Método de classificação com rejeição por indecisão e observações atípicas.

Carla M. Santos Pereira
Universidade Portucalense Infante D. Henrique
Centro de Matemática e Aplicações do Instituto Superior Técnico

Ana M. Pires
Universidade Técnica de Lisboa, Instituto Superior Técnico
Departamento de Matemática e Centro de Matemática e Aplicações

Resumo: Num problema típico de classificação, o objectivo é criar uma regra de decisão que permita afectar um objecto, de origem desconhecida, a uma de c classes pré-definidas, a partir dos valores observados de um conjunto de p variáveis numa amostra de treino. Na impossibilidade de separação absoluta ou numa situação de dúvida (quando as funções de decisão assumem valores muito idênticos ou aquando da existência de observações atípicas- *outliers*) poderá ser preferível não classificar do que optar por classificar com uma probabilidade de erro elevada. Nesse caso introduz-se uma opção de rejeição, por indecisão ou por existência de observações atípicas, pelo que de uma forma genérica teremos um classificador em $c+2$ classes. Neste trabalho apresenta-se um método de classificação em $c+2$ classes com especial realce no tratamento das observações atípicas. Propõe-se uma nova regra de rejeição de *outliers*, RRO, baseada em análise de *clusters* e utilização de distâncias tipo Mahalanobis com estimadores clássicos e robustos que demonstrou ter bom comportamento em simulações de dados normais e não normais, com e sem *outliers*. Como métodos de clustering utilizaram-se o *k-means*, *pam* (*partitioning around medoids*) e *mclust* (*model based clustering*) e para estimadores do vector de médias e da matriz de covariâncias o RMCD25 (*Reweighted Minimum Covariance Determinant* com um ponto de rotura aproximado de 25%), os estimadores clássicos e o estimador OGK de Maronna e Zamar. O método apresentado é ilustrado com dois exemplos práticos.

Palavras–chave: Análise multivariada, Análise discriminante, Classificação supervisionada, Região de indecisão, Detecção de *outliers*, Estimadores robustos, Métodos de *clustering*, Modelos de misturas

Abstract: The aim of a supervised classification problem is to build a decision rule according to which a new object is assigned to one of a set of c predefined classes on the basis of an observed p-dimensional feature vector (training sample). In the absence of absolute separation or when there is some uncertainty it may be better not to classify. In that case we can introduce a rejection option either in cases of doubt or of atypical observations (outliers). This work presents a method for classifying a new object into one of $c+2$ classes. Special emphasis is given to the treatment of atypical observations: we propose a new outlier rejection rule, based on clustering analysis and Mahalanobis type distances with classical and robust estimators, which performed

well in a simulation study with normal and non-normal data, with and without outliers. We considered three clustering methods: *k-means*, *pam* and *mclust*; and three pairs of location-scatter estimators: classical, Reweighted Minimum Covariance Determinant with an approximate 25% breakdown point (RMCD25) and Orthogonalised Gnanadesikan-Kettenring estimator (OGK) of Maronna and Zamar. The method is illustrated with two applications.

Keywords: Multivariate analysis, Discriminant analysis, Supervised classification, Region of doubt, Detection of outliers, Robust estimators, Clustering methods, Mixture models

1 Introdução

Considere-se um problema típico de classificação supervisionada, onde o objectivo é criar uma regra de decisão que permita afectar um objecto, de origem desconhecida, a uma de c classes pré-definidas, a partir dos valores observados de um conjunto de p variáveis numa amostra de treino (isto é, formada por objectos de origem conhecida).

Dados um vector de características $\mathbf{x}= (x_1, x_2, ..., x_p)^T$ e um conjunto de c classes distintas $G_1, G_2, ..., G_c$, e designando por $d_i(\mathbf{x})$ uma função de decisão associada à classe G_i, tem-se o seguinte classificador genérico:

$$\mathbf{x} \in G_i \text{ se } d_i(\mathbf{x}) > d_j(\mathbf{x}), \forall j \neq i; i, j = 1, 2, ..., c.$$

Pretende-se alargar o problema de classificação a $c+2$ classes, $G_1, G_2, ..., G_c, I$, O, onde I representa a classe de indecisão e O a classe de *outliers*.

As componentes básicas de um problema de classificação no sentido deste estar definido de forma completa (por todos os elementos serem conhecidos) são:

- Probabilidades a *priori* das classes, designadas por π_i,
- f.d.p. condicionais às classes, $f_i(\mathbf{x})$, ou probabilidades *a posteriori* $P(G_i|\mathbf{x})$,
- Matriz de custos com elemento genérico $c(i|j)$, representando o custo de classificar erradamente na classe i um objecto de facto pertencente à classe j.

As funções de decisão, $d_i(\mathbf{x})$, são determinadas por estes elementos de acordo com o critério de decisão adoptado. Por exemplo, para o critério de maximização da probabilidade *a posteriori*, tem-se $d_i(\mathbf{x}) = P(G_i|\mathbf{x})$.

2 Método genérico

O método genérico de classificação em $c+2$ classes pode ser definido da seguinte forma:

(i) "Limpeza" da amostra de treino → Definição da classe O,

(ii) Construção da regra de decisão (com ou sem indecisão) → Definição das classes $G_1, G_2, ..., G_c, I$,

(iii) Classificação de um novo objecto de origem desconhecida como *outlier*, ou numa das classes $G_1, G_2, ..., G_c, I$.

Para definir a classe de *outliers* aplica-se uma regra de rejeição de *outliers* (RRO) a cada classe. Uma observação é classificada como *outlier* se e só se for *outlier* para todas as classes.

3 Rejeição por observações atípicas (RRO)

As ideias básicas da RRO, que demonstrou ter bom comportamento em simulações de dados normais e não normais, com e sem *outliers* (Santos Pereira e Pires, 2002) podem ser descritas em quatro passos:

(i) Segmentar as n observações (de cada classe) em k nuvens através da utilização de um método de *clustering* na esperança de que cada nuvem pareça "mais normal" do que a original.

(ii) Aplicar uma regra de detecção simultânea de *outliers* multivariados a cada nuvem através do cálculo de distâncias de Mahalanobis de cada observação a cada nuvem. Uma observação é considerada *outlier* se é *outlier* para todas as nuvens.
Todas as observações de uma nuvem podem também ser consideradas *outliers* se o tamanho relativo da nuvem é pequeno (propomos $n < 2p + 2$ já que para um número menor de observações as estimativas da matriz de covariâncias não são fiáveis).

(iii) Remover as observações detectadas em 2. e repetir 1. e 2. até que não sejam detectadas mais observações.

(iv) A decisão final sobre considerar todas as observações duma dada nuvem como *outliers* é baseada numa tabela de distâncias tipo Mahalanobis.

Uma regra de detecção simultânea de *outliers* multivariados é tal que para uma amostra normal multivariada de dimensão n_j, a probabilidade de não existirem observações detectadas como sendo *outliers* é $1 - \alpha_j$ para $j = 1, 2, ..., k$ (Davies e Gather, 1993). Se $\alpha_j = 1 - (1 - \alpha)^{\frac{1}{k}}$ pode ser garantido um nível de significância global α para misturas de k distribuições normais multivariadas.
Um observação x tal que

$$d^2 = (\mathbf{x} - \hat{\mu}_j)^T \hat{\Sigma}_j^{-1} (\mathbf{x} - \hat{\mu}_j) \geq c(p, n_j, \alpha_j)$$

é considerada *outlier* relativamente à j-ésima nuvem.

A constante $c(p, n_j, \alpha_j)$ é assintoticamente $\chi^2_{p;\beta}$ onde $\beta = (1-\alpha_j)^{1/n_j}$. No entanto, para amostras de dimensão não muito grandes e dependendo dos estimadores $\hat{\mu}_j$ e $\hat{\Sigma}_j$, podem existir grandes diferenças relativamente aos valores assintóticos. Becker e Gather (2001) sugerem que se recorra à simulação para obtenção de constantes mais fiáveis.

No passo 4, sejam $\mathbf{x}_{ij}$ as observações finais, onde os n_j representam os tamanhos das k nuvens finais (note-se que $\sum_j n_j$ é em geral menor do que n) e defina-se:

$$D_{lm} = \min_{i=1,\ldots,n_l} (\mathbf{x}_{il} - \hat{\mu}_m)^T \hat{\Sigma}_m^{-1} (\mathbf{x}_{il} - \hat{\mu}_m), l, m = 1, ..., k.$$

Diz-se que uma dada nuvem, l, é desligada dos restantes dados se $D_{lm} > c(p, n_m, \alpha_m)$, para todo o $m \neq l$ e ligada caso contrário. Nuvens desligadas são suspeitas de conterem apenas *outliers*. No entanto, a decisão final relativamente à rejeição destas observações, deve depender do tipo de dados.

4 Rejeição por indecisão

Considere-se o caso particular de duas classes. Conhecidas as componentes básicas do problema de classificação, utilizando a regra de Bayes para minimização do risco com opção de rejeição (Santos Pereira e Pires, 2001) e considerando $c(i|i) = 0$, $c(i|j) = 1$ e $c(I|j) = t$ para $i, j = 1, 2$, obtém-se a seguinte regra de decisão:

$$\mathbf{x} \in G_1 \quad \text{se} \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{1-t}{t} \times \frac{\pi_2}{\pi_1} = t_1$$

$$\mathbf{x} \in G_2 \quad \text{se} \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \leq \frac{t}{1-t} \times \frac{\pi_2}{\pi_1} = t_2$$

$$\mathbf{x} \in I \quad \text{se} \quad t_2 < \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < t_1.$$

Estimação

Até agora admitiu-se que as probabilidades *a priori*, o custo de rejeição t e as densidades condicionais às classes eram conhecidas. No entanto, tanto nos exemplos apresentados como na maior parte das situações reais este não é o caso. Propomos que se proceda da seguinte forma:

- Probabilidades a priori : $\hat{\pi}_i = \frac{n_i}{n}$ onde n_i representa o número de objectos da amostra de treino que pertencem à classe G_i, $i = 1, 2$.

- Podem ser experimentados vários valores para o custo t. Note-se que $0 \leq t \leq 0.5$. Se $t = 0.5$ cai-se no caso usual de inexistência de rejeição.

- As densidades, ou equivalentemente, as probabilidades *a posteriori* de cada classe, podem ser estimadas por vários processos (discriminante linear, quadrática ou logística, redes neuronais, árvores de classificação, etc.).

Uma proposta alternativa consiste na estimação da densidade $f_i(\mathbf{x})$, através de um modelo de mistura de normais com k_i componentes correspondentes às nuvens fixadas na *i-ésima* classe, da seguinte forma:

$$\hat{f}_i(\mathbf{x}) = \sum_{j=1}^{k_i} \frac{n_{ij}}{n_i} f_{N(\hat{\boldsymbol{\mu}}_{ij}, \hat{\boldsymbol{\Sigma}}_{ij})}(\mathbf{x}) \quad , \quad \forall i = 1, 2, \ldots, c. \qquad (1)$$

5 Exemplos

Os conjuntos de dados utilizados foram os seguintes:

Exemplo 1: Mulheres portadoras de Hemofilia (Johnson e Wichern, 1992)
Amostra: 105 mulheres;
Variáveis: 2 obtidas em análises ao sangue e relacionadas com a capacidade de coagulação: Factor VIII- actividade (x_1) e Factor VIII- antigene (x_2);
Classes:
G_1 - normais (60 mulheres)
G_2 - portadoras de hemofilia (45 mulheres).

Exemplo 2: Dados simulados
Amostra: 320 observações;
Variáveis: 2;
Classes:
G_1 - 50 observações $t_{2,3}(\mu_1, \Sigma_1)$, 50 observações $t_{2,3}(\mu_2, \Sigma_2)$ com $\mu_1 = (0, 12)^T$, Σ_1 =diag$(1, 0.3)$, $\mu_2 = (1.5, 6)^T$ e Σ_2 =diag$(0.2, 9)$ e 20 observações *outliers* $t_{2,3}((-2, 6)^T, 0.01\mathbf{I})$, onde $t_{2,3}$ representa a distribuição t-Student bivariada com três graus de liberdade.
G_2 - 100 observações $N_2((7, 5)^T, 0.5\ \mathbf{I})$ e 100 observações $N_2((2, 3)^T, \Sigma_1)$ com Σ_1 =diag$(0.2, 9)$

Com o propósito de obtenção da regra de classificação foram feitas as seguintes escolhas:

→ Três métodos de *clustering*: $k-means$, *pam* (*partition around medoids* de Kaufman e Rousseeuw, 1990), *mclust* (*model based clustering* de Banfield e Raftery, 1992), cada um deles com $k = 3, 4, 5$.

→ Três estimadores de localização-dispersão: clássicos ($\overline{\mathbf{x}}$, $\mathbf{S}$) com limites de detecção assintóticos; RMCD25 (*Reweighted Minimum Covariance Determinant* com ponto de rotura aproximado de 25%, Rousseeuw e van Driessen, 1999) com limites de detecção determinados previamente através de uma simulação com 1000 conjuntos de dados normais; estimador OGK $_{(2)}(0.9)$ (*orthogonalised*

Gnanadesikan-Kettenring com duas iterações e reponderado a 90%- Maronna e Zamar, 2002) com limites de detecção determinados previamente através de uma simulação com 1000 conjuntos de dados normais.

→ Nível de significância global: $\alpha = 0.1$.

→ Classificar as nuvens desligadas (no passo 4) como *outliers*.

→ Estimação das densidades através da mistura de normais (1).

→ Critério de selecção para escolha da melhor combinação, na "limpeza" dos *outliers* (método de *clustering* /estimador/ k): AIC (McLachlan e Peel, 2000).

Resultados

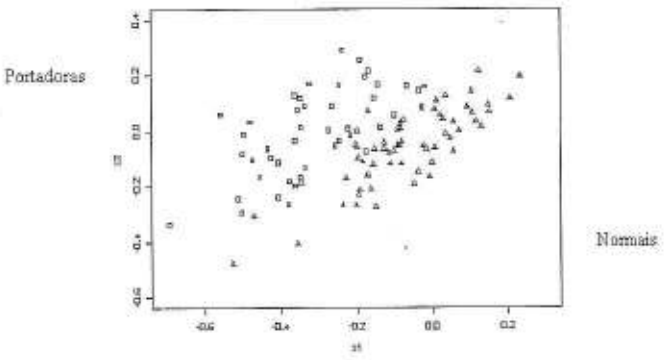

Figura 1: Dados do Exemplo 1.

A utilização do critério AIC conduziu à escolha daquela que se passa a designar por "melhor" combinação. Apresenta-se ainda o número de *outliers* detectados para cada classe.

• "Melhor" combinação para o Exemplo 1:

G_1 : $k-means$ / estimador RMCD25 / $k = 4$ → "limpeza" de 3 *outliers*;

G_2 : $k-means$ / estimador clássico/ $k = 4$ → "limpeza" de 3 *outliers*.

Na Figura 2 apresenta-se, para cada classe, o conjunto de dados "limpos" com os respectivos contornos. Note-se que apenas aparecem três contornos porque a quarta nuvem tinha apenas três observações (classificadas como *outliers* e representadas com símbolos diferentes).

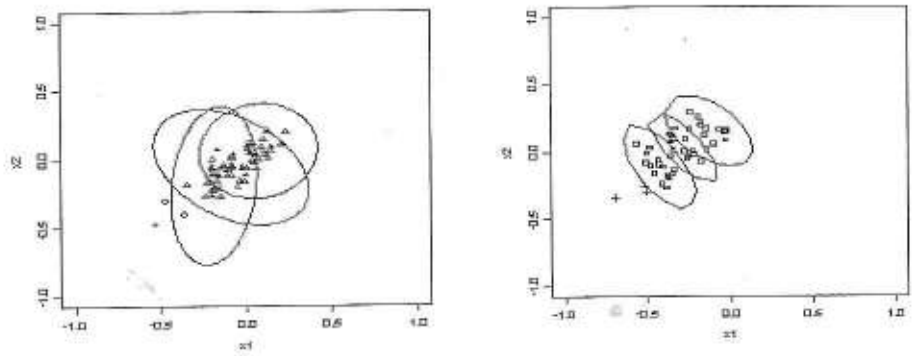

Figura 2: Dados do Exemplo 1 com contornos para a classe G_1 (à esquerda) e para a classe G_2 (à direita).

As regiões de classificação com indecisão, para dois custos diferentes, são apresentadas nas Figuras 3 e 4.

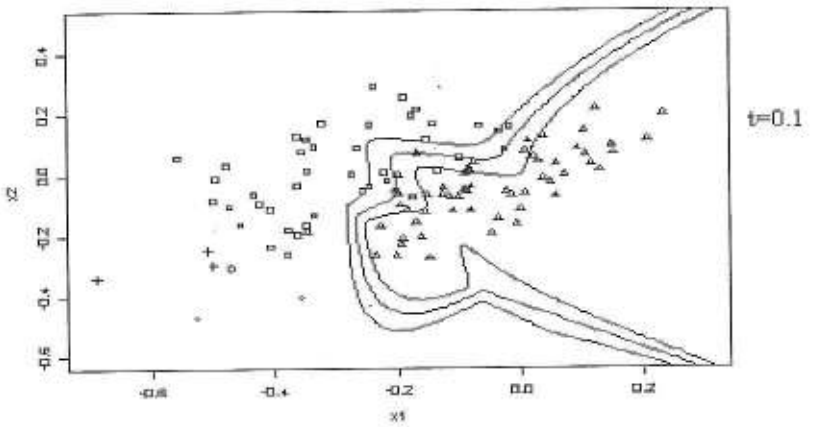

Figura 3: Regiões de classificação com custo de indecisão $t = 0.1$ para o Exemplo 1.

- "Melhor" combinação para o Exemplo 2:
G_1 : *mclust* / estimador clássico / $k = 3 \rightarrow$ "limpeza" de 21 *outliers*;
G_2 : *mclust* / estimador clássico/ $k = 2 \rightarrow$ não foram detectados *outliers*.

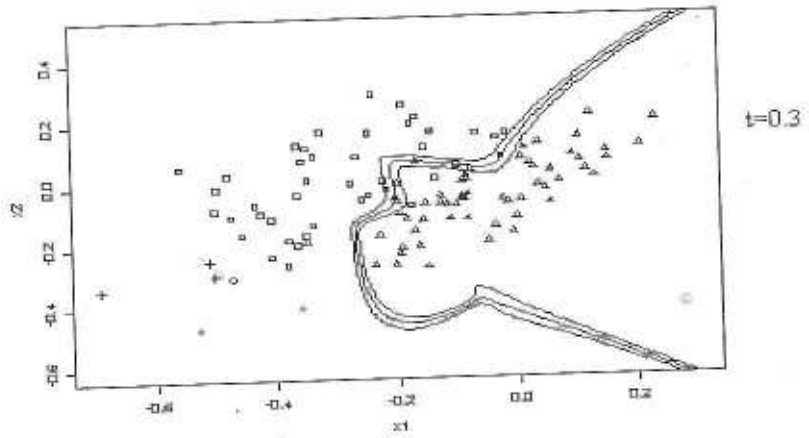

Figura 4: Regiões de classificação com custo de indecisão $t = 0.3$ para o Exemplo 1.

Na Figura 5 são apresentados os dados completos do Exemplo 2 e na Figura 6 apresentam-se as regiões de classificação com custo de indecisão $t = 0.1$ (os *outliers* aparecem com um símbolo diferente).

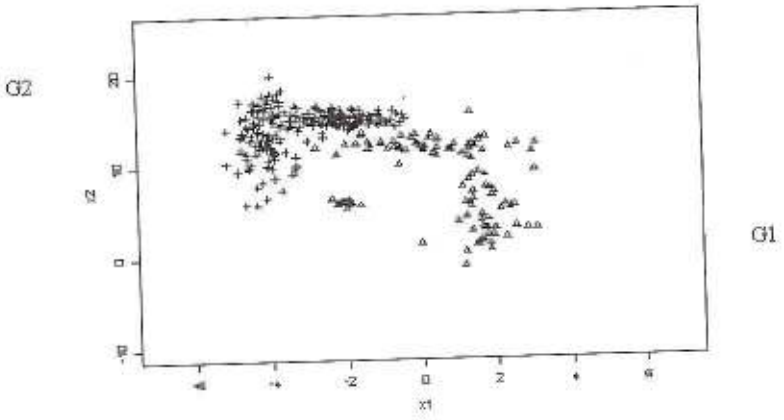

Figura 5: Dados do Exemplo 2.

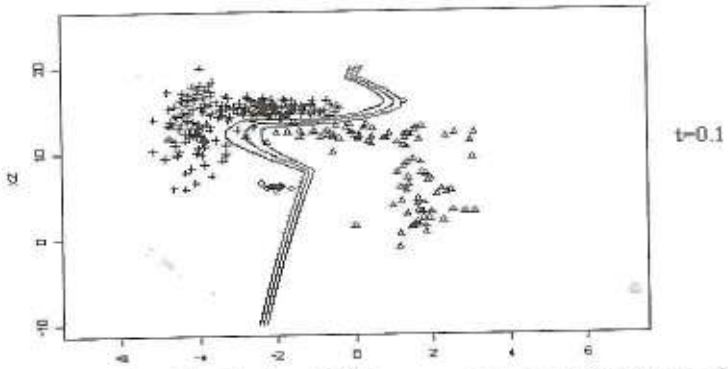

Figura 6: Regiões de classificação com custo de indecisão $t = 0.1$ para o Exemplo 2.

Bibliografia

[1] Banfield, J. e Raftery, A. (1992). *Model-based Gaussian and non-Gaussian clustering. Biometrics* 49, p. 803-822.

[2] Becker, C., e Gather, U. (2001). *The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. Computational Statistics and Data Analysis,* 36, p. 119-127.

[3] Davies, L. e Gather, U. (1993). *The Identification of multiple outliers. Journal of the American Statistical Society,* 88, p. 782-801.

[4] Johnson, R. e Wichern, D. (1992). *Applied Multivariate Statistical Analysis.* Prentice Hall. Englewood Cliffs.

[5] Kaufman, L. e Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis.* Wiley, New York.

[6] Maronna, R. e Zamar, R. (2002). *Robust estimates of location and dispersion for high-dimensional data sets. Technometrics,* 44, p. 307-317.

[7] McLachlan, G. e Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

[8] Rousseeuw, P. J. e Van Driessen, K. (1999). *A fast algorithm for the minimum covariance determinant estimator. Technometrics,* 41, p. 212-223.

[9] Santos-Pereira, C. M. e Pires, A. M. (2001). *Classificação supervisionada com dúvidas: compromisso erro/rejeição.* Em *A Estatística em Movimento: Actas do VIII Congresso Anual da Sociedade Portuguesa de Estatística.* M. Neves et al. (editores), p. 313-322. Edições SPE, Lisboa.

[10] Santos-Pereira, C. M. e Pires, A. M. (2002). *Detection of outliers in multivariate data: a method based on clustering and robust estimators. Proceedings in Computational Statistics. COMPSTAT 2002, Berlim.* Hardle, W. e Ronz, B. (editores), p. 291-296. Physica-Verlag, Heildelberg.