UTFPR – CPGEI – Campus Curitiba
Mineração de Dados e Descoberta de Conhecimento

Profs. Heitor S. Lopes & Thiago H. Silva (UTFPR, 2025) - Exercise #6

Feature Selection and Dimensionality Reduction

The main objective of this exercise is to find genes sensitive to the chemical treatment of breast cancer.

A) Informations:

a. Breast cancer has one of the highest incidence in the world population, and the second highest mortality rate among all types of cancer. Chemotherapy can significantly reduce the risk of death for breas cancer cases with tumors that can be removed by surgery.

b. Docetaxel is one of the most active chemical agents in such treatment. However, resistance to treatment is frequent.

c. The dataset used in this exercise is the **Breast Cancer and Docetaxel Treatment**, directly available in the Orange datasets. It has 24 clinical cases each one with 9,486 attributes. Therefore it is a typical horizontal dataset.

d. Each attribute corresponds to the value of normalized gene expression for a given gene, and the dataset contains 14 patients with treatment-**resistant** tumors and 10 with treatment-**sensitive** tumors.

e. The objective is to determine, among the 9,486 genes, which are the most relevant for inducing resistance to treatment.

B) Procedure:

f. In this exercise, three attribute selection methods based on the Filter approach will be used to reduce the dimensionality of the dataset: InfoGainRatio, Relief and FCBF.

g. For the classification task, the following methods should be used: Decision Tree (DT), Support Vector Machine (SVM) and Neural Network (NN), always with their respective **default** parameters.

h. In the classification, always use 5-fold stratified cross-validation.

i. Repeat the experiments according to the following table: in the green part, the name of the gene indicated as the most discriminant by the attribute selection methods is recorded. In the gray part, the F1 value is recorded for the following various combinations: first using the gene expression of the top-1 gene (OneRule), then the top-10 genes, top-100 genes, and with all genes (without attribute selection).

j. After the experiments, analyze the results and report which combination of attribute selection method, number of attributes selected, and classification method led to the best classification result.

k. Finnaly, based on the top 10 genes indicated by the methods, list the genes that have the greatest consensus according to the attribute selection methods. These are possibly the genes that most influence resistance to cancer treatment.

l. **Optionally**, search in the internet further information about the genes found regarding their relationship with the breast cancer treatment.

			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											