Resultados parciais

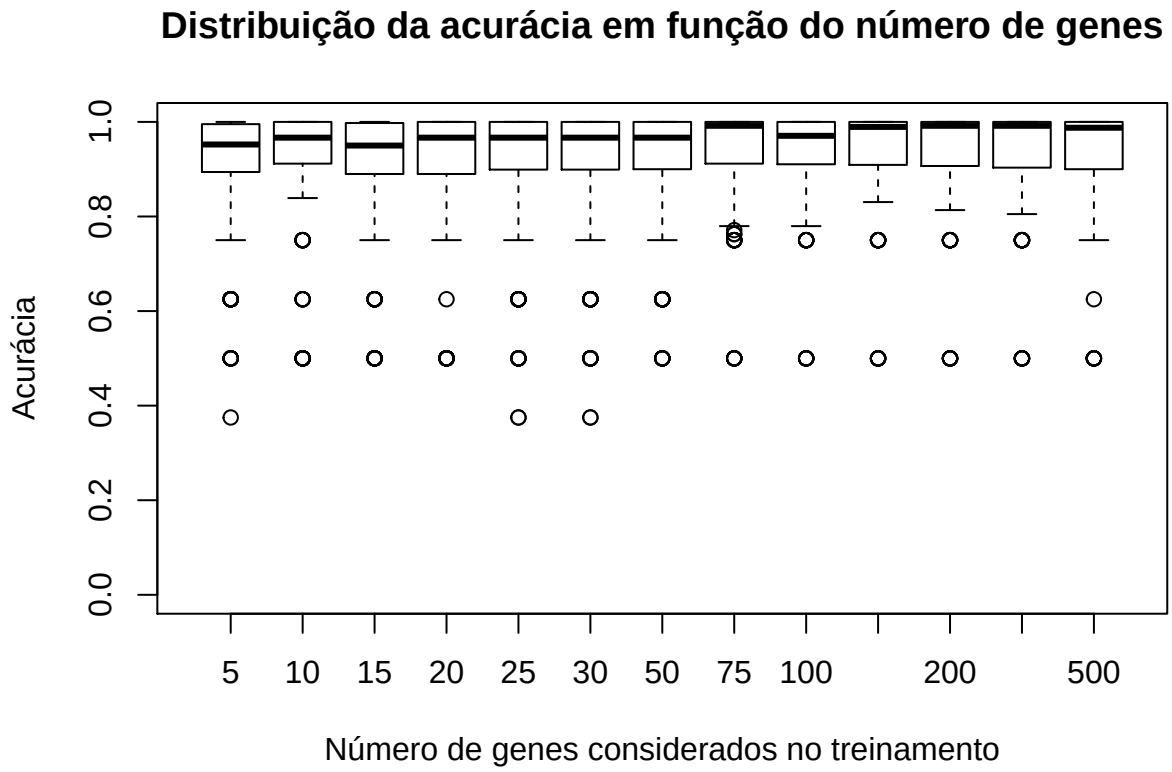
Resultados parciais

Parâmetros testados:

- proporção de treino = 0.6
- número de genes = 5, 10, 15, 20, 25, 30, 50, 75, 100, 150, 200, 250, 500
- classificador = svmRadial, J48, knn, nnet, OneR
- número de bags = 50, 100, 150, 200
- algoritmo de seleção de atributos = infoGain, chiSquared
- algortimo de agregação = média

Número de genes

O número de genes parece não influenciar na performance. Porém, a maioria dos classificadores está atigindo quase 100% de acurácia. Isso deve ser investigado melhor. Dessa forma, nada podemos concluir, por enquanto.



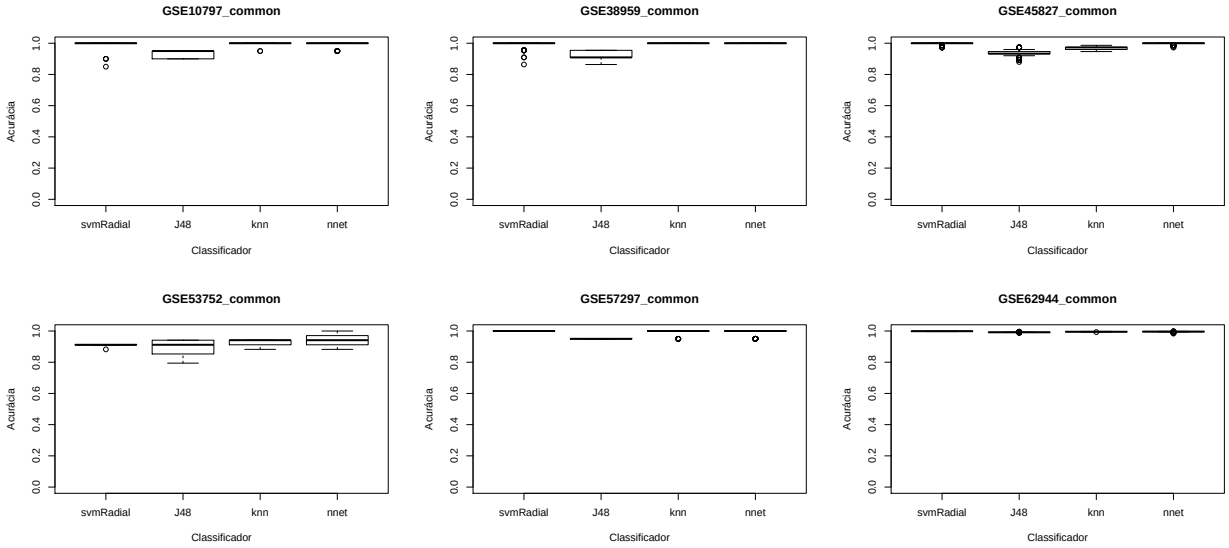
Datasets

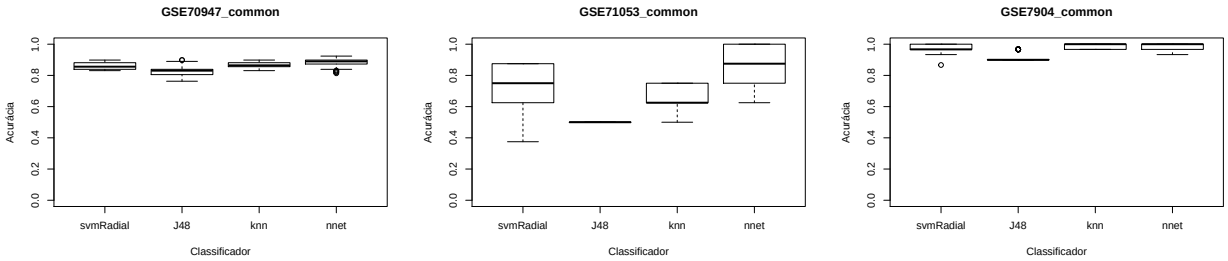
Características

	Number.of.tumor.samples	Number.of.normal.samples
GSE10797_common	27	5
GSE38959_common	30	13
GSE45827_common	98	17
GSE53752_common	46	21
GSE57297_common	25	7
GSE62944_common	1119	113
GSE70947_common	148	148
GSE71053_common	6	12
GSE7904_common	42	18

	Number.of.samples	Number.of.genes	Tumor.Ratio
GSE10797_common	32	7973	0.84375
GSE38959_common	43	7973	0.6976744
GSE45827_common	115	7973	0.8521739
GSE53752_common	67	7973	0.6865672
GSE57297_common	32	7973	0.78125
GSE62944_common	1232	7973	0.9082792
GSE70947_common	296	7973	0.5
GSE71053_common	18	7973	0.3333333
GSE7904_common	60	7973	0.7

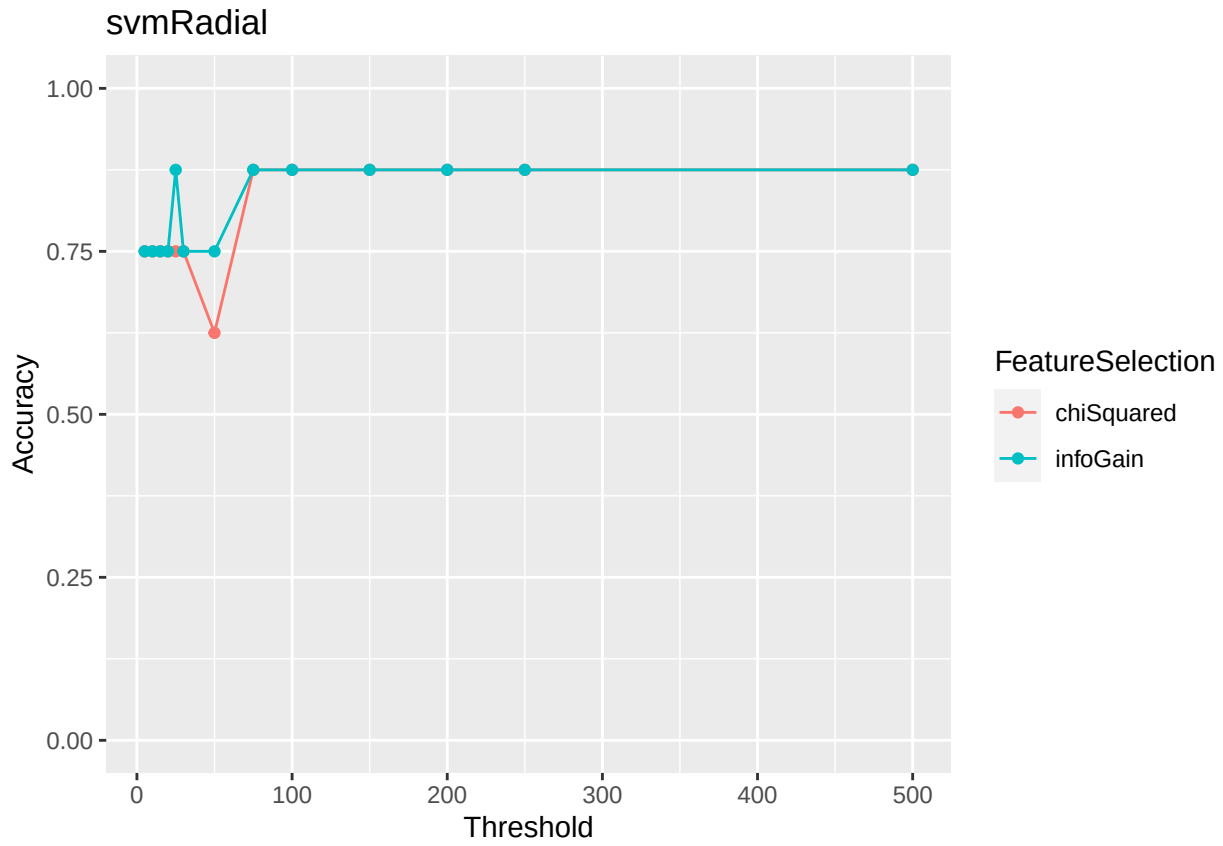
A maioria dos datasets e classificadores possuem um ótimo desempenho, estranhamente. Os métodos apresentam um desempenho quase constante na maioria dos modelos, independente do número de genes utilizados no treinamento. Apenas o dataset GSE71053 possui uma variação maior no desempenho de acordo com o classificador. Analisaremos esse datasets mais porfundamente em seguida.

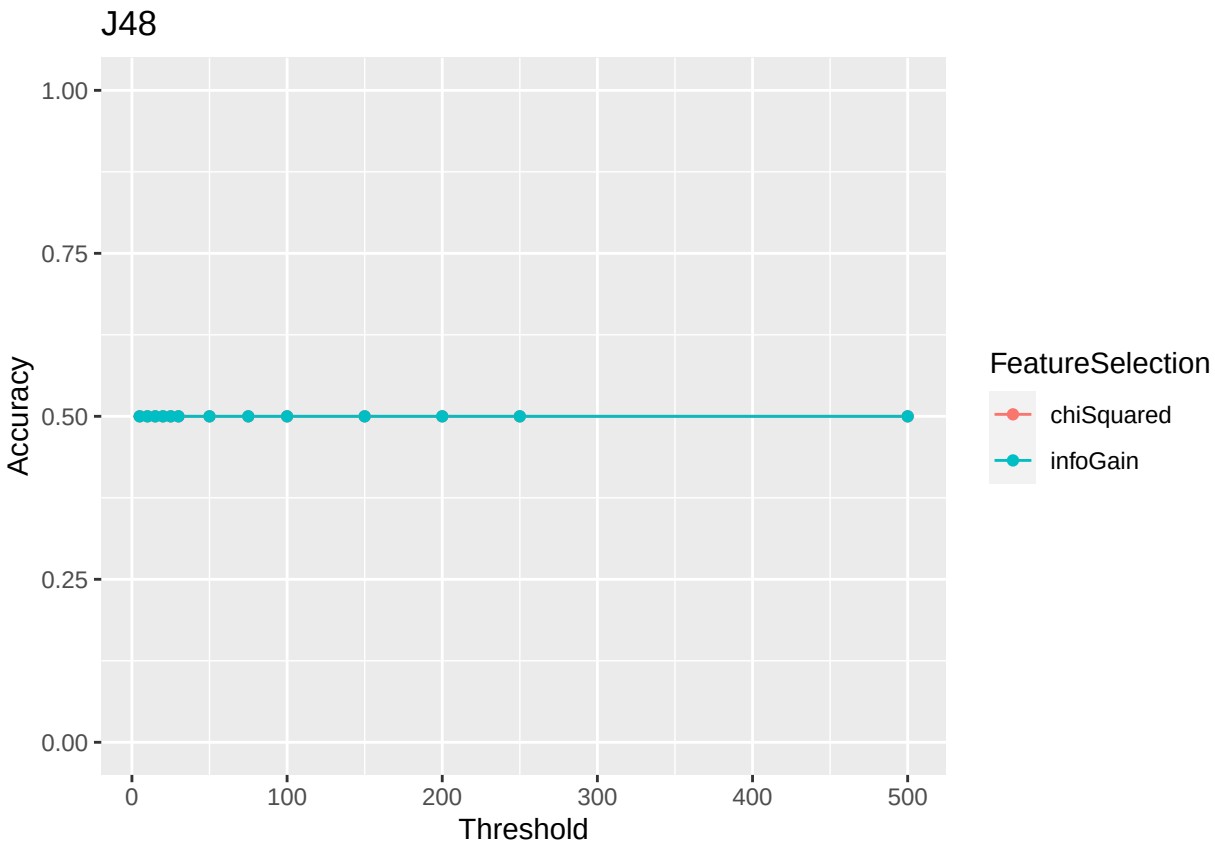


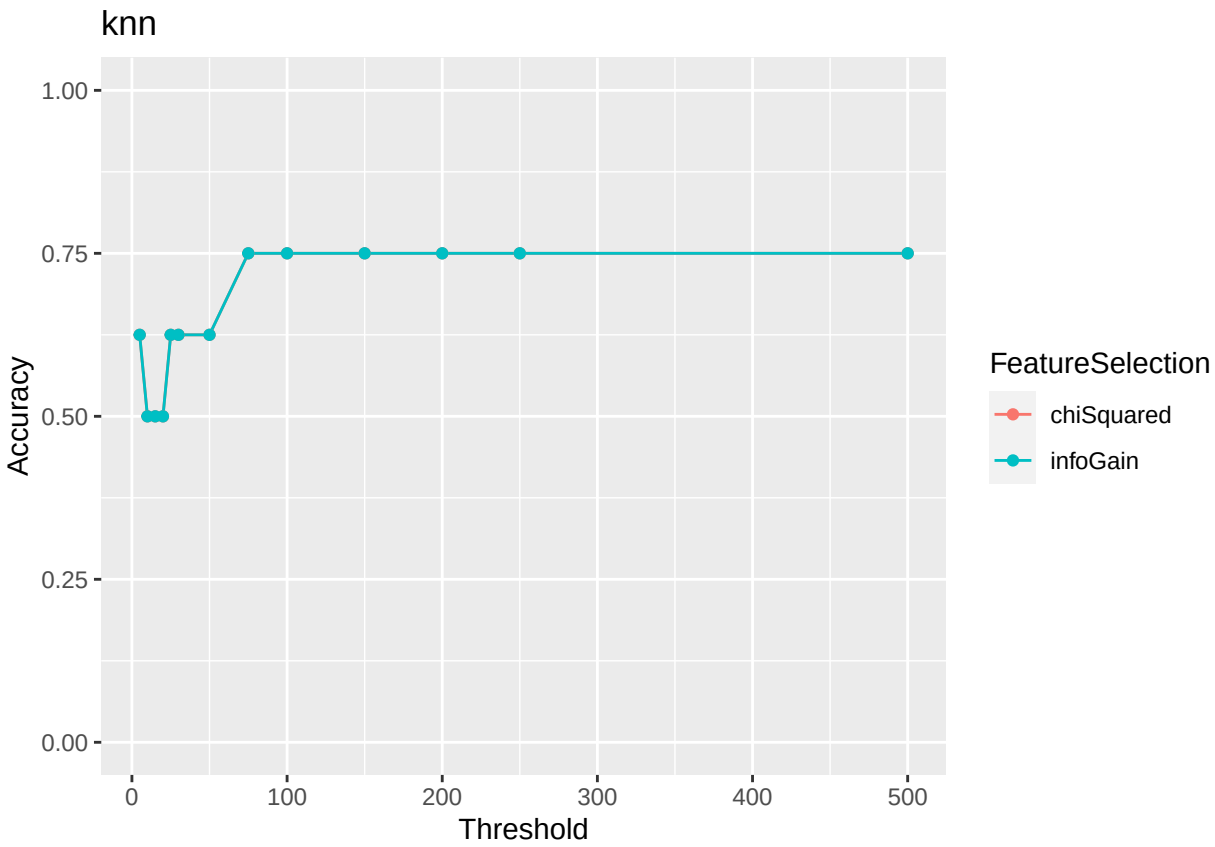


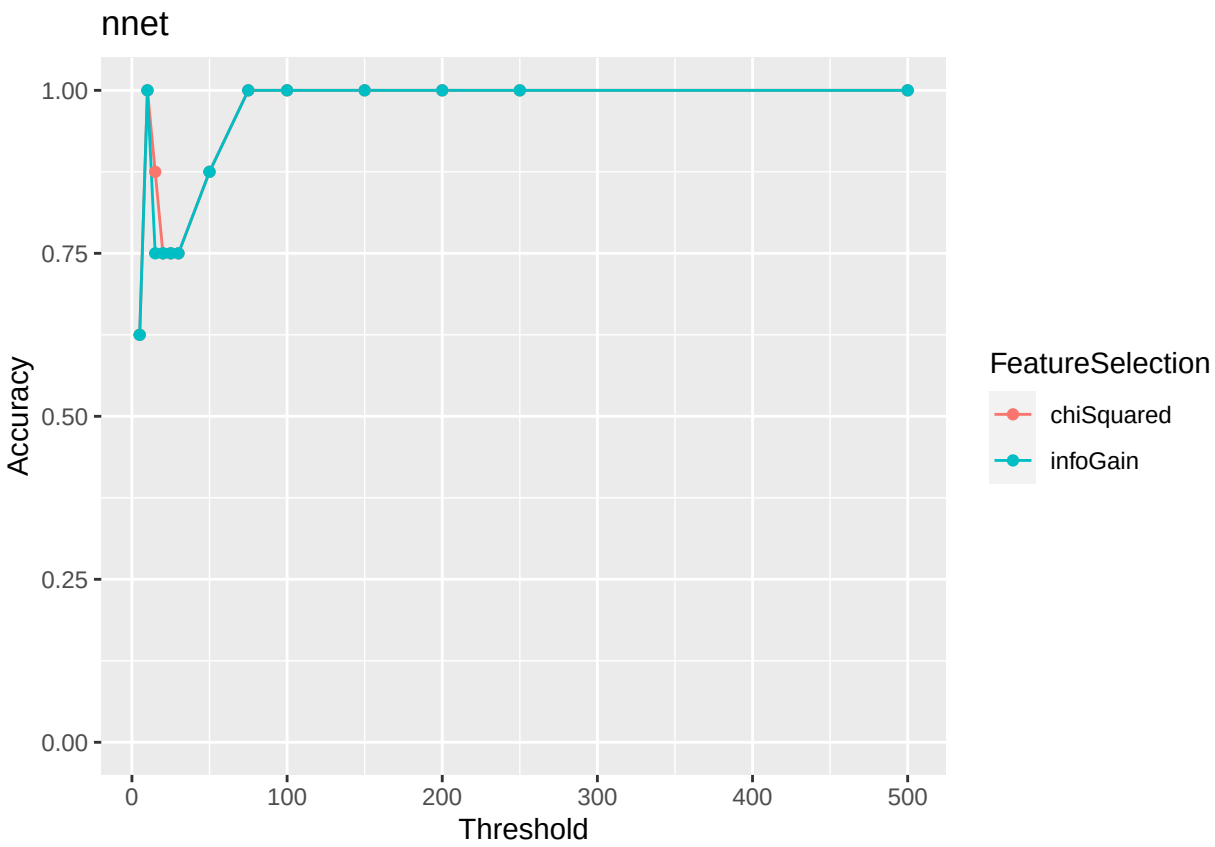
GSE71053

A diferença do GSE71053 para os outros datasets é a quantidade de amostras com tumor em relação ao total. O GSE71053 é o único dataset com menor número de amostras com tumor do que amostras normais, o que *poderia* explicar a variação no desempenho. Na maioria dos casos, a acurácia entre **chiSquared** e **infoGain** são as mesmas em função do número de genes utilizados (threshold). O caso do modelo J48 é interessante, pois não importa o número de genes, o algoritmo não aprende a classificar corretamente as amostras. Nesse caso, a acurácia é constante (50%), ou seja, a classificação é aleatória. Como esperado, a maior variação de desempenho em todos os modelos, com exceção do J48, concentra-se nos números menores de threshold, aitingindo a acurácia máxima com 75 genes.




J48
1.00
0.75
0.50
0.25
0.00
Accuracy
0
100
200
300
400
500
Threshold
FeatureSelection
chiSquared
infoGain


knn
1.00
0.75
0.50
0.25
0.00
Accuracy
0
100
200
300
400
500
Threshold
FeatureSelection
chiSquared
infoGain


nnet
1.00
0.75
0.50
0.25
0.00
Accuracy
0
100
200
300
400
500
Threshold
FeatureSelection
chiSquared
infoGain