# METODOLOGIAS DE CLASSIFICAÇÃO SUPERVISIONADA PARA ANÁLISE DE DADOS DE *MICROARRAYS*

**Sílvia Pedro Rebouças, Lisete Sousa e Ana Pires**

CEAUL, Faculdade de Ciências da Universidade de Lisboa, smdpedro@gmail.com
CEAUL, Faculdade de Ciências da Universidade de Lisboa, lmsousa@fc.ul.pt
CEMAT, Instituto Superior Técnico, Universidade Técnica de Lisboa, apires@math.ist.utl.pt

SUMÁRIO

## SUMÁRIO

O desenvolvimento da tecnologia de *microarrays* tornou possível monitorizar o nível de expressão de milhares de genes em simultâneo, tendo revolucionado a investigação em Biologia e em todas as áreas relacionadas, incluindo a Medicina. Esta tecnologia lançou a necessidade de tratar conjuntos de dados complexos, com um número muito elevado de variáveis para um número geralmente reduzido de observações. Diversos métodos estatísticos e de aprendizagem automática têm sido desenvolvidos com o objectivo de extrair informação desses conjuntos de dados. De realçar que a elevada dimensionalidade dos dados de *microarrays* tornam a aplicação dos métodos de classificação morosa e por vezes inviável, requerendo uma redução prévia de dimensionalidade.

Neste trabalho ilustra-se a aplicação de algumas técnicas de classificação supervisionada, nomeadamente, regressão logística penalizada, árvores de classificação, redes neuronais e classificador dos *k* vizinhos mais próximos, a dados de *microarrays* na área da Oncologia. Os métodos de classificação aplicados permitiram classificar com grande precisão indivíduos quanto ao tipo de leucemia ou identificar indivíduos com cancro do cólon a partir da expressão dos seus genes. Os modelos ajustados apresentaram melhor desempenho no conjunto de dados da leucemia, sendo de realçar o desempenho do classificador dos vizinhos mais próximos.

**Palavras-chave:**

Árvores de Classificação, Classificador dos Vizinhos mais Próximos, Classificação Supervisionada, *Microarrays*, Redes Neuronais, Regressão Logística Penalizada.

# 1. INTRODUÇÃO

Os estudos desenvolvidos na área da análise de dados de *microarrays* tentam dar resposta a 3 questões fundamentais (Stekel, 2003):

- Quais os genes com expressão diferencial num conjunto de dados relativamente a outro?
- Quais as relações presentes entre os genes ou entre os indivíduos em estudo?
- Como classificar indivíduos tendo por base a quantificação da expressão dos seus genes?

Os métodos de classificação supervisionada são úteis para responder à 3ª questão, tendo sido aplicados em estudos em que se utilizam dados de indivíduos com patologias, resultados ou fenótipos conhecidos e se pretendem desenvolver modelos que permitam, a partir de um número reduzido de genes, predizer a que grupo pertence cada indivíduo. O desenvolvimento destes modelos preditivos depende de técnicas estatísticas e computacionais que continuam a ser alvo de investigação activa (Stekel, 2003).

Nos estudos de dados de *microarrays*, o número de observações é relativamente pequeno comparado com o número de genes, ou seja, com o número de variáveis, que são geralmente milhares. A não ser que seja feita uma redução prévia do número de variáveis, os métodos de classificação usuais não podem ser aplicados ou apresentam um fraco desempenho, verificando-se problemas inerentes ao facto do número de variáveis ser muito superior ao de observações, tais como a multicolinearidade. Por esta razão, as metodologias de selecção de genes e de redução de dimensionalidade têm merecido a atenção de vários investigadores. A estratégia implementada influencia o desempenho dos métodos de classificação a aplicar posteriormente.

Alguns métodos de classificação requerem uma selecção prévia de variáveis ou a aplicação de técnicas de redução de dimensionalidade, outros têm uma selecção de variáveis intrínseca, podendo ser aplicados directamente.

A selecção de variáveis pode ser feita por métodos univariados ou multivariados. Os métodos de selecção univariados baseiam-se na utilidade marginal de cada variável, sendo as variáveis ordenadas de acordo com determinado critério que reflicta a sua associação com o fenómeno de interesse. As primeiras variáveis do conjunto ordenado são então seleccionadas. De entre os critérios mais usuais encontram-se os valores *p* obtidos para testes *t* ou análise de variância, consoante o número de grupos, ou para testes não paramétricos tais como o de Mann-Whitney ou o de Kruskall-Wallis. O recurso a métodos bayesianos (Antunes & Sousa, 2008) ou à *false discovery rate* (Benjamini & Hochberg, 1995) são outras das possibilidades. Contudo, os métodos de selecção univariados não têm em conta correlações ou interacções entre variáveis, pelo que, o conjunto de variáveis com melhor poder discriminante univariado não é necessariamente o melhor subconjunto de variáveis (Boulesteix et al., 2008).

Os métodos multivariados de selecção de variáveis são caracterizados pelo critério usado para ordenar os subconjuntos de variáveis e pelo algoritmo aplicado (Boulesteix et al., 2008). O critério pode ser baseado na precisão da classificação (*wraper criteria*) ou no poder de discriminação de cada subconjunto de variáveis sem

recorrer ao classificador (*filter criteria*). Alguns algoritmos aplicados para encontrar os subconjuntos de variáveis restringem a busca a pares de variáveis ou subconjuntos de variáveis pouco correlacionadas (Jaeger et al., 2003), outros, tais como os algoritmos moleculares (Ooi & Tan, 2003), procuram os melhores subconjuntos da globalidade das variáveis.

Os métodos de redução de dimensionalidade procuram sintetizar em poucas variáveis a informação contida nas variáveis originais. De entre estes métodos, destacam-se a análise de componentes principais ou o método equivalente da análise de valores e vectores singulares e, com melhores resultados nos estudos de *microarrays*, o método dos mínimos quadrados parciais (Nguyen & Rocke, 2004). Neste tipo de dados, existem geralmente *outliers*, que distorcem os resultados, mas podem ser as observações mais importantes. A solução sugerida na literatura é o uso de métodos robustos, nomeadamente, a análise de componentes principais robustas (Branco & Pires, 2009). Boulesteix & Tutz (2006) propõem reduzir a dimensionalidade aplicando os métodos de classificação a IPs (Padrões de Interacção) em alternativa às covariáveis originais, recorrendo às árvores de classificação para identificar esses IPs. Após a selecção de variáveis ou a aplicação de uma técnica de redução de dimensionalidade, os métodos de classificação podem ser aplicados de forma usual.

Alguns métodos de classificação não requerem uma selecção prévia de variáveis nem a redução prévia de dimensionalidade, desenvolvendo uma selecção de variáveis intrínseca. Estes métodos podem ser divididos em dois grupos (Boulesteix et al., 2008): métodos estatísticos baseados em penalização ou contracção, que permitem distinguir as variáveis irrelevantes das relevantes através da modificação dos seus coeficientes, tais como a regressão logística penalizada e as máquinas de suporte vectorial; e métodos de aprendizagem automática, tais como, os procedimentos *bagging* (Dudoit et al., 2002) e *random forests* (Diaz-Uriarte & Andrés, 2006).

De entre os métodos de classificação supervisionada aplicados a dados de *microarrays*, destacam-se: análise discriminante linear de Fisher (Satagopan & Panageas, 2003), análise discriminante linear diagonalizada e análise discriminante quadrática (Lee et al., 2005), regressão logística penalizada (Zhu & Hastie, 2004; Shen & Tan, 2005; Liao & Chin, 2007), árvores de classificação (Boulesteix et al., 2003; Boulesteix & Tutz, 2006), modelos bayesianos (Roth & Lange, 2004), classificador dos vizinhos mais próximos (Boulesteix & Tutz, 2006; Li et al., 2001), médias difusas (Asyali et al., 2005), modelos factoriais de misturas (Martella, 2006), redes neuronais artificiais (O'Neill & Song, 2003) e máquinas de suporte vectorial (Pirooznia & Deng, 2006).

Dudoit et al. (2002) e Lee et al. (2005) realizaram estudos comparativos de alguns métodos de classificação supervisionada. Estes autores concluíram que os métodos mais sofisticados revelam melhor desempenho que os clássicos e que este desempenho é influenciado pelo método de selecção de genes escolhido.

Boulesteix et al. (2008) reviram vários aspectos estatísticos da avaliação e validação dos classificadores de um ponto de vista prático e Dupuy & Simon (2007) sugeriram linhas orientadoras para boas práticas na análise de dados de *microarrays*, inclusive, na predição de classes.

Alguns autores, entre os quais, Li et al. (2004), Boulesteix & Tutz (2006) e Kim et al. (2006), aplicaram métodos de classificação supervisionada em mais de dois grupos. Este é um problema mais complexo e menos estudado, verificando-se que a capacidade preditiva dos modelos diminui com o aumento do número de grupos.

Lee et al. (2005) referem a importância de serem realizados estudos de comparação extensiva de técnicas de classificação supervisionada, para que os investigadores disponham de informação que os apoie na escolha da técnica mais apropriada para determinada situação. É neste sentido que se desenvolve o presente trabalho que pretende ilustrar a aplicação de técnicas de classificação supervisionada a dois conjuntos de dados da literatura e comparar os respectivos desempenhos no que diz respeito à capacidade de ajustamento, quantificado pela proporção de erros de classificação na amostra de modelação e, à capacidade de generalização, quantificada pela proporção de erros de classificação na amostra de validação.

## 2. MÉTODOS DE CLASSIFICAÇÃO SUPERVISIONADA

A classificação de indivíduos em grupos pode ser supervisionada ou não supervisionada. Na classificação supervisionada admite-se conhecida a classe que gerou cada padrão na amostra de modelação. O classificador é treinado para replicar a decisão correcta para novas amostras. Na classificação não supervisionada os padrões de treino não se encontram classificados, pelo que, os algoritmos têm que encontrar uma estrutura nos dados que permita dividi-los em grupos. Uma vez que a informação disponível é menor, a classificação é menos precisa do que a obtida com os métodos supervisionados, contudo, esta é a única solução possível para problemas em que não se dispõe de informação acerca dos grupos que geraram os dados.

Seja **X** uma matriz contendo a informação referente à quantificação da expressão de $p$ genes para $n$ indivíduos, na qual cada elemento $x_{ij}$ representa o nível de expressão do *j-ésimo* gene (variável) para o *i-ésimo* indivíduo (observação). Para cada indivíduo têm-se $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ e $y_i$, onde $y_i$ identifica o grupo a que pertence o indivíduo. Na classificação supervisionada pretende-se treinar classificadores numa amostra de modelação (*learning set*) $L = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{nL}, y_{nL})\}$, na qual $nL$ é a dimensão da amostra de modelação e utilizá-los para classificar novas amostras, denominadas amostras de validação ou teste (*test set*) $T = \{\mathbf{x}_1, \ldots, \mathbf{x}_{nT}\}$ com dimensão $nT$.

### 2.1. REGRESSÃO LOGÍSTICA PENALIZADA

Em muitos estudos estatísticos, somos confrontados com problemas em que se pretende estudar a relação entre variáveis, ou mais particularmente, analisar a influência que uma ou mais variáveis (explicativas), medidas em indivíduos ou objectos, têm sobre uma variável de interesse (resposta). A abordagem deste problema é frequentemente feita através do estudo de um modelo de regressão apropriado. Quando se pretende

relacionar uma variável resposta dicotómica e variáveis explicativas qualitativas ou quantitativas, recorre-se geralmente à regressão logística.

A regressão logística foi aplicada por Zhu & Hastie (2004) e Liao & Chin (2007) para desenvolver modelos de predição de uma variável resposta binária a partir de dados de *microarrays*.

Representando a variável resposta a presença ou ausência de doença (cancro, por exemplo), esta é definida de forma adequada do seguinte modo:

$$Y = \begin{cases} 1, & \text{se foi diagnosticada a doença} \\ 0, & \text{se não foi diagnosticada a doença} \end{cases}$$

O vector das variáveis explicativas é $\mathbf{x} = (x_1, ..., x_p)$, onde $x_j$ representa o nível de expressão do *j-ésimo* gene.

O modelo de regressão logística, que pretende estimar $\pi(\mathbf{x}) = P(Y = 1 | \mathbf{x})$, é construído a partir de uma amostra de modelação, podendo ser aplicado a um novo conjunto de dados (amostra de validação) para estimar a probabilidade de doença associada a cada observação desse conjunto.

A especificidade dos dados de *microarrays*, nomeadamente o elevado número de genes (*p*) e o reduzido número de indivíduos (*n*), requer a realização de uma etapa adicional de selecção de genes. A selecção de modelos desenvolvidos segundo esta abordagem em duas etapas requer novas ferramentas estatísticas, uma vez que a estimação do erro de predição deve ser feita tendo em conta a etapa de selecção de genes (Liao & Chin, 2007).

A primeira etapa consiste na selecção de um subconjunto de genes para incluir na regressão logística. Um método consiste em seleccionar os *q* genes com maior significância univariada (Dudoit et al., 2002), representando-se o nível de expressão do *j*-ésimo gene seleccionado por $x_j^*$, $j$ = 1, …, $q$. Na segunda etapa procede-se ao ajustamento do modelo de regressão logística, dado por:

$$\text{logit}\{\pi(\mathbf{x})\} = \beta_0 + \sum_{j=1}^{q} \beta_j x_j^* \qquad (2.1)$$

maximizando a log-verosimilhança penalizada:

$$l(\beta_0, \beta) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\} - \frac{1}{2\tau^2} \| \beta \|^2 \qquad (2.2)$$

onde $\pi_i = \pi(\mathbf{x}_i)$, $\| \beta \|$ é a norma euclideana de $\beta = (\beta_1, ..., \beta_q)$ e $\tau$ é um parâmetro de contracção positivo que controla o grau de contracção de $\beta$ em torno de 0.

Existem dois problemas chave para resolver: a escolha do número de genes (*q*) a seleccionar na primeira etapa e a escolha do melhor parâmetro de contracção ($\tau$) para um dado *q*. Para apoiar a escolha destes dois parâmetros, importa estimar o erro de

predição associado a determinada estratégia de construção do modelo. A estimação do erro de predição tendo em conta o processo de selecção tem sido realizada por validação cruzada ou por *bootstrap* não paramétrico (Ambroise & McLachlan, 2002; Simon et al., 2003; Braga-Neto & Dougherty, 2004; Efron, 2004).

Liao & Chin (2007) implementaram o método de regressão logística penalizada na biblioteca *GeneLogit* do *R: A Language and Environment for Statistical Computing*, tendo proposto um modelo *bootstrap* paramétrico, que se descreve a seguir, para uma estimação mais precisa do erro de predição. O modelo proposto contribui para orientar dois aspectos críticos da selecção de modelos: o número de genes a incluir no modelo e o contracção óptimo para a regressão logística penalizada. Os autores mostraram que seleccionar um número de genes superior a 20 contribui pouco para a redução do erro de predição.

Dado um vector de expressão de genes para o *i*-ésimo indivíduo ($\boldsymbol{x}_i$), $y_i$ segue uma distribuição Bernoulli($\pi_i$) com:

$$\text{logit}(\pi_i) = b_0 + \sum_{j=1}^{p} b_j x_{ij}, \quad i = 1,...,n \tag{2.3}$$

sendo $b_1, \ldots, b_p$ os coeficientes da regressão e $b_0$ a ordenada na origem.

Seja $n_1$ o número de indivíduos com cancro e $n_0$ o número de indivíduos sem cancro, a probabilidade de $\boldsymbol{y}$ condicionada a $y_1 + \ldots + y_n = n_1$ é dada por:

$$\frac{p(\mathbf{y} \mid \pi)}{\sum_{\mathbf{u} \in S} p(\mathbf{u} \mid \pi)} \tag{2.4}$$

onde $p(\mathbf{y} \mid \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_i^{\,y_i} (1-\pi_i)^{1-y_i}$, $\boldsymbol{\pi} = (\pi_1,...,\pi_n)$, $u_i$ assume o valor 1 para os indivíduos com cancro e 0 para os sem cancro, $\mathbf{u} = (u_1,...,u_n)$ e $S = \left\{ \mathbf{u} : \sum_{i=1}^{n} u_i = n_1 \right\}$.

Para desenvolver um modelo para os coeficientes $b_1, \ldots, b_p$, considere-se o subconjunto de genes com coeficiente de regressão não nulo $M = \{j: j = 1, ..., p \wedge b_j \neq 0\}$. Então:

$$\text{logit}(\pi_i) = b_0 + \sum_{j \in M} s_j \mid b_j \mid x_{ij}, \quad i = 1,...,n \tag{2.5}$$

onde $s_j = \dfrac{\mid b_j \mid}{b_j}$ é o sinal de $b_j$.

Se a expressão dos genes seguir uma distribuição normal multivariada, com valor esperado dependente do grupo e matriz de variâncias-covariâncias, $\mathbf{V}$, comum aos dois grupos, isto é, $\mathbf{X}|y = 1 \sim \text{N}(\boldsymbol{\mu}_1, \mathbf{V})$ e $\mathbf{X}|y = 0 \sim \text{N}(\boldsymbol{\mu}_0, \mathbf{V})$, onde $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})$, $\boldsymbol{\mu}_0 = (\mu_{01}, \ldots, \mu_{0p})$, tem-se que $(b_1, \ldots, b_p)^t = \mathbf{V}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ e $b_0$ depende da razão entre as probabilidades a priori dos dois grupos, a qual é geralmente estimada pela razão entre as dimensões amostrais dos dois grupos. Se além disso se verificar independência entre

as expressões dos diversos genes, isto é, entre as componentes do vector aleatório $\mathbf{X}|y = j$, $j$=0,1, então $\mathbf{V}$ = diag($v_1$, …, $v_p$) e tem-se que $b_j = \frac{\mu_{1j} - \mu_{0j}}{v_j}$.

Seja $H_j$, $j$ = 1, …, $p$ a hipótese de que o $j$-ésimo gene não apresenta expressão diferencial nos dois grupos, $H_j^+$ a hipótese alternativa de que a expressão é superior no grupo com cancro e $H_j^-$ a hipótese de que é inferior, estas três hipóteses correspondem a $b_j = 0$, $s_j = 1$ e $s_j = -1$, respectivamente.

Seja $z_j$ o valor $p$ correspondente ao teste de $H_j$, Efron et al. (2001) modelaram $z_j$, $j$ = 1, …, $p$ como gerado por uma distribuição de mistura $\eta_0 f_0(z) + (1-\eta_0) f_1(z)$, onde $\eta_0$ é a proporção de genes sem expressão diferencial, $f_0$ é a função de densidade uniforme em [0, 1] para valores $p$ sob a hipótese nula e $f_1$ é a função densidade sob a hipótese alternativa.

A *local false discover rate* é dada por:

$$fdr_j = p(H_j \text{ é verdadeira} \mid z_j) = \frac{\eta_0}{\eta_0 + (1-\eta_0) f_1(z_j)} \tag{2.6}$$

e pode ser estimada usando o método aplicado em Liao et al. (2004).

Para modelar o subconjunto $M$, este pode ser gerado como um conjunto aleatório tal que cada $j$ tem uma probabilidade 1- $\hat{fdr}_j$ de estar em $M$, independentemente para $j$ = 1, …, $p$.

Liao & Chin (2007) atribuíram a $s_j$, $j \in M$, os valores -1 ou 1 dependendo da direcção da diferença de expressão dos genes na amostra de modelação.

Para completar a especificação de $b_j$, modela-se $|b_j|$, $j \in M$, como efeitos aleatórios independentes de uma distribuição N (0, $\theta_0^2$) truncada em (0, ∞), em que $\theta_0$ quantifica o tamanho de $|b_j|$ e é estimado a partir dos dados.

Para construir uma amostra *bootstrap* $\mathbf{y}^* = (y_1^*, ..., y_n^*)$ segundo o método proposto por Liao & Chin (2007), estima-se a *local fdr* ( $\hat{fdr}_j$ ), estima-se $\theta_0$ ( $\hat{\theta}_0$ ) que maximiza (2.7) e seguem-se os passos descritos a seguir.

O primeiro passo consiste em gerar $M$ tal que cada $j$ tem uma probabilidade 1- $\hat{fdr}_j$ de pertencer a $M$; atribuir o valor 1 a $s_j$ caso a expressão do $j$-ésimo gene seja superior no grupo com cancro e o valor -1 caso contrário e gerar $|b_j|$ a partir de uma distribuição N(0, $\theta_0^2$) truncada em (0, ∞), para $j \in M$.

O segundo passo é construir y* a partir da distribuição logística condicionada (2.4) com π dado pelo modelo (2.5) gerado no primeiro passo.

Para estimar $\theta_0$ , seja $\tilde{\mathbf{y}}$ uma amostra *bootstrap* construída com $|b_j|$ gerado de uma N(0, $\theta^2$) truncada. Seja $h(\theta)$ a verosimilhança de $\theta | \tilde{\mathbf{y}} = \mathbf{y}$, em que a probabilidade incorpora ambos os passos na construção de $\tilde{\mathbf{y}}$ . $\hat{\theta}_0$ maximiza

$$\theta^{-1/3} h(\theta) \tag{2.7}$$

onde $\theta^{-1/3}$ é acrescentado à função de verosimilhança $h(\theta)$ para melhorar a estabilidade da maximização.

Seja $q$ o número de genes mais significantes incluído no modelo de regressão logística penalizada e $\tau$ o parâmetro de contracção, para cada par ($q$, $\tau$) o erro de predição pode ser estimado começando por construir uma amostra *bootstrap* $\mathbf{y}^* = (y_1^*,...,y_n^*)$ seguindo o procedimento descrito anteriormente. Dados os valores gerados para *M*, $s_j$ e $|b_j|$, escolhe-se $b_0$ de modo a que (2.5) satisfaça $\pi_1 + ... + \pi_n = n_1$, que é a estimativa de máxima verosimilhança de $b_0$. Deste modo, a partir de (2.5), $\pi_1,...,\pi_n$ ficam especificados. Segue-se a aplicação das duas etapas do desenvolvimento do modelo de regressão logística à amostra *bootstrap* $\mathbf{y}^*$. $\hat{\pi}^* = (\hat{\pi}_1,...,\hat{\pi}_n)$ estima $\pi = (\pi_1,...,\pi_n)$ a partir do modelo desenvolvido. O *Brier score* esperado é dado por:

$$n^{-1} \sum_{i=1}^{n} \left\{ (\hat{\pi}_i^* - \pi_i)^2 + \pi_i (1 - \pi_i) \right\} \tag{2.8}$$

que é tanto menor quanto mais próximo do verdadeiro valor $\pi$ for o valor estimado $\hat{\pi}^*$. Repetindo este procedimento um grande número de vezes (cerca de 10000) e calculando a média dos valores obtidos para o *Brier score* esperado, obtém-se um valor estimado para o erro de predição.

## 2.2. ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO

As árvores de classificação e regressão, CART (*Classification and Regression Trees*) são uma das metodologias mais utilizadas nos estudos de *Data Mining* devido à simplicidade de interpretação e à boa capacidade de ajustamento que, em geral, oferecem. Para além disso, os métodos baseados em árvore, como as CART, têm a capacidade de revelar estruturas de interacções, o que os torna muito interessantes para investigadores de diversas áreas, entre as quais a Medicina e as Ciências Biomédicas. No âmbito da aplicação a *microarrays*, importa identificar como é que combinações de genes estão associadas a doenças específicas (Boulesteix & Tutz, 2006).

Podendo ser consideradas como modelos de regressão não-paramétricos, as CART têm como objectivo estabelecer uma relação entre o vector de variáveis independentes (covariáveis) e a variável resposta.

Nestas árvores, quer as variáveis explicativas quer a resposta podem assumir valores contínuos ou categóricos. Nos estudos de classificação supervisionada, como a variável resposta é categórica, o modelo designa-se árvore de classificação, caso contrário, designar-se-ia árvore de regressão.

Estes modelos são ajustados mediante sucessivas divisões binárias no conjunto de dados, de modo a tornar os subconjuntos resultantes cada vez mais homogéneos em relação à variável resposta. Essas divisões são representadas por uma estrutura de árvore

binária, na qual cada nó corresponde a uma divisão numa covariável particular. Cada nova divisão é escolhida de modo a minimizar um determinado critério. Um dos critérios mais utilizados é a desviância (*deviance*), também designada por *cross-entropy*. Para duas classes, se $p$ for a proporção na segunda classe, a desviância é dada por: $-p \log p - (1-p) \log (1-p)$ (Hastie et al., 2001).

As componentes básicas de uma árvore de classificação são os nós e as regras de divisão (*splitting rules*). Os nós estão associados aos subconjuntos resultantes da aplicação de uma regra de divisão ao conjunto de dados. O primeiro nó de uma árvore, que corresponde ao conjunto de dados completo, é chamado de nó raiz e os nós terminais denominam-se folhas. Os nós gerados pela divisão de um nó já existente designam-se descendentes e o nó que os originou é chamado de ascendente ou pai (Ferreira et al., 2001).

Quando o número de variáveis é elevado e o número de observações é pequeno, como acontece nos dados de *microarrays*, as CART apresentam frequentemente um pobre desempenho porque utilizam apenas uma pequena parte da informação disponível, sendo construídas com poucas divisões. Se pararmos o crescimento da árvore muito tarde, algumas divisões podem ser estatisticamente insignificantes. Se o processo parar muito cedo, as regras de divisão dependem de poucas variáveis e não utilizam a maioria das variáveis potencialmente interessantes (Boulesteix & Tutz, 2006).

Boulesteix & Tutz (2006) propõem um método baseado no algoritmo CART para encontrar padrões de interacção (IPs) em conjuntos de dados. Os padrões detectados podem ser usados para definir novas covariáveis com o objectivo de melhorar o desempenho dos métodos de classificação.

Os IPs têm a forma $\{x_1 > \theta_1\} \cap \{x_2 \leq \theta_2\} \cap \cdots \cap \{x_d > \theta_d\}$, onde $x_1, \ldots, x_d$ são covariáveis, $\theta_1, \ldots, \theta_d$ são estimados e $d$ é o número de covariáveis envolvidas. O método proposto permite identificar candidatos a padrões e seleccionar como IPs apenas aqueles que verificam determinado critério estatístico. É utilizado um critério de *pruning* para evitar IPs muito longos e irrelevantes. Uma versão mais simples do algoritmo proposto por Boulesteix & Tutz (2006), restrita ao caso de 2 classes, é dada em Boulesteix et al. (2003).

Utilizar árvores para encontrar IPs tem o problema da construção ser por partição recursiva, o que faz com que todos os nós se dividam segundo as mesmas variáveis. Em particular, todas as folhas têm origem na mesma divisão da raiz, o que faz com que padrões que não envolvam a variável que divide a raiz nunca sejam encontrados. O algoritmo proposto por Boulesteix & Tutz (2006) é baseado no crescimento de várias árvores que usam diferentes variáveis a partir das quais a divisão inicia. Cada vez que é construída uma árvore, armazenam-se as folhas e elimina-se a variável que definiu a primeira divisão. Repete-se este procedimento até não haver mais nenhuma covariável. Os candidatos a padrões, cuja relevância é posteriormente analisada, são as folhas das árvores construídas.

As Árvores de Classificação foram construídas recorrendo à biblioteca *tree* do *R*.

## 2.3. CLASSIFICADOR DOS VIZINHOS MAIS PRÓXIMOS

O classificador dos *k* vizinhos mais próximos (*k-Nearest Neighbor*, *k*-NN) é um método de classificação baseado numa função de distância entre pares de observações, sendo a distância euclideana a mais usual. Este método foi aplicado por Deegalla & Boström (2007) e Boulesteix & Tutz (2006), entre outros, na classificação supervisionada de dados de *microarrays*.

Sejam $\mathbf{x}_1 = (x_{11}, \ldots, x_{1p})$ e $\mathbf{x}_2 = (x_{21}, \ldots, x_{2p})$ duas observações *p*-dimensionais, a distância euclideana entre elas é dada por:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{p} (x_{1i} - x_{2i})^2} \tag{2.9}$$

Para uma nova observação *x*, são encontradas na amostra de modelação as *k* observações mais próximas $x_1, \ldots, x_k$. A classificação ($\hat{y}$) será a moda dos valores $y_1, \ldots, y_k$.

Os classificadores dos vizinhos mais próximos foram construídos recorrendo à biblioteca *class* do *R*.

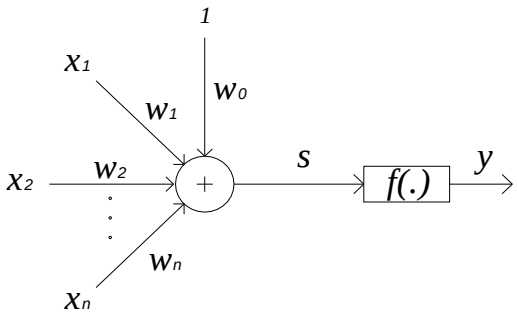
## 2.4. REDES NEURONAIS

As redes neuronais são uma técnica de inteligência artificial frequentemente utilizada em *Data Mining* e surgiram como uma tentativa de simular parcialmente o funcionamento do cérebro humano.

Acredita-se que o cérebro é composto por aproximadamente $10^{11}$ neurónios de muitos tipos diferentes. Os neurónios são compostos por um corpo central, onde se encontra o núcleo da célula, a partir do qual partem, ramificadamente, fibras nervosas chamadas dendrites. Do neurónio parte ainda uma longa fibra nervosa, chamada axónio, que por sua vez também se ramifica. No fim de cada ramificação encontram-se as sinapses, junções, através das quais se transmitem sinais deste para outros neurónios (Amaral, 1993).

O interesse em redes neuronais data do início da década de 1940, com o modelo proposto por McCulloch e Pitts para reproduzir as características básicas de um neurónio. O modelo é constituído por uma combinação linear de entradas ($\boldsymbol{x}$) produzidas por outras células, seguida de uma decisão binária, conforme ilustrado na figura 2.1 (Marques, 1999).

Figura 2.1 – Modelo de um neurónio.




O valor da saída, *y*, é dado por:

$$y = f(s) = f\left(w_0 + \sum_{i=1}^{n} w_i x_i\right) \tag{2.10}$$

em que $f(s)$ é a função de Heaviside:

$$f(s) = \begin{cases} 1, & se\ s \geq 0 \\ 0, & se\ s < 0 \end{cases} \tag{2.11}$$

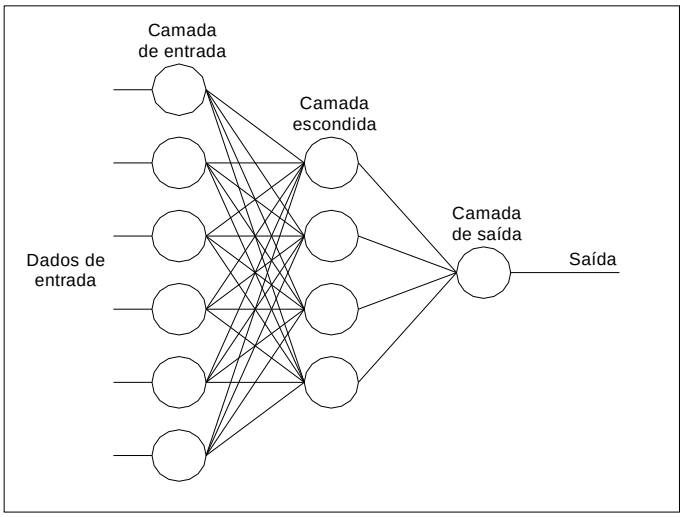
Os pesos (***w***) são constantes e conforme o seu sinal definem dois tipos de sinapses: sinapses de excitação, em que os pesos são positivos e todos de igual valor; e sinapses de inibição, sendo neste caso os pesos negativos e os seus valores tais que se alguma delas estiver activa o resultado do somatório é negativo. Isto é, nenhum neurónio está activo quando alguma das suas sinapses inibidoras se encontra activa.

Do ponto de vista das redes neuronais o seu resultado fundamental foi mostrar que elementos computacionais bastante simples, quando agregados num vasto sistema de interligações devidamente pesadas, eram capazes de realizar tarefas imensamente poderosas, e a sua apresentação teve grande repercussão na comunidade de investigadores da então nascente ciência da computação (Amaral, 1993).

O perceptrão é uma rede constituída por várias unidades interligadas. Frequentemente, as unidades estão organizadas em camadas. Assim, as entradas das unidades de uma camada são saídas das unidades da camada anterior. Cada unidade é descrita através do modelo de McCulloch e Pitts com função de activação binária ou contínua (sigmóide).

O perceptrão simples é uma rede com uma única camada, em que cada peso influencia uma única saída. As capacidades do perceptrão simples são limitadas a fronteiras de decisão linear, podendo-se ultrapassar esta limitação através do uso de camadas escondidas, o que permite obter regiões de decisão arbitrárias desde que se utilize um número suficiente de unidades em cada camada. Uma rede constituída por mais que uma camada denomina-se perceptrão de multicamada, conforme ilustrado na figura 2.2.

Figura 2.2 – Perceptrão de Multicamada.




A escolha da arquitectura que conduzirá a um melhor desempenho da rede, ou seja, do número de camadas, do número de unidades por camada e das ligações, não é fácil. Um bom ajustamento do classificador à amostra de modelação pode corresponder a um excesso de adaptação da rede aos padrões do treino.

No treino de uma rede, pretende-se evitar a sobreaprendizagem (*overfitting*), situação em que o desempenho da rede em dados não pertencentes ao conjunto de treino começa a degradar-se, embora o algoritmo de aprendizagem continue a garantir bons resultados no conjunto de modelação.

A estimação dos parâmetros de uma rede neuronal faz-se por um algoritmo recursivo designado por método de retropropagação do erro (Ripley, 1996).

Dado um conjunto de amostras, defina-se como medida de desempenho da rede uma função de custo:

$$C(\boldsymbol{w}) = \sum_{p} C^{p} \tag{2.12}$$

com

$$C^{p} = \sum_{i} E(y_i^{p}, d_i^{p}) \tag{2.13}$$

em que $y_i{}^p$ designa o valor da saída $i$ e $d_i{}^p$ o valor desejado dessa mesma saída, quando na entrada da rede é apresentado o padrão $p$ e E() designa a função de erro.

Ao estender a arquitectura da rede a um sistema multicamada é necessário que se considerem unidades de resposta não linear, pois, de outro modo, o sistema multicamada seria sempre equivalente a um sistema de uma camada apenas, não introduzindo portanto a capacidade de classificar padrões não linearmente separáveis. Devido ao facto das funções de activação serem diferenciáveis, também o é $C(\boldsymbol{w})$ em relação a cada peso. Deste modo poder-se-á empregar para método de minimização da função de custo o método de gradiente, pois é possível calcular a contribuição de cada peso para o erro total, calculando as derivadas parciais deste em relação a cada peso. A

regra delta generalizada não é mais que uma forma expedita e simples de calcular este gradiente, baseada na regra de derivação da função composta (Amaral, 1993).

Em cada época são dados pequenos passos na direcção em que $C(\boldsymbol{w})$ mais acentuadamente desce, modificando ligeiramente os diversos pesos na direcção oposta à do gradiente da função de custo da rede: o incremento $\Delta w_{ij}(n)$ do peso $w_{ij}(n)$ é proporcional ao gradiente da função de custo no espaço dos pesos:

$$\Delta w_{ij}(n) = w_{ij}(n+1) - w_{ij}(n) = -\eta\, g_{ij}(n) \tag{2.14}$$

em que $n$ designa o número da época, $\eta$ é uma constante positiva denominada passo de aprendizagem e $g_{ij}$ representa o gradiente da função de custo $C(\boldsymbol{w})$ em ordem ao peso $w_{ij}$

$$g_{ij}(n) = \frac{\partial C(n)}{\partial w_{ij}} \tag{2.15}$$

Para unidades lineares, a função de custo define uma superfície côncava com um único mínimo global e o processo de minimização da função segundo o gradiente garante que esse mínimo é encontrado. Para redes multicamada com funções de activação não lineares, o processo de convergência é muito mais complexo, dado que a função de custo define agora uma superfície mais sinuosa, onde podem ocorrer mínimos locais, máximos locais e pontos de sela (Amaral, 1993).

Quando a superfície da função de custo se caracteriza por vales estreitos e profundos, as técnicas de gradiente apresentam um desempenho baixo, devido às frequentes oscilações do vector gradiente. Uma forma simples de evitar este tipo de fenómenos é proceder a uma filtragem "passa-baixo" do vector de adaptação de pesos:

$$\Delta w_{ij}(n) = -\eta\, g_{ij}(n) + \alpha\, \Delta w_{ij}(n-1) \tag{2.16}$$

em que $\alpha$ representa o parâmetro de momento, o qual deve tomar um valor entre 0 e 1 de modo a garantir a estabilidade do processo (Amaral, 1993).

Dada a morosidade com que o processo de aprendizagem através do algoritmo de retropropagação se desenvolve, vários autores têm proposto métodos de aceleração deste algoritmo, integrando ou não técnicas clássicas da área da optimização numérica.

Sendo o treino das redes neuronais através do algoritmo de retropropagação dominado pela forma da superfície de erro no espaço dos pesos, os algoritmos de aceleração procuram de algum modo recolher informação útil sobre essa superfície, nomeadamente informação de 2ª ordem.

O objectivo da fase de aprendizagem do algoritmo de retropropagação aplicado a uma rede neuronal multicamada é o da modificação dos pesos da rede de modo a que as saídas da rede para cada padrão que lhe é apresentado coincidam com as saídas que são desejadas. Do ponto de vista da optimização numérica este processo pode ser interpretado como um problema clássico de optimização não linear, uma vez que envolve a modificação, de um modo sistemático, de um conjunto de variáveis independentes, os pesos da rede, com o objectivo de minimizar uma função de custo.

Neste artigo aplicaram-se redes neuronais recorrendo à biblioteca *nnet* do *R*. Em problemas de classificação binária, esta biblioteca usa uma função de custo de entropia relativa. Para acelerar a convergência do algoritmo de retropropagação implementado, é adoptado um método quasi-Newton de segunda ordem.

## 3. BASES DE DADOS

Os métodos de classificação descritos anteriormente foram aplicados a dois conjuntos de dados de *microarrays* utilizados na literatura e disponíveis na Internet. Em ambos, os níveis de expressão dos genes foram medidos utilizando *Affymetrix high-density oligonucleotide arrays*. Um dos conjuntos permite classificar os indivíduos segundo o tipo de leucemia e o outro permite distinguir os indivíduos com cancro do cólon dos saudáveis. A descrição pormenorizada destes dados encontra-se em Golub et al. (1999) e em Alon et al. (1999), respectivamente.

### 3.1. Leucemia

Golub et al. (1999) demonstraram que os dados de *microarrays* podem ser utilizados para classificar os doentes em dois grupos: ALL (Leucemia Linfoblástica Aguda) ou AML (Leucemia Mielóide Aguda). Desde então, a classificação de doenças a partir de dados de *microarrays* tem sido alvo de investigação intensa, sendo este um dos conjuntos de dados mais utilizados (Dudoit et al., 2002; Liao & Chin, 2007; entre outros). Este conjunto de dados consiste no nível de expressão de 7129 genes para 38 indivíduos, 27 (71%) ALL e 11 (29%) AML, que constituem a amostra de modelação, e para 34 indivíduos, 20 (59%) ALL e 14 (41%) AML, que constituem a amostra de validação.

### 3.2. Cancro do cólon

Os dados do cancro do cólon, apresentados em Alon et al. (1999) e utilizados também por Boulesteix & Tutz (2006) e Li et al. (2001), consistem no nível de expressão de 2000 genes para 62 indivíduos, dos quais 40 (65%) têm cancro do cólon e 22 (35%) são saudáveis. Optou-se por efectuar duas divisões deste conjunto de dados. A primeira consistiu na divisão aleatória da amostra em duas de 31 indivíduos cada, que constituem a amostra de modelação e a amostra de validação. A amostra de modelação inclui 10 (32%) indivíduos saudáveis e 21 (68%) com cancro do cólon e a amostra de validação inclui 12 (39%) saudáveis e 19 (61%) com cancro. A segunda divisão apresenta uma proporção de aproximadamente 2:1 (modelação:validação), tendo-se incluído 40 indivíduos na amostra de modelação, 15 (38%) saudáveis e 25 (62%) com cancro, e 22 na de validação, 7 (32%) saudáveis e 15 (68%) com cancro. Em ambas as divisões, a selecção dos indivíduos foi feita de forma aleatória e a proporção de

indivíduos com cancro na amostra de modelação e na amostra de validação foi deixada ao acaso.

## 4. PRÉ-PROCESSAMENTO

O primeiro método de pré-processamento dos dados aplicado foi o adoptado na generalidade dos artigos que analisam os dados de leucemia de Golub et al. (1999), entre os quais, Dudoit et al. (2002) e Liao & Chin (2007). Este método consiste em:

- *Thresholding*: considerando um limiar inferior para os níveis de expressão dos genes igual a 100 e um limiar superior igual a 16000;
- Filtragem: exclusão de genes com máx/mín $\leq 5$ e máx-mín $\leq 500$, onde máx e min referem-se respectivamente ao máximo e ao mínimo dos níveis de expressão de um gene em particular;
- Transformação logarítmica de base 10

A forma mais usual de lidar com a elevada dimensionalidade dos dados de *microarrays* é seleccionar apenas os genes com expressão diferencial. Para tal, pode-se recorrer ao teste t ou à análise de variância, consoante o número de grupos, ou, alternativamente aos testes não paramétricos de Wilcoxon/Mann-Whitney ou de Kruskall-Wallis. Neste artigo optou-se por aplicar o teste t e seleccionar os genes para os quais o valor observado da estatística de teste foi superior a 2. Esta selecção constitui o segundo método de pré-processamento referido no decorrer do artigo.

No conjunto de dados da leucemia, o primeiro método de pré-processamento conduziu a uma redução do número de genes de 7129 para 3571, que reduziu para 471 com a aplicação do segundo método.

No que diz à primeira divisão dos dados do cancro do cólon, o primeiro método reduziu de 2000 para 1224 o número de genes. Aplicando o segundo método aos 1224 genes obteve-se um conjunto com apenas 43 genes e aplicando-o aos 2000 genes originais obtiveram-se 60 genes.

Para os dados resultantes da segunda divisão, da aplicação do primeiro método aos 2000 genes resultaram também 1224. Aplicando o segundo método aos 2000 genes seleccionaram-se 86 genes e aplicando o segundo método aos 1224 seleccionaram-se 64 genes.

Para efeitos comparativos, os métodos de classificação supervisionada foram aplicados aos vários conjuntos de genes resultantes da etapa de pré-processamento.

## 5. RESULTADOS

Apresenta-se nesta secção uma síntese dos resultados obtidos através da aplicação dos quatro métodos de classificação descritos anteriormente (regressão logística penalizada, árvores de classificação, classificador dos vizinhos mais próximos e redes neuronais), antecedida ou não das duas etapas de pré-processamento descritas.

Os modelos de regressão logística penalizada foram construídos com $q = 20$ para os dados da leucemia e $q = 2$ para os dados do cancro do cólon por se terem revelado os mais parcimoniosos, quando comparados com outros valores entre 1 e 100.

Para efeitos comparativos, os genes revelados pelas árvores de classificação como mais importantes, foram utilizados para treinar redes neuronais.

No caso da leucemia, não foi possível classificar os indivíduos com base na quantificação da expressão dos 3751 genes através de redes neuronais por implicar a estimação de um número de pesos superior ao admissível pela livraria *nnet* do *R*. Pela mesma razão, não se aplicou este método aos conjuntos compostos por 2000 ou 1224 genes referentes ao cancro do cólon.

No que diz respeito ao classificador dos *k* vizinhos mais próximos, apresentam-se apenas os melhores resultados obtidos para valores de *k* entre 1 e 10.

### 5.1. Leucemia

A qualidade do ajustamento obtida através dos vários modelos pode ser comparada a partir das proporções de erros de classificação (*misclassification error rate*) obtidas na amostra de modelação, enquanto que a capacidade preditiva pode ser quantificada pela proporção de erros na amostra de validação. A tabela 5.1 sintetiza os resultados obtidos para as duas amostras de dados da leucemia, consoante o método de classificação e o de pré-processamento aplicado.

Tabela 5.1 – Proporção de erros de classificação, consoante o método de classificação e de pré-processamento aplicados.

| **Método de classificação** | | **Pré-proces-samento** | **Número de genes** | **Proporção de erros na amostra de modelação** | **Proporção de erros na amostra de validação** |
|---|---|---|---|---|---|
| Regressão logística penalizada | | Mét. 1 | 3571 ($q = 20$) | 0,0000 | 0,0294 |
| | | Méts. 1 e 2 | 471 ($q = 20$) | 0,0526 | 0,0588 |
| Árvore de classificação | | Mét. 1 | 3571 | 0,0000 | 0,0882 |
| | | Méts. 1 e 2 | 471 | 0,0526 | 0,2647 |
| Rede neuronal (Perceptrão simples) | | Mét. 1 e 2 | 471 | 0,0000 | 0,2059 |
| | | Mét. 1 | 1 | 0,0000 | 0,0882 |
| | | Mét. 1 | 2 | 0,1053 | 0,0882 |
| Rede neuronal (5un) | | Mét. 1 | 2 | 0,0000 | 0,2059 |
| Classificador vizinhos mais próximos | $k = 1$ | Mét. 1 e 2 | 471 | 0,0000 | 0,0882 |
| | $k = 4$ | Mét. 1 e 2 | 471 | 0,0526 | 0,0000 |
| | $k = 3$ | Mét. 1 | 3571 | 0,0000 | 0,0294 |

De um modo geral, os métodos de classificação aplicados apresentaram bom desempenho, especialmente quando aplicados às amostras com 3571 genes sujeitas apenas ao primeiro método de processamento.

A maioria dos modelos ajustou-se perfeitamente aos dados da amostra de modelação, tendo-se conseguido uma capacidade de generalização perfeita apenas com o classificador dos 4 vizinhos mais próximos aplicado ao conjunto de 471 genes.

Tendo em conta ambas as proporções de erro de classificação, pode-se afirmar que a regressão logística penalizada e o classificador dos 3 vizinhos mais próximos são os métodos com melhor desempenho, ajustando-se perfeitamente à amostra de modelação e cometendo menos de 3% de erros na amostra de validação. Contudo, importa referir que, ao contrário do que acontece com os classificadores dos vizinhos mais próximos, o ajuste de modelos de regressão logística penalizada demora várias horas.

A árvore de classificação ajustada à amostra de modelação com 3571 genes recorreu apenas ao gene X95735_at e obteve uma capacidade de ajustamento perfeita e uma capacidade preditiva muito boa, cometendo erros em cerca de 9% das classificações de indivíduos da amostra de validação. Perante a importância deste gene, treinou-se um perceptrão simples considerando-o como única variável explicativa, tendo-se registado proporções de erros idênticas às obtidas com a árvore de classificação. Este gene não está incluído no conjunto dos 471 com expressão diferencial, sendo os genes M92287_at e D13627_at os que se revelaram com maior poder de discriminação dos dois tipos de leucemia. Contudo, não revelaram proporções de erros de classificações tão baixas quanto as obtidas com o gene X95735_at, quer nas árvores de classificação, quer no perceptrão simples. Uma vez que o perceptrão simples não se ajustou perfeitamente à amostra de modelação, observando-se uma proporção de erros de 0,1053, treinou-se uma rede neuronal com 5 unidades na camada escondida. Apesar de se ter obtido um ajuste perfeito à amostra de modelação, regista-se a ocorrência de *overfitting*, tendo a proporção de erros na amostra de validação aumentado de 0,0882 para 0,2059.

### 5.2. Cancro do cólon

A tabela 5.2 permite comparar a qualidade do ajustamento e a capacidade preditiva dos vários modelos ajustados aos dados do cancro do cólon, utilizando as amostras resultantes da primeira divisão efectuada (31 indivíduos na amostra de modelação e 31 na de validação).

Tabela 5.2 – Proporção de erros de classificação, consoante o método de classificação e de pré-processamento aplicados (cancro do cólon – primeira divisão).

| **Método de classificação** | | **Pré-proces-samento** | **Número de genes** | **Proporção de erros na amostra de modelação** | **Proporção de erros na amostra de validação** |
|---|---|---|---|---|---|
| Regressão logística penalizada | | Nenhum | 2000 ($q$ = 2) | 0,0645 | 0,1935 |
| | | Mét. 1 | 1224 ($q$ = 2) | 0,3226 | 0,3871 |
| | | Mét. 2 | 60 ($q$ = 2) | 0,0645 | 0,2581 |
| | | Méts. 1 e 2 | 43 ($q$ = 2) | 0,3226 | 0,3871 |
| Árvore de classificação | | Nenhum | 2000 | 0,0526 | 0,3226 |
| | | Mét. 1 | 1224 | 0,0645 | 0,3226 |
| | | Mét. 2 | 60 | 0,0645 | 0,3226 |
| | | Méts. 1 e 2 | 43 | 0,0645 | 0,3226 |
| Rede neuronal (Perceptrão simples) | | Mét. 1 | 2 | 0,0968 | 0,1613 |
| | | Mét. 2 | 60 | 0,1935 | 0,3226 |
| | | Méts. 1 e 2 | 43 | 0,0000 | 0,3871 |
| Rede neuronal com 5 unidades na camada escondida | | Mét. 1 | 2 | 0,0000 | 0,2903 |
| | | Mét. 2 | 60 | 0,1290 | 0,3871 |
| | | Méts. 1 e 2 | 43 | 0,0323 | 0,3548 |
| Classificador vizinhos mais Próximos | $k$ = 1 | Nenhum | 2000 | 0,0000 | 0,1481 |
| | $k$ = 1 | Mét. 1 | 1224 | 0,0000 | 0,1935 |
| | $k$ = 1 | Mét. 2 | 60 | 0,0000 | 0,2258 |
| | $k$ = 1 | Mét. 1 e 2 | 43 | 0,0000 | 0,2258 |

Conforme referido por Boulesteix & Tutz (2006), é difícil obter para os dados do cancro do cólon resultados tão bons quanto os obtidos para os dados da leucemia.

Os modelos de classificação ajustados aos dados revelaram fraca capacidade preditiva, variando as proporções de erro na amostra de validação entre os 0,1481 obtidos com a aplicação do classificador do vizinho mais próximo aplicado ao conjunto de 2000 genes e os 0,3871 obtidos para vários métodos.

O modelo de regressão logística penalizada não se ajustou bem aos dados pré-processados com o método 1. Tendo por base a expressão dos 1224 ou dos 43 genes, todos os indivíduos foram classificados como tendo cancro do cólon, quer na amostra de modelação, quer na amostra de validação. Este facto sugere que o método 1 pode não ser adequado a este conjunto de dados.

As árvores de classificação apresentaram desempenhos muito semelhantes independentemente do método de pré-processamento aplicado.

Quer aplicadas ao conjunto dos 2000 genes, quer aplicadas ao conjunto dos 1224 (resultantes do primeiro método de pré-processamento), as árvores de classificação revelaram o poder explicativo de dois genes: T72863 e R87126. Curiosamente, um

deles foi seleccionado na regressão logística penalizada aplicada ao conjunto dos 2000 e o outro foi seleccionado no conjunto dos 60. Comparativamente às árvores de classificação, a rede neuronal treinada com estes dois genes apresentou uma proporção de erros inferior na amostra de validação (0,1613), contudo revelou um ajustamento um pouco pior à amostra de modelação (0,0968). O uso de 5 unidades na camada escondida não se mostrou favorável registando-se aumentos nas proporções de erros de classificação na amostra de validação.

Os conjuntos dos 60 e dos 40 genes com expressão diferencial não incluem o gene T72863, evidenciando-se a relevância do gene X15880 juntamente com o R87126. Contudo, o poder explicativo desta dupla é inferior ao da inicial.

Importa referir a perfeita capacidade de ajustamento e a razoável capacidade preditiva dos classificadores do vizinho mais próximo.

Os resultados obtidos utilizando uma amostra de modelação de 40 e uma amostra de validação de 22 indivíduos (segunda divisão dos dados do cancro do cólon) estão expressos na tabela 5.3.

Tabela 5.3 – Proporção de erros de classificação, consoante o método de classificação e de pré-processamento aplicados (cancro do cólon – segunda divisão).

| **Método de classificação** | | **Pré-proces-samento** | **Número de genes** | **Proporção de erros na amostra de modelação** | **Proporção de erros na amostra de validação** |
|---|---|---|---|---|---|
| Regressão logística penalizada | | Nenhum | 2000 ($q$ = 2) | 0,1000 | 0,3636 |
| | | Mét. 1 | 1224 ($q$ = 2) | 0,0750 | 0,6818 |
| | | Mét. 2 | 86 ($q$ = 2) | 0,1250 | 0,1364 |
| | | Méts. 1 e 2 | 64 ($q$ = 2) | 0,3750 | 0,6818 |
| Árvore de classificação | | Nenhum | 2000 | 0,0750 | 0,4091 |
| | | Mét. 1 | 1224 | 0,0750 | 0,4091 |
| | | Mét. 2 | 86 | 0,0750 | 0,4091 |
| | | Méts. 1 e 2 | 64 | 0,0750 | 0,4091 |
| Rede neuronal (Perceptrão simples) | | Mét. 1 | 3 | 0,0750 | 0,4091 |
| | | Mét. 2 | 86 | 0,0000 | 0,5909 |
| | | Méts. 1 e 2 | 64 | 0,0250 | 0,3636 |
| Rede neuronal com 5 unidades na camada escondida | | Mét. 1 | 3 | 0,0500 | 0,3636 |
| | | Mét. 2 | 86 | 0,0250 | 0,3636 |
| | | Méts. 1 e 2 | 64 | 0,0500 | 0,3636 |
| Classificador vizinhos mais Próximos | $k$ = 1 | Nenhum | 2000 | 0,0000 | 0,2273 |
| | $k$ = 1 | Mét. 1 | 1224 | 0,0000 | 0,1818 |
| | $k$ = 1 | Mét. 2 | 86 | 0,0000 | 0,2727 |
| | $k$ = 1 | Mét. 1 e 2 | 64 | 0,0000 | 0,1818 |

Comparativamente aos resultados expressos na tabela 5.2, verifica-se que a maioria dos métodos apresentou melhor ajustamento (menores proporções de erro na amostra de modelação) e pior capacidade preditiva (maiores proporções de erro da amostra de validação). Este facto era espectável uma vez que tendo a amostra de modelação uma dimensão superior (quase o dobro da de validação) existe maior risco de *overfitting*.

Tal como com a divisão anterior, o modelo de regressão logística penalizada não apresenta bom desempenho quando aplicado aos dados pré-processados com o método 1 (exclusivamente ou e conjunto com o método 2), tendo classificado como normais todos os indivíduos da amostra de validação. Contudo, o modelo com melhor ajustamento e com a segunda melhor capacidade preditiva, foi o classificador dos vizinhos mais próximos, considerando apenas um vizinho, aplicado justamente aos dados processados pelo método 1 (exclusivamente ou não).

A melhor capacidade preditiva foi apresentada pelo modelo de regressão logística penalizada ajustado aos dados processados pelo método 2. Este modelo apresentou proporções de erro muito próximas na amostra de modelação (0,1250) e na de validação (0,1364).

As árvores de classificação evidenciaram o poder explicativo de 3 genes: R87126, U14973 e R80612. Destes genes, apenas o primeiro foi seleccionado pelo método 2 de pré-processamento e, para além disso, foi o único que também se revelou importante na classificação dos dados resultantes da primeira divisão. Utilizando apenas estes 3 genes, a rede neuronal apresenta um desempenho semelhante ao da árvore de classificação. O uso de 5 unidades na camada escondida traduziu-se numa melhoria na capacidade preditiva das redes neuronais.

## 6. DISCUSSÃO E CONCLUSÃO

A classificação supervisionada de dados de *microarrays* constitui um desafio, levantando diversos problemas de operacionalização dos métodos geralmente aplicados a outros tipos de dados, por apresentar um número de variáveis muito grande.

A regressão logística, apresenta comparativamente às restantes metodologias, a vantagem de haver uma teoria matemática bastante sólida por detrás da sua fundamentação, o que permite obter estimadores óptimos para os coeficientes do modelo. Construir um modelo de regressão logística usando dados de *microarrays* requer alterações técnicas consideráveis por se trabalhar com um valor de $p$ alto e um valor de n baixo, sendo necessária uma estrutura adicional para $b_0$, $b_1$, …$b_p$. A abordagem proposta por Liao & Chin (2007) conduz a bons resultados, contudo, revelou-se computacionalmente muito intensa e morosa. Apesar de no método proposto por Liao & Chin (2007) se assumir independência, os genes são geralmente muito correlacionados, sendo este um aspecto a explorar em trabalhos futuros. Seguir uma abordagem Bayesiana para definir uma estrutura para os coeficientes do modelo de regressão logística será uma possibilidade a analisar.

As árvores de classificação apresentaram capacidades de ajustamento muito boas. Dispondo de um elevado número de variáveis é possível conseguir para a amostra de modelação uma separação das classes quase perfeita, recorrendo a poucas variáveis. Contudo, as árvores obtidas têm ramos curtos e nalguns casos, um fraco desempenho em novos conjuntos de dados. Este método permite construir modelos simples e fáceis de interpretar, sendo este um dos seus principais atractivos.

As redes neuronais são um método emergente e com bom desempenho, sendo de salientar a sua capacidade de formar fronteiras de decisão não lineares no espaço dos atributos. Contudo, o uso de camadas escondidas não se revelou vantajoso, conduzindo a problemas de sobreaprendizagem, o que sugere que para estes dados, as superfícies de separação das classes seja aproximadamente linear. O principal problema inerente à aplicação deste método é o facto do número de pesos a estimar nos dados de *microarrays* ultrapassar geralmente as capacidades computacionais da biblioteca *nnet* do *R*, sendo necessário proceder a uma redução prévia do número de variáveis.

O classificador dos vizinhos mais próximos, apesar de muito simples, revelou-se um método de classificação supervisionada com excelente desempenho, tendo-se evidenciado positivamente quer para os dados da leucemia, quer para os do cancro do cólon.

Para todas as amostras utilizadas, algum dos genes com maior poder explicativo nos modelos ajustados não faz parte do conjunto de genes com expressão diferencial, não sendo retido com o segundo método de pré-processamento. Este facto permite-nos constatar que, conforme referido por outros autores (Jaeger et al., 2003), ao seleccionar os genes com expressão diferencial pode-se estar a perder genes com elevada capacidade explicativa e, para além disto, os genes seleccionados podem apresentar correlações elevadas, não se resolvendo o problema da multicolinearidade típico dos dados de *microarrays*. Estes problemas indiciam que não seja este o procedimento mais adequado para lidar com a elevada dimensionalidade dos dados. Como continuação deste trabalho, está planeado ensaiar e avaliar a aplicação de técnicas multivariadas de redução de dimensionalidade, entre as quais, a análise de componentes principais robusta.

Observou-se que os resultados diferem consoante a divisão efectuada nas bases de dados em amostra de modelação e amostra de validação, pelo que, pretende-se aplicar em trabalhos futuros o método da validação cruzada para evitar este problema.

Em trabalhos futuros pretende-se desenvolver, testar e comparar modelos de classificação supervisionada de séries temporais de dados de *microarrays*, sendo esta uma área emergente na genómica funcional, que constitui uma importante fonte de informação para o conhecimento dos processos biológicos e para o desenvolvimento de fármacos e terapêuticas eficientes. As especificidades presentes nestes conjuntos de dados requerem métodos especializados de selecção de genes, redução de dimensionalidade e classificação.

**Agradecimentos:** Este trabalho foi parcialmente subsidiado pelos projectos FCT/POCI 2010 e PTDC/MAT/64353/2006 e pela bolsa de doutoramento SFRH/BD/36606/2007.

## REFERÊNCIAS BIBLIOGRÁFICAS:


- Alon, U.; Barkai, N.; Notterman, D.; Gish, K.; Ybarra, S.; Mack, D. & Levine, A. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci.* USA, 96, 6745-6750.
- Amaral, J. (1993) *Métodos de aceleração do algoritmo de retropropagação*, Dissertação para obtenção do grau de mestre em Engenharia Electrotécnica e de Computadores, Instituto Superior Técnico.
- Ambroise, C. & McLachlan, G. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data, *PNAS*, 99 (10), 6562-6566.
- Antunes, M. & Sousa, L. (2008) Bayesian classification and non-bayesian label estimation via EM algorithm to identify differentially expressed genes: a comparative study, *Biometrical Journal*, 50 (5), 824-836.
- Asyali, M. & Alci, M. (2005) Reliability analysis of microarray data using fuzzy c-means and normal mixture modelling based classification Methods, *Bioinformatics*, 21, 644-649.
- Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, Series B, 57, 289 -300.
- Boulesteix, A.-L.; Strobl, C.; Augustin, T. & Daumer, M. (2008) Evaluating microarray-based classifiers: An overview, *Cancer Informatics*, 6, 77-97.
- Boulesteix, A.-L. & Tutz, G. (2006) Identification of interaction patterns and classification with applications to microarray data, *Computational Statistics & Data Analysis*, 50, 783-802.
- Boulesteix, A.-L.; Tutz, G. & Strimmer, K. (2003) A CART-based approach to discover emerging patterns in microarray data, *Bioinformatics*, 19, 18, 2465-2472.
- Braga-Neto, U. & Dougherty, E. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20 (3), 374-380.
- Branco, J.A. & Pires, A.M. (2009) Robust principal component analysis for high-dimensional data. Trabalho submetido.
- Deegalla, S. & Boström, H. (2007) Classification of microarrays with kNN: comparison of dimensionality reduction methods, *Intelligent Data Engineering and Automated Learning*, *LNCS*, 4881, 800-809.
- Díaz-Uriarte, R. & Andrés, S. (2006) Gene selection and classification of microarray data using random forest, *BMC Bioinformatics*, 7 (3), 1-13
- Dudoit, S.; Fridlyand, J. & Speed, T. (2002) Comparison of discrimination methods for the classification of tumours using gene expression data, *Journal of the American Statistical Association*, 97 (457), 77-87.
- Dupuy, A. & Simon, R. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting, *Journal of National Cancer Institute*, 99, 147-157.

- Efron, B. (2004) The estimation of prediction error: covariance penalties and cross-validation, *Journal of the American Statistical Association*, 99 (467), 619-632.
- Efron, B.; Tibshirani, R.; Storey, J. & Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association*, 96 (456), 1151-1160.
- Ferreira, C.; Soares, J. & Cruz, F. (2001) Reconhecimento de padrões em Estatística: Uma abordagem comparativa, *Proceedings of the V Brazilian Conference on Neural Networks*, 409-414.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.
- Hastie, T.; Tibshirani, R. & Friedman, J. (2009) *The elements of Statistical Learning: Data mining, inference and prediction*, 2nd ed., Springer, New York.
- Jaeger, J.; Sengupta, R. & Ruzzo, W. (2003) Improved gene selection for classification of microarrays, *Pacific Symposium on Biocomputing*, 8, 53-64.
- Kim, Y.; Kwon, S. & Song, S. (2006) Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data, *Computational Statistics & Data Analysis*, 51, 1643-1655.
- Lee, J.; Lee, J.; Park, M. & Song, S. (2005) An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics & Data Analysis*, 48, 869-885.
- Li, L.; Weinberg, R.; Darden, T., & Pedersen, L. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/kNN method, *Bioinformatics*, 17, 1131-1142.
- Li, T.; Zhang, C. & Ogihara, M. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics*, 20, 15, 2429-2437.
- Liao, J. & Chin, K.-V. (2007) Logistic regression for disease classification using microarray data: model selection in a large *p* and small *n*, *Bioinformatics*, 23, 1945-1951.
- Liao, J.; Lin, Y.; Selvanayagam, Z. & Shih, W. (2004) A mixture model for estimating the local false discovery rate in DNA microarray analysis, *Bioinformatics*, 20 (16), 2694-2701.
- Marques, J. (1999) *Reconhecimento de padrões: Métodos estatísticos e neuronais*, IST Press, Lisboa.
- Martella, F. (2006) Classification of microarray data with factor mixture models, *Bioinformatics*, 22, 2, 202-208.
- Nguyen, D. & Rocke, D. (2004) On partial least squares dimension reduction for microarray-based classification: a simulation study, *Computational Statistics & Data Analysis*, 46, 407-425.
- O'Neill, M. & Song, L. (2003) Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect, *BMC Bioinformatics*, 4: 13.

- Ooi, C. & Tan, P. (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Bioinformatics*, 19 (1), 37-44.
- Pirooznia, M. & Deng, Y. (2006) SVM classifier – a comprehensive Java interface for support vector machine classification of microarray data, *BMC Bioinformatics*, 7, Suppl 4, S25.
- Ripley, B. (1996) *Pattern recognition and neural networks*, Cambridge University Press.
- Roth, V. & Lange, T. (2004) Bayesian class discovery in microarray datasets, *IEEE Transactions on Biomedical Engineering*, 51, 5, 707-718.
- Satagopan, J. & Panageas, K. (2003) Tutorial in Biostatistics: a statistical perspective on gene expression data analysis, *Statistics in Medicine*, 22, 481-499.
- Shen, L. & Tan, E. (2005) Dimension reduction-based penalized logistic regression for cancer classification using microarray data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2, 2, 166-175.
- Simon, R.; Radmacher, M.; Dobbin, K. & McShane, L. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, *Journal of the National Cancer Institute*, 95 (1), 14-18.
- Stekel, D. (2003) *Microarray Bioinformatics*, Cambridge University Press.
- Zhu, J. & Hastie, T. (2004) Classification of gene microarrays by penalized logistic regression, *Biostatistics*, 5, 3, 427-443.