



Desenvolvimento de metodologias de seleção de amostras

Alisson Silva de Araújo

Adriano de Araújo Gomes

Agência financiadora: CNPq

Resumo: O uso de um conjunto de amostras representativas na etapa de calibração é um ponto chave para obtenção de modelos com elevada capacidade preditiva. Neste contexto este trabalho reporta o desenvolvimento de uma nova estratégia de seleção de amostras baseada no algoritmo das projeções sucessivas para partição de conjuntos de amostras. O método proposto, denominado de SS-SPA, foi desenvolvido em ambiente MatLab e avaliado em um estudo de caso simulado envolvendo calibração e classificação. Em comparação com outras estratégias, o critério de seleção de amostras baseado em projeção de vetores, se mostrou eficaz, selecionando amostras representativas, levando resultados comparáveis a método consolidados.

Palavras chave: Seleção de amostras, algoritmo das projeções sucessivas, modelos multivariados.

1. INTRODUÇÃO

O sucesso de modelos multivariados passa, sem dúvida, pela qualidade das amostras que compõem o conjunto de calibração/treinamento. O conjunto de amostras calibração/treinamento deve ser representativo o suficiente para garantir bons estimadores populacionais dos parâmetros do modelo e deve assegurar a modelagem de todo espaço amostral, afim de se evitar extrapolações na etapa de predição. Nos casos em que as amostras de calibração são misturas de padrões, é recorrente o uso de planejamentos experimentais para assegurar as características citadas acima (ARAÚJO et al., 2001). Contudo, em muitas aplicações em calibração multivariada e em todos os casos que envolve classificações, amostras reais, que foram analisadas previamente por um método oficial e/ou de referência, são usadas para compor o conjunto de calibração. Neste contexto, assegurar representatividade das amostras não é uma tarefa trivial (GALVÃO et al., 2001).

Algumas alternativas são apresentadas na literatura para este propósito, como a seleção aleatória, algoritmo de Kernard-Stone (KS), SPXY e DUPLEX (GALVÃO et al., 2007; PEREIRA et al., 2008; FERNANDES et al., 2011). O Algoritmo das Projeções Sucessivas (SPA), proposto originalmente como estratégia de seleção de variáveis, foi usado na seleção de amostras para transferências de calibração via modelos de regressão linear múltipla (MLR) (PONTES et al., 2011; DI NEZIO et al., 2007). Desta forma, neste trabalho, o SPA é caracterizado como método de partição de conjunto de amostras e comparado as principais estratégias descritas na literatura em estudos de caso simulados envolvendo classificação e calibração.

2. MATERIAL E MÉTODOS

A rotina, denominada SPASS, foi implementada em ambiente MatLab 2010a com possibilidades de seleção de amostras para o desenvolvimento de modelos de classificação e calibração. O usuário conta ainda com a possibilidade de partição das amostras em subconjuntos de treinamento/calibração, teste/predição bem como e validação. Inicialmente a matriz de respostas instrumentais $\mathbf{X}_{(I \times J)}$, com I amostras analisadas em J sensores é transposta para gerar $\mathbf{X}^T_{(J \times I)}$. A matriz $\mathbf{X}^T$ é submetida ao processo de projeções sucessivas, tendo como vetor de partida ($\mathbf{z}^1$) o vetor amostra de maior norma quadrada. Após a etapa de projeções, um vetor $\mathbf{s}$ (1 x I) contendo o sequenciamento crescente de projeções de todas as I amostras no

plano P^1 ortogonal a z^1 . Com base nas quantidades de amostras predefinidas para compor cada subconjunto, (treinamento/calibração, validação externa e teste/predição), estes são gerados como output da rotina SS-SPA. Na Figura 1a, é mostrado um esquema de funcionamento do método SS-SPA.

O algoritmo SS-SPA, inicialmente implementado em linhas de comando, foi também implementada posteriormente na forma de interface gráfica para melhor difusão e facilitação do uso. A Figura 1b ilustra a janela principal do algoritmo desenvolvido. Nessa janela o analista pode carregar os dados e efetuar a partição das amostras conforme configurações escolhidas pelo mesmo.

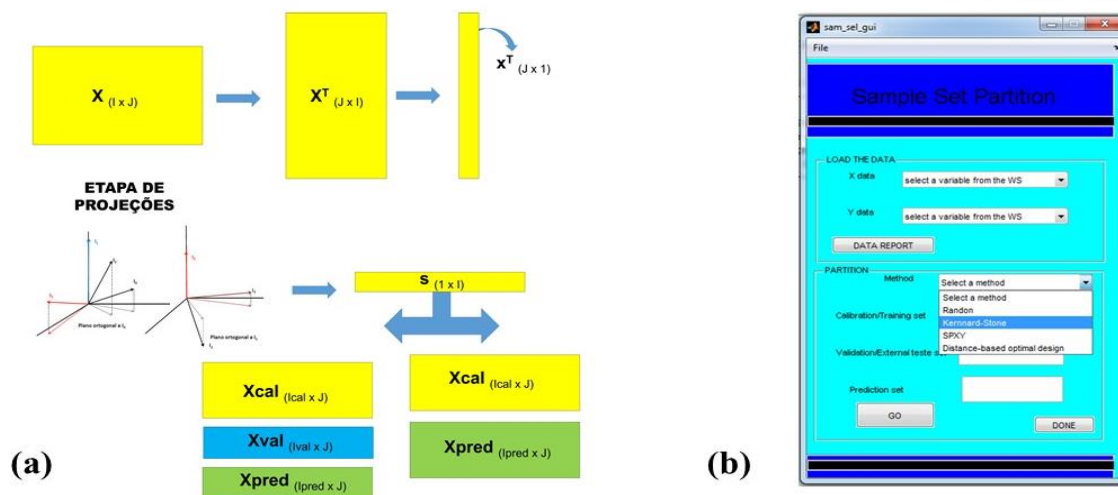


Figura 1: Em (a) mostra o esquema do funcionamento do SPASS e em (b) a janela principal do software.

Estudo simulado

Dois estudos de caso simulados, envolvendo classificação e calibração, foram realizados. Para ambos os casos foram empregues somas de perfis gaussianos para obter espectros simulados. Para avaliar a classificação foram gerados espectros simulados de 90 amostras pertencentes a duas classes (45 amostras por classe). Estas amostras foram particionadas em conjuntos de Treinamento (25 amostras por classe), validação (10 amostras por classe) e teste (10 amostras por classe) empregando o método proposto. Para propósito de comparação foram também avaliadas partições envolvendo os métodos randômicos (RS) e Kernnard-Stone (KS). Para todos os casos modelos PLS-DA foram construídos para avaliação do desempenho dos mesmos em função da partição do conjunto de amostras.

Para o estudo de caso simulado de calibração multivariada 45 amostras foram geradas com contrações de um analito variando em três concentrações, segundo um planejamento experimental fatorial completo. A partição das amostras se deu em calibração (25 amostras), validação (10 amostras) e Predição (10 amostras). Para todos os casos modelos PLS foram construídos.

3. RESULTADOS E DISCUSSÃO

Estudo de caso (I): Classificação

A taxa de acerto dos modelos PLS-DA com as amostras selecionadas pelos métodos Randômico, KS e SS-SPA se encontram na Tabela 1. O método proposto alcançou modelos altamente seletivos e específicos, para um menor número de fatores. Todas as amostras foram atribuídas as classes corretas, levando a 100% de acerto no ajuste, na validação cruzada e na

predição de um conjunto de teste externo. Em adição, o modelo PLS-DA, baseado em partição randômico, apresentou erros de classificação no conjunto de teste, possivelmente por efeitos de extrapolação. O Método KS, obteve desempenho similar ao método propostos, 100% em todas as etapas. Contudo, este método de partição levou a modelos mais complexo requerendo 6 fatores.

Método	Treinamento	Validação	Teste
		cruzada	
	Taxa de acerto	Taxa de acerto	Taxa de acerto
Randômico/PLS (5)*	100%	100%	96%
KS/PLS (6)*	100%	100%	100%
SPASS/PLS (5)*	100%	100%	100%

Tabela 1: Resumo estatístico dos modelos PLS-DA * Número de fatores nos modelos PLS-DA

Estudo de caso (II): Calibração

Neste estudo de caso, envolvendo a calibração multivariada, para todos os casos o número ótimo de fatores estimado por validação cruzada foi de 6 fatores. Este valor foi usado para ajustar os modelos PLS, cujos resultados estão sumarizados na Tabela 2.

Método	Calibração	Validação cruzada			Predição		
					REP (%)	bias	
	RMSEC	REP (%)	RMSECV	REP (%)	RMSEP		
Randômico/PLS (6)*	0,049	1,28	0,146	3,72	0,100	2,75	ausente
KS/PLS (6)*	0,061	1,73	0,133	3,74	0,112	2,90	presente
SPASS/PLS (6)*	0,053	1,45	0,122	3,31	0,103	2,71	ausente

Tabela 2: Resumo do ajuste e predição para os modelos PLS.* Números de fatores nos modelos PLS-DA.

Como pode ser observado pelos resultados do ajuste (RMSEC), todos os modelos PLS, baseados em diferentes estratégias de partição de conjuntos de amostras, mostraram ajuste satisfatórios (erro relativo menor que 5%). Os valores de REP (erro relativos) foram menores para os métodos propostos, indicando que as amostras selecionadas pelo SS-SPA, para compor o conjunto de calibração, levam a um modelo PLS com maior capacidade preditiva e sem presença de bias significativo.

4. CONCLUSÃO

Neste trabalho foi apresentado uma nova proposta de algoritmo de seleção de amostra em classificação e calibração multivariada via algoritmo das projeções sucessivas e avaliados em estudos de caso simulados. Quando comparados a estratégias consolidadas na literatura, o critério de seleção de amostras baseado em projeção de vetores se mostrou eficaz, selecionando amostras representativas levando resultados comparáveis ou ligeiramente melhores que KS e a seleção randômica.

5. REFERÊNCIAS E CITAÇÕES

- ARAÚJO, M. C. U., et al. *The successive projections algorithm for variable selection in spectroscopic multicomponent analysis*. Chemometrics and Intelligent Laboratory Systems. 57: 65, 2001.
- DANTAS FILHO, H. A., et al. *Simultaneous Spectrometric Determination of Cu^{2+} , Mn^{2+} and Zn^{2+} in Polivitaminic/Polimineral Drug Using SPA and GA Algorithms for Variable Selection*. J. Braz. Chem. Soc. 16: 58, 2005.
- DI NEZIO, M. S., et al. *Successive projections algorithm improving the multivariate simultaneous direct spectrophotometric determination of five phenolic compounds in sea water*. Microchemical Journal. 85: 194, 2007.
- FERNANDES, D. D. S., et al. *Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection*. Talanta 87: 30, 2011.
- GALVÃO, R. K. H., et al. *Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry*. Analytica Chimica Acta. 443: 107, 2001.
- GALVÃO, R. K. H., et al. *Cross-Validation for the Selection of Spectral Variables Using the Successive Projections Algorithm*. Journal of Brazilian Chemical Society. 18: 1580, 2007.
- PEREIRA, A. F. C., et al. *NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection*. Food Research International. 41: 341, 2008.
- PONTES, M. J. C., et al. *Determining the quality of insulating oils using near infrared spectroscopy and wavelength selection*. Microchemical journal 98: 254, 2011.