação das classes envolvidas, como o uso dos princípios Bayesianos. Há dois principais de algoritmos dentro da técnica *Filter*: **Univariados** e **Multivariados**. Os multivariados representam os algoritmos que consideram a interação entre os genes, o que não acontece nos univariados. A principal vantagem das técnicas de *filtering* é que sua complexidade é bem escalada em relação ao total de genes considerado no problema. Porém, sua principal desvantagem é que por não se considerar nenhum classificador durante a seleção, não se sabe qual será o real impacto da seleção no processo de classificação.
- ***Wrapper***: São técnicas onde há a dependência de ao menos um classificador. Cada subconjunto presente no espaço total de características do problema é considerado como possibilidade do conjunto selecionado ao final do algoritmo. É fácil de ver que o tamanho desse conjunto cresce exponencialmente em relação ao total de genes considerado no estudo, o que implica na necessidade da elaboração de heurísticas para a redução da complexidade computacional que podem ser **determinísticas** ou **aleatorizadas**. Nos métodos aleatórios, algoritmos genéticos por exemplo, os subespaços são

gerados aleatoriamente e posteriormente são avaliados em relação aos seus desempenhos em um classificador. Os métodos determinísticos partem da mesma ideia de algoritmos de retro substituição, onde há um direcionamento ou convergência para a solução. Um exemplo dentro dos métodos determinísticos é o algoritmo *Sequential backward elimination*. A principal vantagem dessas técnicas é que há a consideração da influência do desempenho do algoritmo de classificação nos conjuntos selecionados. Porém, é necessário um grande custo computacional para que isso seja satisfeito e também há o risco de que a solução seja específica de mais para a base de treinamento, podendo causar problemas de *overfitting*.

- ***Embedded***: Semelhante à classe *Wrapper*. As técnicas *embeddeds* também dependem de um classificador em suas definições porém, neste caso o subespaço de características escolhidas será decidido por estruturas internas dos classificadores. Um exemplo de algoritmo considerado dentro desta classe é a arvore de decisão. Suas vantagens e desvantagens são semelhantes às abordagens *wrappers*, mas o custo computacional é reduzido pois não é necessário construir um classificador para cada subespaço de genes encontrado. Como desvantagem, deve-se considerar que a solução é dependente de um único classificador, que criou os critérios de seleção de genes.

III. Classificadores

Nesta pesquisa será abordado dois dos diversos classificadores presentes na literatura: *Random forest* e *Suport Vector Machine* (SVM).

A escolha desses classificadores se deu devido ao alto relato de seus usos em diversos trabalhos da área.

A. Random forest

Random forest (RF) é um algoritmo de classificação, baseado no uso do método *ensemble* em diversas instâncias de árvores de decisão. Assim como ilustra a figura abaixo, as RFs são geradas aleatoriamente e se baseiam em variedade e não dependem de muitos parâmetros para serem ”tunadas”durante a etapa de treinamento, diferentemente do SVM que será comentado posteriormente.

Uma característica muito interessante que deve ser ressaltada no algoritmo RF é que por meio modelo gerado após o treinamento, pode-se consultar as camadas mais baixas das árvores criadas e podadas. Ao consultar essas camadas, podemos obter os critérios de decisões mais comuns que foram modeladas, ou seja, as características mais descriminantes para se classificar ou predizer no contexto onde o mesmo foi treinado.

Outras vantagens do classificador *random forest* são [3]:

- Pode-se usar tanto para problemas de classificação binária como problemas multiclasses.
- Possui bom desempenho quando há ”ruídos”em algumas características das instâncias.
- Não entra em *overfitting*.

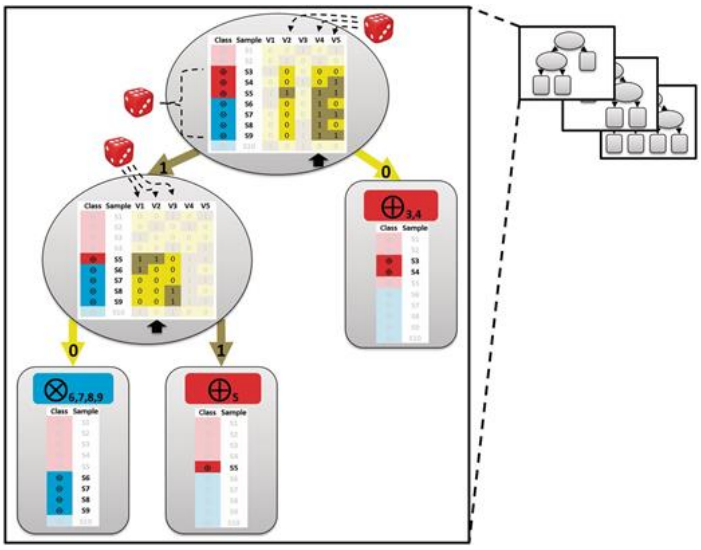

Figura 2: Treino de uma única árvore em uma modelo de Random Forest. [10, Figure 2]

.

- Incorpora a relação entre características determinantes para classificação.
- Pode retornar a quantificação da importância de uma característica para a predição de problemas.
- Pode ser ”tunado”para aumentar seu desempenho (como o número de árvores e profundidade máxima das árvores por exemplo).

B. SVM

O SVM (Support Vector Machines) é um classificador baseado na aproximação de uma fronteira com base em vetores de suporte adquiridos da base de dados. Seu uso consiste principalmente em aplicações de regressão linear e aprendizado de máquina supervisionado. Diversas técnicas são empregadas junto ao SVM, aumentando sua eficácia.

Devido à sua sensibilidade à dimensionalidade das características de uma base de dados o SVM pode ser utilizado para seleção de características e redução da dimensionalidade em problemas de classificação, como o algoritmo SVM-RFE [5].

IV. SVM-RFE como seleção de genes

O algoritmo *SVM Recursive Feature Elimination* (SVM-RFE) se baseia no ranqueamento e ordenação do impacto das características em um problema de classificação. A ideia base do algoritmo é a quantificação da influência de um gene na classificação por meio de uma instância de SVM, onde considera-se principalmente o peso atribuído para cada gene no modelo construído para um determinado subconjunto. As etapas do algoritmo SVM-RFE consistem em:

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

O algoritmo pode ser realizado até atingir um determinado critério, como um limiar para a significância dos genes ou uma quantidade mínima de atributos selecionados. Com isso, teremos como resultado um subconjunto de genes do conjunto inicial de forma que seja otimizado o desempenho do classificador em futuras instâncias do problema.

V. Revisão de literatura

No trabalho [4], vemos a aplicação de diferentes variações do SVM como classificador e seletor de características (inclusive o SVM-RFE). Neste trabalho, vemos a comparação de três metodologias do uso do SVM para a classificação de câncer em diferentes bases de dados. As métodologias são: SVM sem seleção de característica; SVM-RFE; MSVM-RFE, que é uma variação do algoritmo SVM-RFE onde se aplica uma técnica de *resampling*, como *Cross Validation*, para explorar melhor o conceito de que diferentes subconjuntos do conjunto geral dos genes possuem bons resultados como parâmetros em classificações. Na Figura 3, os resultados obtidos no trabalho [4] são apresentados, apontando o número de genes selecionados e o impacto no resultado do classificador. Pode-se perceber que não há, nos resultados apresentados, nenhuma relação entre superioridade na questão de número de genes selecionados entre as diferentes técnicas. Porém, em relação ao desempenho do classificador há uma relação notável na taxa de erro (principalmente quando não se usa seleção de características). O MSVM-RFE apresentou melhores resultados após a etapa de classificação, porém, não há indícios se isso é causado por uma real melhora que a técnica propicia ao processo ou se aconteceu um caso de especificação para a base de estudo, acarretando em *overfitting*.

Dataset	Measurement	SVM	SVM-RFE	MSVM-RFE
Breast	♯ Genes	Full (24481)	81	161
	Test Error	35.26±9.71	3.95±4.57	**3.74±4.32**
	Sensitivity	58.75±13.42	94.58±6.94	**94.83±6.68**
	Specificity	75.00±17.27	98.57±4.31	**98.71±4.11**
Colon	♯ Genes	Full (2000)	7	3
	Test Error	18.30±6.86	7.90±4.78	**6.55±4.48**
	Sensitivity	74.00±16.67	**83.71±11.84**	83.57±12.08
	Specificity	85.85±7.63	96.62±5.82	**98.77±3.04**
Leukemia	♯ Genes	Full (12582)	95	37
	Test Error	12.91±6.37	3.76±4.35	**2.38±4.32**
	Sensitivity	99.40±1.92	**100.00±0.00**	**100.00±0.00**
	Specificity	69.50±14.64	90.86±10.55	**94.21±10.49**
Lung	♯ Genes	Full (12533)	31	33
	Test Error	0.48±0.78	**0.32±0.34**	**0.32±0.34**
	Sensitivity	96.80±7.37	**96.87±3.34**	**96.87±3.34**
	Specificity	99.83±0.38	**100.00±0.00**	**100.00±0.00**

Figura 3: Desempenho do algoritmo SVM-RFE como seletor de genes. [4, Table 2]

Em [2], é apresentado o algoritmo *Regularized Random Forest* (RRF), uma variação do algoritmo *Random Forest*, onde há uma alteração sutil na forma que as árvores são construídas dentro do algoritmo. Semelhante ao trabalho apresentado anteriormente, aqui são também comparadas diferentes variações de um mesmo algoritmo de seleção, cada um com suas particularidades. Nas figuras 4, 5 e 6 o autor demonstra a comparação dos resultados de cada seletor. Pode-se observar que nesse caso, há uma hierarquia do número de genes selecionados pelas variações dos seletores, o que não foi observado no trabalho [4]. Entretanto, a figura 6 demonstra que tal peculiaridade não afeta necessariamente no resultado da classificação. Um fato que pode ser destacado, é que a variação GRRF(0.1) apresentou melhores resultados na classificação (Figura 6), comparado com os outros seletores, nas bases *colon* e *nci*, onde foram exceções no número de características selecionadas (Figura 5) quando o GRRF(0.1) selecionou mais genes que os demais seletores.

	Number of features		Average error rates	
	RRF(1)	All	RRF(1)-RF	All-RF
adenocarcinoma	86	9868	0.158	0.159
brain	97	5597	0.159	0.170
breast.2.class	210	4869	0.352	0.371○
breast.3.class	253	4869	0.397	0.415○
colon	92	2000	0.162	0.158
leukemia	24	3051	0.053	0.064
lymphoma	31	4026	0.018	0.006•
nci	197	5244	0.332	0.321
prostate	88	6033	0.082	0.109○
srbct	51	2308	0.027	0.032

Figura 4: Comparação entre o uso do algoritmo Random Forest para classificação com e sem seleção de características por meio do algoritm RRF (Número de genes e Taxa de erro). [2, Table 3]

	All	GRRF(0.1)	GRRF(0.2)	RRF(0.9)	varSelRF	LASSO
adenocarcinoma	9868	20	15	23	4	2
brain	5597	22	12	27	28	24
breast.2.class	4869	59	25	60	7	7
breast.3.class	4869	77	31	78	12	18
colon	2000	29	13	27	4	7
leukemia	3051	6	4	6	2	8
lymphoma	4026	5	4	4	81	25
nci	5244	63	26	61	60	53
prostate	6033	18	13	19	6	12
srbct	2308	13	9	14	34	28

Figura 5: Número de genes selecionados por algoritmos derivados do RF. [2, Table 4]

No trabalho [8] foi realizada uma comparação direta entre o desempenho dos classificadores SVM e RF em base de dados de cânceres. O autor explicitou que seu objetivo não era necessariamente observar o impacto dos algoritmos de seleção de genes no resultado da classificação, mas sim a análise direta do desempenho dos classificadores, entretanto foi utilizado no estudo o uso de algoritmos baseados no RF, SVM-RFE e *Backward eliminations* para a seleção de atributos. As figuras 7 e 8 ilustram o desempenho dos classificadores em

	GRRF(0.1) -RF	GRRF (0.2) -RF	RRF (0.9) -RF	varSelRF -RF	LASSO -RF
adenocarcinoma	0.169	0.168	0.160	0.212 ○	0.189 ○
brain	0.214	0.259 ○	0.234	0.231	0.259 ○
breast.2.class	0.345	0.359 ○	0.367 ○	0.386 ○	0.366 ○
breast.3.class	0.387	0.403 ○	0.410 ○	0.418 ○	0.400
colon	0.175	0.186 ○	0.190 ○	0.232 ○	0.180
leukemia	0.080	0.093	0.091	0.107 ○	0.076
lymphoma	0.067	0.076	0.098 ○	0.022 ●	0.009 ●
nci	0.389	0.452 ○	0.405	0.418 ○	0.396
prostate	0.085	0.085	0.101 ○	0.085	0.088
srbct	0.064	0.072	0.074	0.035 ●	0.007 ●
win-lose-tie	–	1-9-0	1-9-0	3-7-0	3-7-0

Figura 6: Desempenho do classificador Random Forest utilizando variações do algoritmo RF como seletor de genes (Taxa de erro). [2, Table 5]

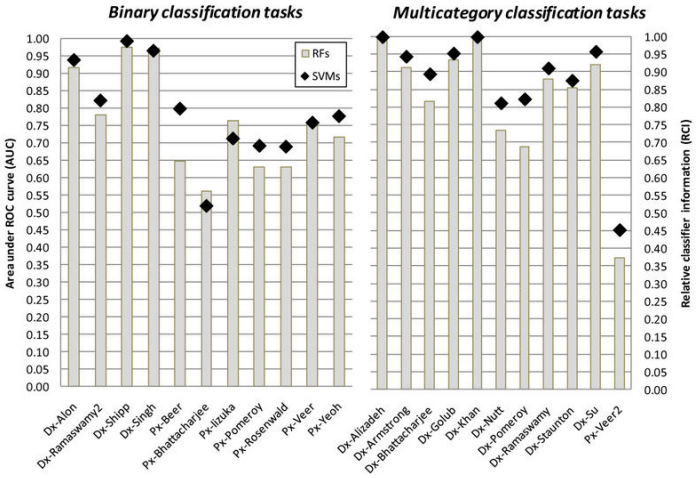

Figura 8: Desempenho dos classificadores RF e SVM com seleção de genes. [8, Figure 2]

diferentes base de dados, utilizando a seleção de características anteriormente à classificação (figura 8) e a aplicação dos classificadores mencionados em todos os genes das bases (figura 7). Segundo os autores do trabalho, não há diferença estatística significativa em nenhum dos casos em que o algoritmo RF obteve maior taxa de acerto nas bases de dados do estudo, sobressaindo-se o melhor desempenho do SVM nesses casos. Comparando a figura 7 e 8, também podemos observar que o uso de seleção de genes melhorou o desempenho de ambos classificadores na maioria das bases, principalmente para o SVM que teve sua melhora em até 10,06% na métrica AUC sendo que o RF teve o aumento máximo de 5,4%. Na figura 9, podemos observar que em duas das bases com maiores quantidades de genes (*Px-Bhattacharjee* e *Px-Veer2*) houve as piores taxas de acerto na etapa de classificação, o que pode indicar dúvidas a respeito de que poderiam existir outros subconjuntos de genes que poderiam maximizar os resultados dos classificadores.

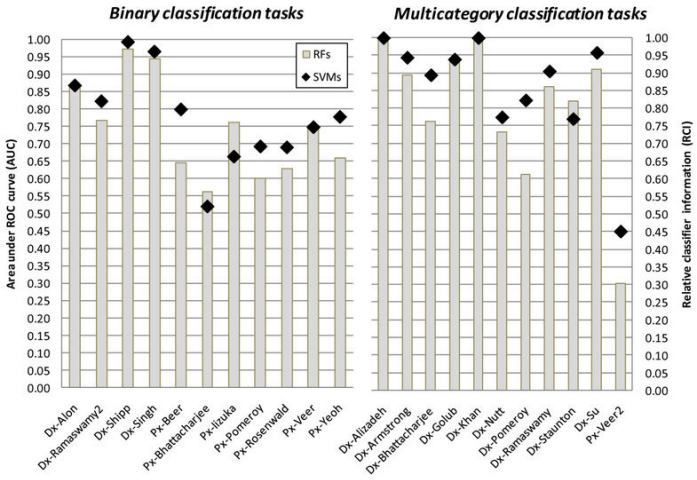

Figura 7: Desempenho dos classificadores RF e SVM sem seleção de genes. [8, Figure 1]

Task & dataset	*No gene selection*	*RFE*	*RFVS1*	*RFVS2*	*KW*	*S2N*
Dx-Alizadeh	4026	12	62	73	19	15
Dx-Alon	2000	105	16	3	15	13
Dx-Armstrong	11225	74	709	57	106	48
Dx-Bhattacharjee	12600	289	27	15	1864	653
Dx-Golub	5327	12	456	336	42	4
Dx-Khan	2308	28	17	18	15	11
Dx-Nutt	10367	1598	126	101	476	926
Dx-Pomeroy	5920	186	34	16	70	435
Dx-Ramaswamy	15009	3346	966	411	8248	10277
Dx-Ramaswamy2	13247	1576	12	4	4129	1364
Dx-Shipp	5469	8	15	6	13	89
Dx-Singh	10509	157	58	21	22	38
Dx-Staunton	5726	169	152	73	93	97
Dx-Su	12533	2429	845	320	1318	1927
Px-Beer	7129	201	15	7	953	1380
Px-Bhattacharjee	12600	21	46	7	138	61
Px-Iizuka	7070	103	38	7	168	185
Px-Pomeroy	7129	70	29	13	445	439
Px-Rosenwald	7399	2338	124	27	3201	3897
Px-Veer	24188	1056	124	20	5388	4405
Px-Veer2	24188	491	149	39	1194	1764
Px-Yeoh	12240	1187	21	6	3077	1869

Figura 9: Quantidade de genes selecionados. [8, Table 3]

VI. Materiais e métodos

O código fonte dos experimentos realizados neste trabalho pode ser acessado em https://github.com/BrunoMeyer/gene-selection-to-classification.

A. *Base de dados*

Foram utilizadas três matrizes de expressão gênica provenientes de uma base de dados pública [7], disponível em http://www.gems-system.org/. Cada base consiste nas seguintes propriedades:

- *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;

B. Classificadores

A biblioteca *scikit-learn* foi utilizada como base para a implementação e avaliação dos classificadores. Por não ser o objetivo deste trabalho, os parâmetros encontrados não foi considerada uma busca por valores ótimos para os parâmetros. Foram utilizados os seguintes algoritmos e parâmetros:

- *Multi Layer Perceptron (MLP)*
    - solver=lbfgs,
    - alpha=1e-4,
    - activation=logistic,
    - hidden_layer_sizes=5,
    - learning_rate_init=0.0001,
    - max_iter=100000,
    - random_state=1

- *Support Vector Machine (SVM)*
    - GridSearch com os seguintes intervalos:
        - C: [1, 10, 100, 1000]
        - kernel: [linear, rbf]
        - gamma: 0.001

- *Gaussian Naive Bayes (GNB)*

- *Bernoulli Naive Bayes (BNB)*

- *Multinomial Naive Bayes (MNB)*

- *k-nearest neighbors (KNN)*
    - K=1

- *Random Forest (RDF)*
    - GridSearch com os seguintes intervalos:
        - max_depth: [2, 40],
        - n_estimators: [4, 30]

C. Métricas

Para avaliar o desempenho de cada classificador, foi utilizada a média da área sobre a curva ROC (AUC) entre um total de 5 *folds* ao se utilizar o *Cross Validation* como validação estatística, mantendo a proporção de instâncias entre as classes de cada base. Devido ao problema de não ser possível obter a curva ROC para um problema multiclasse, nas bases que haviam essa propriedade foi calculada a média entre as áreas sobre as curvas de cada classe.

D. Seleção de genes

Como base, foram utilizadas as bibliotecas *scikit-learn* e DEAP da linguagem de programação *python* para implementar os métodos citados a seguir:

- Filter
    - *Variance Threshold*

- Wrapper
    - Algoritmo Genético - SVM
    - Algoritmo Genético - RDF
    - Algoritmo Genético - MNB

- Embedded
    - *SVM-RFE*
    - *MNB-RFE*
    - *Random Forest Model Selector*

Nos algoritmos baseados em RFE, foram utilizados como parâmetros o estabelecimento da seleção de 150 genes, e passo de execução=1 (número de genes removidos a cada iteração). Nas variações do algoritmo genético, foi utilizado uma população de tamanho 70 por 20000 gerações, probabilidade de mutação de 5% e taxa de cruzamento igual a 30%.

VII. Resultados e discussões

A. Eficiência

As tabelas a seguir ilustram o desempenho da combinação de cada classificador e seletor mencionado na seção de metodologia deste trabalho.

	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**

Figura 10: 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.

	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

Figura 11: 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.

	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

Figura 12: 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.

Pode-se observar que as bases possuem certa afinidade por determinados classificadores. Em contraste, o SVM-RFE apresentou melhores resultados como seletor nos experimentos realizados. Ao comparar o SVM-RFE com o SVM-MNB é possível observar uma notável diferença, provavelmente devido à grande diferença na valoração das importâncias dos atributos calculadas dentro do modelo de cada classificador. Neste quesito, o SVM possui uma vantagem devido ao fato de que em seu modelo é gerada uma fronteira de decisão baseada em um problema de maximização, cujo resultado é refletido no uso desse modelo para considerar a importância de cada atributo a ser selecionado.

O algoritmo MNB apresentou afinidade com o algoritmo genético quando combinados em um sistema de classificação. Talvez, alterações nos parâmetros ou o uso de algoritmos semelhantes possam apresentar resultados próximos ou superiores ao SVM-RFE.

Percebe-se também que apesar do SVM apresentar o melhor resultado como núcleo no RFE dentre os resultados obtidos, o mesmo não teve o melhor desempenho na segunda e terceira base de dados, o que pode ser vistu nas figuras 11 e 12 onde o algoritmo MLP obteve o melhor desempenho como classificador.

B. *Custo computacional*

A tabela a seguir contém o tempo de execução (em CPU) de cada algoritmo utilizado como seletor. Cada coluna tem o seu maior valor em negrito, e o maior resultado da tabela está sublinhado.

	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**

Figura 13: Tempo (segundos) de execução em CPU para cada base de dados utilizada no trabalho.

Vemos que o algoritmo genético apresenta um alto custo computacional no experimento, entretanto, sua escalabilidade é linear em relação aos seus três principais parâmetros, que podem ser controlados: número de gerações, tamanho da população e classificador utilizado como base.

Percebe-se também, de acordo com a Figura 13, uma grande significância no custo em relação ao classificador utilizado como núcleo dos algoritmos RFE e genético, que podem ser simplificados ou ter o aumento de suas complexidades em relação aos parâmetros que são utilizados para o mesmos. A seleção de genes baseada no modelo do classificador *Random Forest* e a seleção baseada na limiarização por variância de atributo demonstraram um baixo custo computacional, porém também apresentaram um mal desempenho em comparação aos outros algoritmos.

Também, ao comparar os tempos de execução entre a primeira base dados, com aproximadamente 6000 genes e 90 instâncias e segunda base de dados, com aproximadamente 10000 genes e 50 instâncias, vemos que o número de instâncias no problema possui uma grande relevância para o custo computacional em relação ao impacto da variação do número de genes.

VIII. Conclusão

Expressão gênica é um problema complexo que em determinados casos podem envolver vários genes. Percebe-se também que na maioria dos casos de estudo, as instâncias (números de exemplos) são poucas, o que permite algumas abordagens específicas para reduzir a dimensionalidade do problema, como o algoritmo SVM-RFE.

Independente da abordagem e classificadores utilizados para a seleção de genes, é possível utilizar qualquer outro classificador ou tecnologia (como algoritmos de clusterização) em etapas posteriores. Um exemplo é o uso do algoritmo *Random Forest* nos resultados obtidos por meio do SVM-RFE, o que permitiria uma análise mais detalhada da interação entre os genes selecionados em relação às classes de estudo como a identificação de expressão de câncer.

Podemos dizer, com base na revisão de literatura do presente trabalho, que os classificadores e algoritmos de seleção de genes se comportam de forma diferente conforme as bases de estudos. Percebe-se também que o SVM possui um melhor desempenho como classificador que o Random Forest em bases de dados de cânceres porém, o mesmo não acontece quando comparamos o algoritmo SVM-RFE e métodos de seleção baseados em Random Forest para a seleção de atributos.

Em relação ao algoritmo SVM-RFE, há um detalhe sutil em sua definição. Diferente de uma abordagem exaustiva, sua execução ocorre proporcionalmente em relação ao número de atributos iniciais. Já na abordagem exaustiva, todas as combinações seriam testadas, ou seja, uma solução inviável para os problemas de expressão gênica que em sua maioria contêm muitas variáveis e demandariam um custo computacional muito elevado no calculo da combinatória de todas as possibilidades de subconjuntos de genes.

Há vantagens e desvantagens em relação às abordagens *Filter*, *Wrapper* e *Embedded* para a seleção de características. As abordagens *Embedded* demonstram ter melhor eficiência para a seleção de atributos em sistemas de classificações, o que pode ser consequência de suas características ao considerar um classificador em sua execução e analisar o modelo gerado pelos classificadores para qualificar a importância dos atributos. Vemos que o algoritmo SVM-RFE apresentou o melhor resultado

durante os experimetos realizados e que algumas combinações como o algoritmo *Multinomial Naive Bayes* (MNB) combinado ao algoritmo genético podem também apresentar bons resultados em algumas bases de dados.

Cada problema de classificação e bases de dados possuem suas peculiaridades e características, o que causa grande impacto no tempo de execução e desempenho dos algoritmos utilizados. Considerando o estado da arte atual na resolução do problema de seleção de genes, pode-se dizer que é possível que ainda sejam criados métodos e técnicas que melhorem as soluções já conhecidas devido ao avanço das diversas áreas que estão relacionadas ao tema.

Referências

[1] Diogo G Barardo et al. "Machine learning for predicting lifespan-extending chemical compounds". Em: *Aging (Albany NY)* 9.7 (2017), p. 1721.

[2] Houtao Deng e George Runger. "Gene selection with guided regularized random forest". Em: *Pattern Recognition* 46.12 (2013), pp. 3483–3489.

[3] Ramón Dıaz-Uriarte e Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest". Em: *BMC bioinformatics* 7.1 (2006), p. 3.

[4] Kai-Bo Duan et al. "Multiple SVM-RFE for gene selection in cancer classification with expression data". Em: *IEEE transactions on nanobioscience* 4.3 (2005), pp. 228–234.

[5] Isabelle Guyon et al. "Gene selection for cancer classification using support vector machines". Em: *Machine learning* 46.1-3 (2002), pp. 389–422.

[6] Yvan Saeys, Iñaki Inza e Pedro Larrañaga. "A review of feature selection techniques in bioinformatics". Em: *bioinformatics* 23.19 (2007), pp. 2507–2517.

[7] A Statnikov, CF Aliferis e I Tsamardinos. "Online supplement". Em: *http:// discover1. mc. vanderbilt. edu/ discover/ public/ GEMS* ().

[8] Alexander Statnikov, Lily Wang e Constantin F Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification". Em: *BMC bioinformatics* 9.1 (2008), p. 319.

[9] Alexander Statnikov et al. "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis". Em: *Bioinformatics* 21.5 (2004), pp. 631–643.

[10] Wouter G Touw et al. "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?" Em: *Briefings in bioinformatics* 14.3 (2012), pp. 315–326.

[11] Gary Zweiger. "Knowledge discovery in gene-expression-microarray data: mining the information output of the genome". Em: *Trends in biotechnology* 17.11 (1999), pp. 429–436.