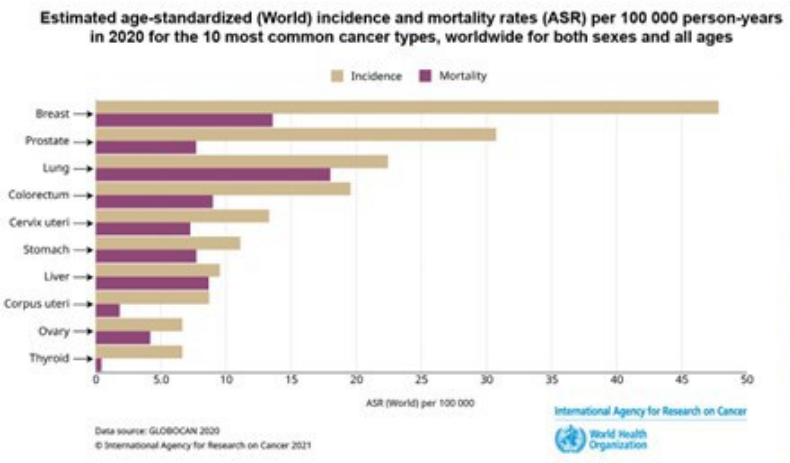
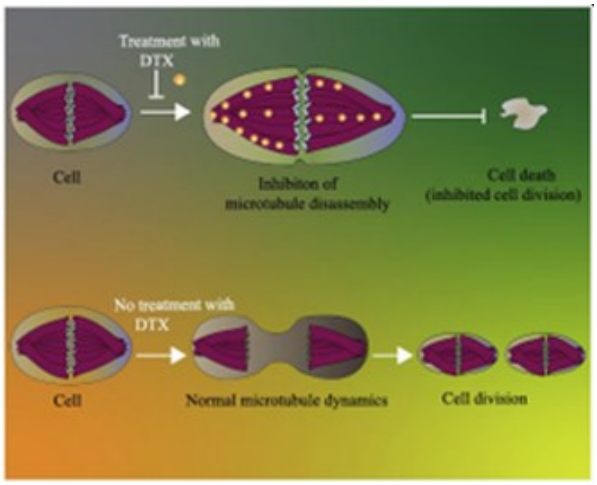
Introdução à Modelagem e Aprendizado
ELTD1 - DAELN UTFPR-CT

Prof. Heitor S. Lopes (2024-1) - Exercício #5





Objetivo: utilizar métodos de seleção de atributos e redução de dimensionalidade

Seleção de genes sensíveis ao tratamento químico de câncer de mama

a. O câncer de mama é o que tem mais alta incidência na população mundial, e a segunda mais alta mortalidade dentre todos os tipos de câncer. A quimioterapia diminui significativamente o risco de morte para os casos de câncer de mama com tumores passíveis de remoção por cirurgia. Docetaxel é um dos agentes químicos mais ativos neste tratamento. Porém, resistência ao tratamento é frequente. O *dataset* utilizado é o *Breast Cancer and Docetaxel Treatment*, disponível no Orange, e tem 24 casos clínicos e 9486 atributos sendo, portanto, um típico *dataset* horizontal. Cada atributo é o valor da expressão gênica normalizada para um determinado gene, e o *dataset* contém 14 pacientes com tumores resistentes ao tratamento, e 10 com tumores sensíveis ao tratamento. O objetivo é determinar, dentre os 9486 genes, quais os mais relevantes para induzir a resistência ao tratamento.

b. Neste exercício serão utilizados três métodos de seleção de atributos baseado na abordagem *Filter* para reduzir a dimensionalidade do *dataset*: InfoGainRatio, Relief e FCBF. Para a classificação, são utilizados: Árvore de Decisão (AD), Support Vector Machine (SVM) e Rede Neural (NN), sempre com os respectivos parâmetros-padrão. Na classificação, utilizar validação cruzada estratificada de 5-*folds*. Repetir os experimentos conforme a tabela a seguir: na parte verde é registrado o nome do gene indicado como o mais discriminante pelos métodos de seleção de atributos. Na parte cinza é registrado o valor de F1 para as diversas combinações a seguir: primeiramente utilizando a expressão gênica do gene top-1 (OneRule), depois dos top-10 genes, top-100 genes, e com todos os genes (sem seleção de atributos).

			Número de features selecionadas											
		Top-1 gene	top-1 (oneRule)			top-10			top-100			todas		
Método de seleção de atributos	InfoGain Ratio													
	Relief													
	FCBF													
			AD	SVM	NN	AD	SVM	NN	AD	SVM	NN	AD	SVM	NN
			Método de classificação											

c. Após os experimentos, analise os resultados e informe qual a combinação de método de seleção de atributos, número de atributos selecionados e método de classificação que levou ao melhor resultado.

d. Com base nos top-10 genes apontados pelos métodos, enumere os genes que sejam de maior consenso pelos métodos de seleção de atributos. Estes, possivelmente, são os genes que mais influenciam a resistência ao tratamento do câncer.