Busca de relações quando o número de variáveis é muito maior que o de observações: o caso de dados hiperespectrais

Alfredo José Barreto Luiz[1]
Aline de Holanda Nunes Maia[1]
Ieda Del'Arco Sanches[2]
Salete Gürtler[3]
Carlos Roberto de Souza Filho[3]

Resumo: Experimentos em que são coletadas muitas medidas em cada unidade experimental têm matrizes de dados nas quais o número de colunas (variáveis) é muito maior que o de linhas (observações). Dados hiperespectrais são habitualmente coletados por instrumentos que medem instantaneamente as reflectâncias de um alvo em milhares de comprimentos de onda e podemos considerar que cada uma delas constitui uma variável prognóstica num modelo de regressão. A facilidade na obtenção de cada vez maior número de variáveis simultâneas não se repete na obtenção das observações dessas variáveis. A análise de componentes principais (ACP) é indicada para tratar tal quantidade de variáveis e reduzir a dimensionalidade dos dados, mas sua aplicação ainda exige a obtenção de um grande número de medidas. Já a análise discriminante é usada na tentativa de classificar diferentes alvos, mas precisa ser precedida da seleção de um pequeno subconjunto de bandas, geralmente escolhidas com base em informações preexistentes e não nos próprios dados. A regressão linear permite empregar o método *stepwise* para selecionar um subconjunto de bandas, mas só é indicada para variáveis dependentes quantitativas. O presente trabalho propõe o uso da regressão logística politômica *stepwise* para selecionar um pequeno conjunto de bandas espectrais que discrimine alvos em *k* classes, quando a variável resposta de interesse é nominal. Apresentamos um exemplo no qual os dados espectrais são utilizados para construção de modelos logísticos com um pequeno número de preditores (bandas) para classificação de folhas verdes em classes correspondentes a três culturas agrícolas: soja perene, milho e braquiária.

Palavras-chave: regressão logística politômica, *stepwise*, seleção de variáveis explicativas

[1]Embrapa Meio Ambiente. e-mail: alfredo.luiz@embrapa.br
[2]Divisão de Sensoriamento Remoto (DSR), Instituto Nacional de Pesquisas Espaciais (INPE).
[3]Instituto de Geociências, Universidade Estadual de Campinas.

Searching relations when the number of variables is much larger than the observations: the case of hyperspectral data

Alfredo José Barreto Luiz[4]
Aline de Holanda Nunes Maia[1]
Ieda Del'Arco Sanches[5]
Salete Gürtler[6]
Carlos Roberto de Souza Filho[3]

Abstract: This paper proposes the stepwise polytomous logistic regression to select small set of spectral bands that discriminate targets in k classes when response variable is nominal. Spectral data are used in logistic models with a small number of predictors (bands) for classification of green leaves in three corresponding classes (crops).
Keywords: polytomous logistic regression, stepwise, selection of explanatory variables

[4]Embrapa Meio Ambiente. e-mail: alfredo.luiz@embrapa.br
[5]Divisão de Sensoriamento Remoto (DSR), Instituto Nacional de Pesquisas Espaciais (INPE).
[6]Instituto de Geociências, Universidade Estadual de Campinas.

Busca de relações quando o número de variáveis é muito maior que o de observações: o caso de dados hiperespectrais

Alfredo José Barreto Luiz[7]
Aline de Holanda Nunes Maia[1]
Ieda Del'Arco Sanches[8]
Salete Gürtler[9]
Carlos Roberto de Souza Filho[3]

1 Introdução

Alguns experimentos ou levantamentos em que são coletadas muitas medidas em cada unidade experimental ou amostral, as matrizes de dados correspondentes tem o número de colunas (variáveis) muito maior do que o número de linhas (observações). Um caso típico é o de dados hiperespectrais, habitualmente coletados por instrumentos que, numa fração de segundos, medem a reflectância de um alvo em centenas ou milhares de comprimentos de onda. Como cada comprimento de onda pode estar relacionado a uma característica do alvo, podemos considerar que cada um constitui uma variável prognóstica num modelo de regressão. A facilidade de obtenção de um número cada vez maior de variáveis simultâneas não é acompanhada por uma maior facilidade de obtenção de um grande número de observações dessas variáveis. Pesquisadores da área de sensoriamento remoto tendem a empregar técnicas de análise de componentes principais (ACP) para tratar de forma compreensível tal quantidade de variáveis, entretanto, o uso de combinações lineares de variáveis ainda requer medidas de um grande número de comprimentos de onda. Neste campo, a justificativa para o uso desta técnica baseia-se na afirmação que a ACP "é usada para avaliar a dimensionalidade dos dados e as componentes principais são usadas para construir mapas" [1]. Outros utilizaram a análise discriminante na tentativa de classificar diferentes alvos [2], precedida pela seleção de poucas bandas a serem usadas na análise, escolhidas com base em indicações existentes na literatura e não nos próprios dados. Em outro exemplo, os autores se utilizaram da técnica de regressão linear e o método *stepwise* para

[7] Embrapa Meio Ambiente. e-mail: alfredo.luiz@embrapa.br
[8] Divisão de Sensoriamento Remoto (DSR), Instituto Nacional de Pesquisas Espaciais (INPE).
[9] Instituto de Geociências, Universidade Estadual de Campinas.

selecionar entre as milhares de bandas disponíveis aquelas com maior capacidade de explicar a variação de um fenômeno contínuo (volume de madeira) observado no alvo de interesse [3].

No presente trabalho, propomos o uso de regressão logística politômica para selecionar, entre milhares de bandas espectrais, um pequeno conjunto que discrimine os alvos em k classes. No presente caso, a variável resposta de interesse é nominal (espécie vegetal), o que inviabiliza o uso de regressão linear múltipla. Apresentamos um exemplo no qual os dados espectrais são utilizados para construção de modelos logísticos com um pequeno número de preditores (bandas) para classificação das observações (folhas verdes de diferentes espécies vegetais) em classes correspondentes a três culturas agrícolas: soja perene, milho e braquiária.

2 Material e Métodos

Foram utilizados dados colhidos a campo de dois experimentos com as culturas de soja perene (*Neonotonia wightii*), milho (*Zea mays*) e braquiária (*Brachiaria brizantha*). Tanto no primeiro experimento, realizado em 2010 [4], como no segundo, conduzido em 2013 [5], em toda data de realização das medidas eram obtidos dados simultâneos de reflectância de 2.151 bandas espectrais para cada folha analisada. Embora os experimentos contivessem outros tratamentos e visassem objetivos diversos, apenas os dados oriundos das parcelas controle de cada espécie foram considerados no presente trabalho, cuja meta é demonstrar a aplicabilidade da regressão logística politômica usando o método *stepwise* para a seleção de preditores na análise de dados hiperespectrais. Em ambos os experimentos foram realizadas medidas em 9 momentos (Tabela 1), sendo que o intervalo entre as datas e os correspondentes dias após o plantio (**dap**) variaram.

Tabela 1. Relação das datas das medições nos dois experimentos analisados.

Data	Medidas	Data	Medidas
20/04/2010	M1_1	28/05/2013	M2_1
27/04/2010	M1_2	04/06/2013	M2_2
29/04/2010	M1_3	11/06/2013	M2_3
06/05/2010	M1_4	18/06/2013	M2_4
10/05/2010	M1_5	02/07/2013	M2_5
12/05/2010	M1_6	10/07/2013	M2_6
17/05/2010	M1_7	16/07/2013	M2_7
20/05/2010	M1_8	30/07/2013	M2_8
28/05/2010	M1_9	06/08/2013	M2_9

No experimento de 2010, em cada data foram mensuradas as reflectâncias de 20 folhas de cada espécie e no ano de 2013 as medidas foram repetidas em 30 folhas. Devido ao pouco desenvolvimento das plantas, só foram utilizadas as medidas a partir da terceira data de medição do primeiro experimento (M1_3), totalizando 7 datas para as quais foram ajustadas as equações de regressão logística. Nos dois experimentos a sétima medição correspondeu aos 105 **dap** da soja perene.

Foi utilizado o procedimento Logistic do SAS/STAT [6] para selecionar entre as 2.151 bandas disponíveis aquelas que melhor discriminam as espécies, em diferentes momentos do ciclo das culturas. Embora a regressão logística seja mais frequentemente utilizada para modelar a relação entre uma variável resposta dicotômica ($k = 2$) e um conjunto de variáveis preditoras, ela também pode ser aplicada quando existem mais de dois níveis ($k > 2$) na variável resposta. Nesse caso é aplicado o modelo de regressão logística politômica [7]. A generalização do modelo logístico nesse caso é direta [8] e resulta que a probabilidade de uma observação $\boldsymbol{y}$ pertencer a uma das classes y_i, dado um vetor de preditores $\mathbf{x}$, é estimada diretamente por meio da seguinte expressão:

$$P(Y = y_i \mid \mathbf{x}) = \frac{exp\{g_i(x)\}}{1 + \sum_{j=1}^{k-1} exp\{g_j(x)\}}$$

$$i = 1, 2, \ldots, k\text{-}1$$

onde a função *logit*, assumindo o nível y_k como base, é dada por:

$$g_i(x) = ln\left[\frac{P(Y = y_i \mid \mathbf{x})}{P(Y = y_k \mid \mathbf{x})}\right] =$$

$$\beta_{i0} + \beta_{i1}x_1 + \ldots + \beta_{ip}x_p$$

$$i = 1, 2, \ldots, k\text{-}1$$

$$g_k(x) = 0 \cdot$$

O código do programa SAS para a leitura dos dados da primeira fase de desenvolvimento das culturas do experimento de 2010 é apresentado na Figura 1. Os dados foram lidos de um arquivo do tipo Excel, com três planilhas, uma para cada cultura. Nesse arquivo, cada planilha apresenta, para cada espécie, na primeira coluna os valores, em nanômetros, dos comprimentos de onda de cada banda espectral e, nas demais colunas, as medidas de reflectância de cada uma das folhas.

```
/* 'Importação dos dados e construção do arquivo para análise'; */
options nonumber nodate pagesize=1000 linesize=80;
run;
proc import datafile="C:\[...]M1_x.xls" out=brac dbms=xls replace;
        sheet="brac";
        getnames=no;
run;
data brac;
  set brac;
   cultura='brac';
   run;
proc import datafile="C:\[...]M1_x.xls" out=milh dbms=xls replace;
        sheet="milh";
        getnames=no;
run;
data milh;
  set milh;
   cultura='milh';
   run;
proc import datafile="C:\[...]M1_x.xls" out=soja dbms=xls replace;
        sheet="soja";
        getnames=no;
run;
data soja;
  set soja;
   cultura='soja';
   run;
proc transpose data=brac out=dois prefix=L_;
by cultura;
   id A;
      run;
proc transpose data=milh out=tres prefix=L_;
by cultura;
    id A;
      run;
proc transpose data=soja out=quatro prefix=L_;
by cultura;
    id A;
      run;
Data um;
  merge dois tres quatro;
  by cultura;
  drop _name_ _label_;
  run;
quit;
```

Figura 1. Programa SAS para importação dos dados para análise.

O ajuste da regressão logística politômica via PROC Logistic, pelo método de máxima verossimilhança, empregou o código apresentado na Figura 2, em continuação ao anterior.

Como resultado desse procedimento o programa nos fornece os valores de intercepto e dos coeficientes lineares correspondentes a cada preditor (banda) selecionado no modelo ajustado. Nesse tipo de análise, os resultados são expressões para estimar a probabilidade de

uma observação pertencer a cada uma das classes, dados os valores observados das bandas selecionadas no modelo.

```
/*'Análise de regressão logística de dados hiperespectrais data M1_x'*/
proc sort data=um;
  by cultura;
  run;
Proc logistic data=um order=data;
class cultura (ref = last) / param = ref;
model cultura = L_350-L_2500
         / selection=stepwise
                   slentry=0.4
                   slstay=0.6
                            link = glogit;
      output out=logreg p=pred  predprobs=(i);
run;
proc print data=logreg;
 var cultura _FROM_ _INTO_;
 run;
quit;
```

Figura 2. Programa SAS para ajuste do modelo de regressão logística politômica usando o método *stepwise* para seleção de preditores em cada momento de avaliação.

Na primeira etapa, utilizaram-se os dados do experimento com três espécies distintas [4] para a seleção das bandas e obtenção das equações. Num segundo passo, a performance do modelo foi avaliada utilizando os dados obtidos num experimento independente [5].

3 Resultados e Discussões

Na Tabela 2, são apresentados os resultados do ajuste do modelo de regressão logística multinomial para as respostas espectrais das culturas de soja perene, milho e braquiária, provenientes de experimento de campo. De posse dos valores encontrados para os parâmetros das regressões, foram calculadas, em cada data e para cada observação, a probabilidade de uma observação, com seus valores de reflectância nas bandas selecionadas, pertencer a cada uma das espécies. De maneira simplificada, podemos dizer que, no caso da data M1_3, as probabilidades são calculadas da seguinte maneira:

$$E(b)_j = \exp [244{,}4 - (717 \times L\_1383_j)]$$

$$E(m)_j = \exp [497{,}5 - (1580 \times L\_1383_j)]$$

$$P(Y=s)_j = 1 / [1 + E(b)_j + E(m)_j]$$

$$P(Y=b)_j = E(b_j / [1 + E(b)_j + E(m)_j]$$

$$P(Y=m)_j = E(m)_j / [1 + E(b)_j + E(m)_j]$$

onde j = 1 até n (n sendo o número de observações) e $P(Y=s)_j$, $P(Y=b)_j$ e $P(Y=m)_j$ são as probabilidades de que a i-ésima observação, com seus valores de reflectância nas bandas selecionadas, pertença às classes: soja perene, braquiária e milho, respectivamente.

Tabela 2 – Resultado da análise de regressão logística *stepwise*: parâmetros estimados, em cada data (M3 a M9), para braquiária e milho, ao considerar a soja perene como referência.

	Fase das culturas													
	M1_3		M1_4		M1_5		M1_6		M1_7		M1_8*		M1_9	
	braq	milh	braq	milh	braq	milh	braq	milh	braq	milh	braq	milh	braq	milh
Intercepto	244,4	497,5	-226,2	-132,4	-64,7	156,2	-496,9	-70,2	223,0	319,9	-183,4	-103,2	-82,3	169,6
Bandas														
355													60	-1010
366													-133	690
371			1151	2975										
373			763	-1923										
418											2406	1256		
445							5246	1510						
499													1324	1166
643											78	1318		
676					1168	1929								
702											129	-534		
727					-27	-842								
1373									3732	-1759				
1383	-717	-1580												
1393							6763	1070						
1660									-5237	756				
1851							-6431	-1194						
1904													-31096	-64945
1905													37736	71817
2490									2300	501				
2492													3916	-7746

* nessa data houve um único erro de classificação, de milho para braquiária; para as demais datas, o acerto foi total.

Utilizando as expressões acima e substituindo o valor da reflectância de cada observação, cada uma é atribuída à classe para a qual ela apresenta maior probabilidade de pertencimento. Para as demais épocas, o procedimento se repete de maneira a classificar cada observação.

Como pode ser observado pelos resultados expostos na Tabela 2, um máximo de 6 bandas (dentre as 2.151 disponíveis) foi suficiente para separar completamente as observações e classificá-las corretamente entre as espécies estudadas. O único erro de classificação aconteceu com uma observação oriunda de uma folha de milho na fase M1_8, que foi classificada pelo modelo como braquiária. O número de bandas incluídas no modelo cresceu

com a idade das folhas. Não foi possível estabelecer um padrão com relação ao posicionamento das bandas ao longo do espectro. Para cada data, o conjunto de bandas selecionado foi diferente.

Para trabalhos futuros, é recomendável que as datas de leitura sejam associadas à soma térmica acumulada no período a partir da emergência das plantas, para que os resultados possam ser comparados com os obtidos em outras regiões ou épocas, cujas condições climáticas durante o ciclo não sejam exatamente iguais.

Embora no presente caso o local tenha sido o mesmo mas as datas de plantio não e, consequentemente a soma térmica deve ter variado de um ano para outro, para efeito de avaliação da análise realizada foi tomada a sétima medição dos dois experimentos que, como mostrado na Tabela 1, ocorreram no mesmo **dap** da soja perene (105). Pelo modelo de regressão logística politômica ajustado aos dados correspondentes a essa medição no primeiro experimento, as bandas selecionadas usando o método stepwise foram usadas no mesmo modelo logístico para o ajuste dos dados do segundo experimento. Os parâmetros estimados são apresentados na Tabela 3.

Tabela 3 – Resultado da análise de regressão logística do experimento de 2013 utilizando as bandas escolhidas com base no experimento de 2010: parâmetros estimados, na medição M2_7, para braquiária e milho, ao considerar a soja perene como referência.

	braquiária	milho
Intercepto	208,1	201,5
Bandas		
1373	3150	28108
1660	-4388	-37828
2490	2009	961

Usando esses parâmetros, das noventa folhas medidas, apenas duas de braquiária foram erroneamente classificadas como milho e uma de milho classificada como braquiária. Todas as 30 folhas de soja perene foram corretamente classificadas. Como tanto o milho como a braquiária são poáceas e a soja perene é uma fabácea, as diferenças fisiológicas que podem afetar na reflectância devem ser maiores entre famílias distintas do que dentro da mesma família e entre espécies diferentes.

4 Conclusões

O uso do modelo de regressão logística politômica usando o método *stepwise* para a seleção de preditores se mostrou adequado para a busca de relações entre um conjunto inicial com grande número

de variáveis independentes numéricas e poucas variáveis dependentes categorizadas nominais, mesmo quando o número de observações é muito menor que o de variáveis.

O método pode ajudar na busca de explicações físicas e fisiológicas ao identificar um pequeno conjunto de variáveis independentes capazes de promover a diferenciação das espécies de plantas.

No presente conjunto de dados foi possível perceber que à medida que as folhas eram provenientes de plantas mais velhas, foi preciso incluir um maior número de bandas no modelo para separar corretamente as espécies.

5 Bibliografia

[1] FLINK, P.; LINDELL, T.; ÖSTLUND, C. Statistical analysis of hyperspectral data from two Swedish lakes. **The Science of the Total Environment,** v. 268, p. 155-169, 2001.

[2] GALVÃO, L. S.; FORMAGGIO, A. R.; TISOT, D. A. Discriminação de variedades de cana-de-açúcar com dados hiperespectrais do sensor Hyperion/EO-1. **Revista Brasileira de Cartografia**, v.57/01, p. 7-14, 2005.

[3] CANAVESI, V.; PONZONI, F.J.;VALERIANO, M. M. Estimativa de volume de madeira em plantios de Eucalyptus spp. utilizando dados hiperespectrais e dados topográficos. **Revista Árvore**, v.34, n.3, p. 539-549, 2010.

[4] SANCHES, I. D.; SOUZA FILHO, C.R.; MAGALHÃES, L. A.; QUITÉRIO, G. C. M.; ALVES, M. N.; OLIVEIRA, W. J. Assessing the impact of hydrocarbon leakages on vegetation using reflectance spectroscopy. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 78, p. 85–101, 2013.

[5] GÜRTLER, S.; SOUZA FILHO, C.R.; SANCHES, I. D.; NOPPER, M. Detecção de estresse por hidrocarbonetos em culturas agrícolas a partir de índices de vegetação de banda estreita. XV SBSR. **Anais**. João Pessoa, PB. INPE, no prelo, 2015.

[6] SAS Institute Inc. 2013. **SAS/STAT**® 13.1 User's Guide. Cary, NC: SAS Institute Inc.

[7] NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. **Applied linear statistical models**. 4th ed. Chicago: Irwin, 1996. 1408p.

[8] BITTENCOURT, H. R. Regressão logística politômica: revisão teórica e aplicações. A**cta Scientiae**. v. 5, n. 1, p. 77-86, 2003.