Anais X SBSR, Foz do Iguaçu, 21-26 abril 2001, INPE, p. 1217-1223, Sessão Técnica Oral - Workshops

Um classificador baseado na Discriminação Logística: vantagens e desvantagens

HÉLIO RADKE BITTENCOURT [1]
ROBIN THOMAS CLARKE [2]

[1]Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia – CEPSRM – UFRGS
Caixa Postal 15044 – CEP 91501-970 – Porto Alegre – RS, Brasil
heliorb@cpovo.net

[2]Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia – CEPSRM – UFRGS
Caixa Postal 15044 – CEP 91501-970 – Porto Alegre – RS, Brasil
clarke@if.ufrgs.br

Abstract: Logistic discrimination can be regarded as a partially parametric approach to pattern recognition. The technique is quite general and robust since it assumes nothing about the probability distribution of variables, and the number of parameters to be estimated is relatively small. Despite its generality and robustness, it is still not widely used for classifying digital images. This paper describes the logistic model and discusses its advantages and disadvantages; it also gives some results obtained when using it to classify Landsat-TM and AVIRIS images.

Keywords: digital image processing, logistic discrimination, pattern recognition, classifier.

1 Introdução

A regressão logística tornou-se uma técnica padrão, sobretudo na área médica, para relacionar um conjunto de variáveis independentes a uma única variável resposta binária. A extensão do modelo logístico para variáveis resposta politômicas (Hosmer e Lemeshow, 1989) possibilita a sua utilização na classificação de imagens digitais.

Vamos considerar w_1 , w_2 , ... , w_k as k classes presentes em uma imagem; X_1 , X_2 , ..., X_P o conjunto de variáveis independentes formado pelos contadores digitais dos pixeis nas p bandas espectrais e x_1 , x_2 ,..., x_p particulares valores das variáveis X_i. Na abordagem estatística para reconhecimento de padrões cada pixel da imagem é visto como um vetor p-dimensional e portanto, denotaremos um pixel i simplesmente por $\underline{x}_i$.

De acordo com o modelo logístico, a probabilidade de um dado pixel $\underline{x}_i$ pertencer a uma das classes w_j pode ser estimada diretamente por meio da seguinte expressão:

$$P(w_j / \underline{x}_i) = \frac{\exp(b_{0j} + b_j^T \underline{x}_i)}{1 + \sum_{c=1}^{k-1} \exp(b_{0c} + b_c^T \underline{x}_i)} \quad \text{onde,} \quad \underline{x} = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \text{M} \\ x_{pi} \end{bmatrix} \quad \underline{\beta}_j = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \text{M} \\ \beta_p \end{bmatrix}$$

O modelo logístico necessita de *k-1* vetores de parâmetros a serem estimados, correspondentes a *k-1* classes presentes na imagem. A *k*-ésima classe é assumida como base e, portanto, o logaritmo natural das razões entre as funções de probabilidade das classes w_j (*j=1,...,k-1*) e do nível base w_k são assumidas como sendo funções lineares:

$$\ln\left[\frac{P(w_j / \underline{x})}{P(w_k / \underline{x})}\right] = \beta_{0j} + \beta_j^T \underline{x}$$

McLachlan (1992) considera essa a suposição fundamental da abordagem logística e, por esse motivo, a chama de modelagem parcialmente paramétrica, porque apenas as razões entre as funções de probabilidade das classes estão sendo modeladas.

A utilização do modelo logístico para discriminação de classes é direta. Os parâmetros $\boldsymbol{\beta}$ serão estimados a partir de uma amostra de treinamento, caracterizando um classificador supervisionado. A regra de decisão para alocar um dado pixel $\underline{x}_i$ numa das classes w_j é muito simples: o pixel $\underline{x}_i$ será alocado na classe onde a probabilidade $P(w_j / \underline{x}_i)$ for mais alta.

O processo de estimação dos parâmetros em regressão logística está baseado na maximização da função de verossimilhança $L(x,\boldsymbol{\beta})$. Para apresentar a função de verossimilhança do modelo, temos de criar *k-1* variáveis *dummy*, as quais chamaremos de $y_1, y_2, \ldots, y_{k-1}$ que assumem o valor 1 se o pixel pertence à classe correspondente e zero em caso contrário.

$$L(x,B) = \prod_{i=1}^{n} \left(\frac{exp(g_1(\underline{x}_i))}{1 + \sum_{c=1}^{k-1} exp(g_c(\underline{x}_i))} \right)^{y_{1i}} \times \mathrm{K} \times \left(\frac{exp(g_{k-1}(\underline{x}_i))}{1 + \sum_{c=1}^{k-1} exp(g_c(\underline{x}_i))} \right)^{y_{k-1;i}} \times \left(\frac{1}{1 + \sum_{c=1}^{k-1} exp(g_c(\underline{x}_i))} \right)^{1 - y_{1i} - \mathrm{K} - y_{k-1;i}}$$

onde $g_j(\underline{x}_i) = \exp\left(\beta_{0j} + \beta_j^T \underline{x}_i\right)$

Tomando o logaritmo natural da função de verossimilhança, chegamos a uma expressão mais simples:

$$\ln L(x,B) = \sum_{i=1}^{n} y_{1i} g_1(\underline{x}_i) + y_{2i} g_2(\underline{x}_i) + \mathrm{K} + y_{(k-1)i} g_{k-1}(\underline{x}_i) - \ln\left(1 + \sum_{c=1}^{k-1} g_c(\underline{x}_i)\right)$$

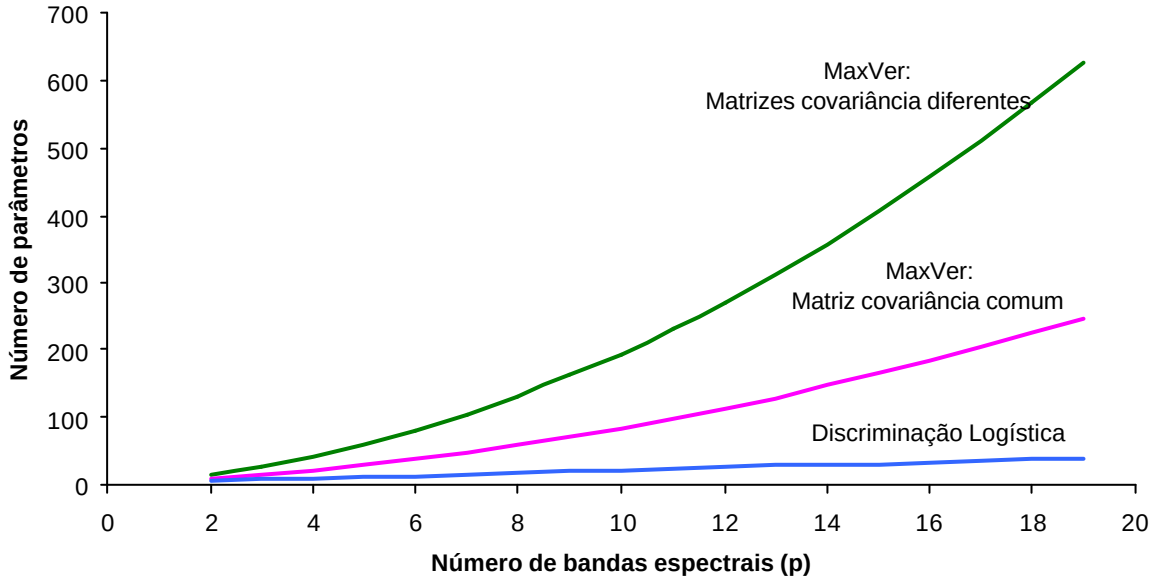
Os *k*-1 vetores de parâmetros $\boldsymbol{\beta}$ serão aqueles que maximizam o logaritmo da função de verossimilhança. Como a função é claramente não linear necessitamos da utilização de métodos numéricos para o processo de maximização. Esses processos são iterativos e estão disponíveis em alguns *softwares* estatísticos. No presente estudo utilizamos o procedimento CATMOD do sistema SAS que utiliza o método de Newton-Raphson, bastante rápido para convergência.

2 Vantagens

Vários autores têm apresentado as vantagens teóricas da discriminação logística, principalmente quando comparada ao método da máxima verossimilhança gaussiana. Efron (1975) diz que, ao menos teoricamente, a discriminação logística é mais robusta do que a análise discriminante gaussiana, sendo válida sobre uma grande variedade de distribuições. O fato de não necessitarmos da suposição de normalidade multivariada, torna a discriminação logística mais genérica. Press & Wilson (1978) e Krzanowaski (1988) consideram que há praticamente um consenso de que a discriminação logística deve ser preferida quando as distribuições são claramente não gaussianas.

Outra importante vantagem do modelo logístico é o reduzido número de parâmetros. Em regressão logística necessitamos de $(k-1)(p+1)$ para discriminação de k classes considerando as p bandas espectrais disponibilizadas pelo sistema sensor. Esse número é muito inferior ao necessário pelo método da Máxima Verossimilhança Gaussiana. A **figura 1** apresenta um comparativo entre o número de parâmetros necessários pela discriminação logística, o método da máxima verossimilhança gaussiana com matriz de covariância comum e a máxima verossimilhança com diferentes matrizes covariância.

Figura 1 – Número de parâmetros necessário para discriminação de 3 classes em função do número de bandas espectrais



É visível que, enquanto o crescimento do número de parâmetros na discriminação logística é linear, o número de parâmetros cresce quadraticamente na máxima verossimilhança gaussiana. Esse resultado transforma-se imediatamente numa vantagem da regressão logística, pois indica que é necessário um número menor de amostras de treinamento para o processo de estimação de parâmetros.

A regressão logística também apresenta uma vantagem na interpretação dos resultados, porque cada pixel terá uma probabilidade de pertencer a cada uma das classes. Sendo assim, interpretações como “o pixel *x* tem 70% de chance de pertencer a classe de vegetação” são perfeitamente possíveis.

Bittencourt e Clarke (2000) apresentaram resultados obtidos com a utilização da discriminação logística na classificação de imagens digitais, mostrando que é possível obter resultados semelhantes ou até melhores do que o método da máxima verossimilhança gaussiana.

3 Desvantagens

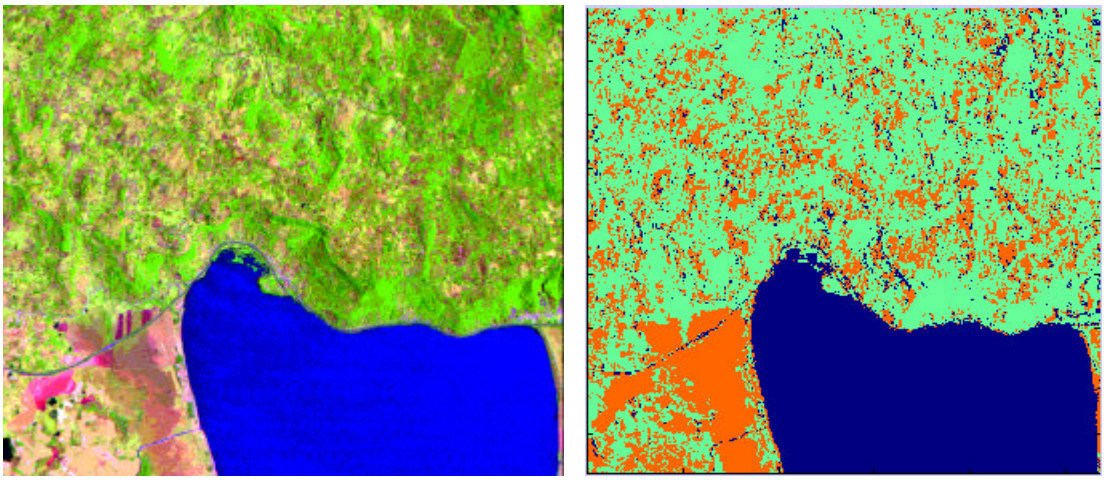
O primeiro problema em regressão logística consiste na obrigatoriedade da utilização de métodos numéricos para obtenção da solução de máxima verossimilhança. Métodos numéricos nem sempre convergem.

O segundo problema é crítico pois se refere a correlação existente entre as *p* bandas espectrais. O modelo logístico é bastante sensível à colinearidade, sendo que as conseqüências são erros padrão extremamente elevados que não permitem a realização de testes de significância para os coeficientes ***b***. Os testes realizados com imagens digitais mostraram que, mesmo com a parte inferencial sacrificada, é possível obter excelentes resultados com o classificador logístico.

4 Resultados

A classificação das imagens digitais foi realizada por rotinas elaboradas em MATLAB, a partir das estimativas dos parâmetros obtidas no SAS. Supomos que não hajam programas de processamento de imagens que implementem a discriminação logística. A **figura 2** apresenta uma imagem Landsat-TM composta de 436 linhas, 535 colunas e 6 bandas espectrais e a respectiva imagem classificada.

Figura 2 – Segmento de uma imagem Landsat-TM, composição colorida 5-4-2 (R-G-B) e imagem classificada por discriminação logística



As três classes consideradas para classificação foram: água, vegetação e culturas. A amostra de treinamento foi de pouco mais de mil pixeis, sendo que a imagem conta com mais de 230 mil

pixeis. Numa amostra de teste de 1992 pixeis, o percentual de acerto da discriminação logística foi de 99,6%. Visualmente, podemos perceber que os principais problemas ocorreram na classificação de uma estrada e em algumas áreas entre colinas que foram classificados como água.

O modelo estimado para classificação foi o seguinte:

$$P\left(w_j / \underline{x}_i\right) = \frac{e^{g_j(x)}}{1 + \sum_{c=1}^{k-1} e^{g_i(x)}}$$

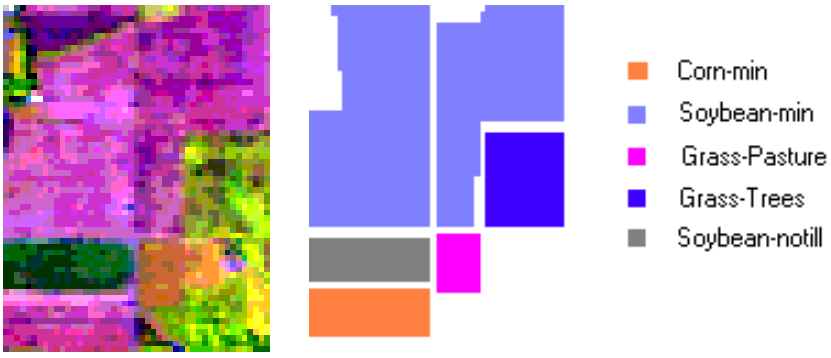
Classes: w_1=*Água*
w_2=*Culturas*
w_3=*Vegetação*

$$g_1(x) = 113.6 - 0.1581*x_1 + 1.1984*x_2 - 0.2760*x_3 - 1.7899*x_4 - 0.1227*x_5 + 0.1929*x_7$$
$$g_2(x) = 141.3 - 0.7992*x_1 + 2.3858*x_2 - 1.9673*x_3 - 3.1077*x_4 + 0.9979*x_5 + 0.1598*x_7$$
$$g_3(x) = 0$$

Na estimação dos parâmetros houve problema de colinearidade, mas apesar dos testes de significância não terem sido realizados para todos coeficientes $\boldsymbol{b}$ do modelo, o resultado da classificação foi bom.

A seguir testamos a discriminação logística para classificação de um segmento de uma imagem AVIRIS onde existem classes com um comportamento espectral médio muito semelhante e a verdade terrestre é conhecida. A **figura 3** apresenta a imagem.

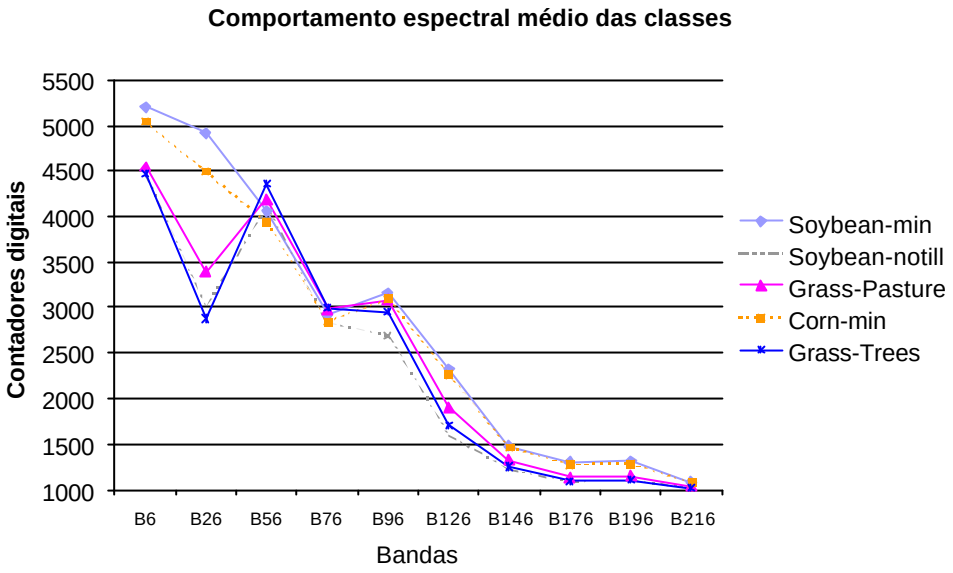
Figura 3 – Imagem AVIRIS – composição colorida 96-56-6 (R-G-B) e verdade terrestre



Apesar da imagem contar com 220 bandas espectrais, optamos por uma amostra sistemática de 10 bandas, o que nos levou a estimação de 44 parâmetros. A **figura 4** apresenta o comportamento espectral médio das cinco classes presentes na imagem.

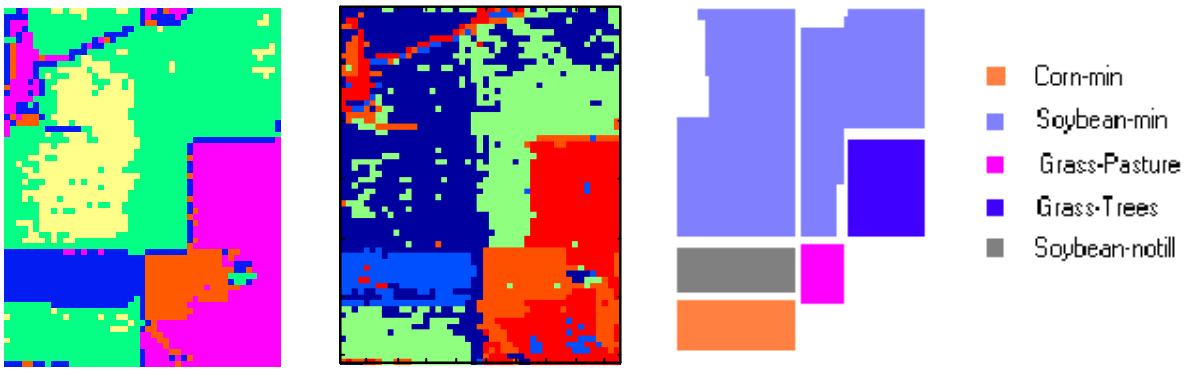
Fig. 04

Figura 4 – Comportamento espectral médio das classes nas dez bandas espectrais



A **figura 5** mostra a imagem classificada via máxima verossimilhança gaussiana e discriminação logística, ao lado da verdade terrestre.

Figura 5 – Imagens temáticas e verdade terrestre: primeira imagem classificada por máxima verossimilhança gaussiana; segunda imagem classificada por discriminação logística.



Os dois modelos confundiram uma das parcelas da imagem correspondente a classe *Soybean-min*. Isso ocorreu porque a amostra de treinamento contemplou apenas o lado esquerdo da imagem. De qualquer forma, o modelo logístico foi capaz de diferenciar classes com comportamento espectral médio muito semelhante, como *Grass-pasture* e *Grass-trees*, o que é altamente desejável. Mesmo com um número bem menor de parâmetros no modelo, a discriminação logística permitiu a obtenção de resultados semelhantes ao método da máxima verossimilhança gaussiana.

5 Considerações finais

O fato do modelo logístico não fazer restrições quanto a forma funcional das variáveis, torna a discriminação logística mais geral quando comparada a métodos convencionais de classificação, como a máxima verossimilhança gaussiana. Além disso, o número de parâmetros do modelo é relativamente baixo, permitindo uma menor quantidade de amostras de treinamento.

Mesmo com a possibilidade de problemas na estimação, ocasionados pela colinearidade, sugerimos que a discriminação logística seja considerada como uma alternativa viável para classificação de imagens digitais. Os estudos experimentais com imagens Landsat e AVIRIS mostraram que, mesmo quando a parte inferencial for sacrificada, o percentual de pixeis corretamente classificados pode ser alto, inclusive quando as classes apresentarem comportamento espectral semelhante.

Referências

Bittencourt, H. e Clarke, R.T. (2000) Estudo comparativo entre o modelo de discriminação logístico e o método da máxima verossimilhança gaussiana. In: IX Simpósio Latioamericano de Percepción Remota, Puerto Iguazú, 2000. **Anais do IX Simposio Latinoamericano de Percepción Remota**. Luján: UNLU. (em fase de publicação)

Efron, B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. **Journal of the American Statistical Association**, vol. 70, no. 352., p. 892-898.

Hosmer, D. and Lemeshow, S.. (1989) **Applied Logistic Regression**. New York: John Wiley & Sons.

Krzanowsky, W. J. (1988) **Principles of Multivariate Analysis**. Oxford: Clarendon Press.

McLachlan, G. (1992) **Discriminant Analysis and Statistical Pattern Recognition**. New York: John Wiley & Sons.

Press, J. and Wilson, S. (1978) Choosing between logistic regression and discrminant analysis. **Journal of the American Statistical Association**, vol. 73, no. 364., p. 699-705.