EMPREGO DAS TÉCNICAS DE REMOSTRAGEM *BAGGING* E *SUBBAGING* EM CALIBRAÇÃO MULTIVARIADA

Anna Lúcia Bezerra da Silva
Leandro Pedrosa
Samuel da Costa Vicente
Clarimar José Coelho

Resumo

Este trabalho apresenta o uso de técnicas de calibração multivariada em quimiometria ambiental. É feita uma pequena introdução sobre a regressão linear múltipla (*Multiple Linear Regression*, MLR) e sobre as técnicas de reamostragem *bootsrap*, *bagging* e *subbaging*. Como exemplo de aplicação é feita a calibração para amostras de aço-ligas contendo Manganês (Mn), Molibdênio (Mo), Cromo (Cr), Níquel (Ni) e Ferro (Fe). Os resultados obtidos com a RLM combinada com as técnicas de reamostragem *bagging* e *subbaging* são melhores que os resultados obtidos com a MLR-tradicional.

1. Introdução

A quimiometria é definida como a disciplina da química que usa métodos matemáticos e estatísticos para planejar ou selecionar experimentos de forma otimizada e para fornecer o máximo de informação química na análise de dados de natureza multivariada, ou como veículos que auxiliam os químicos a se moverem de forma mais eficiente na direção do maior conhecimento (FERREIRA, 1999).

A quimiometria é uma área da química analítica quantitativa que aplica métodos estatísticos e matemáticos associados à computação. A ênfase é dada aos sistemas multivariados onde é possível medir muitas variáveis simultaneamente ao analisar uma amostra qualquer (LAWSON; HANSON, 1974; NETO et al, 2006). Nesses sistemas, a conversão da resposta instrumental no dado químico de interesse, requer a utilização de técnicas de estatística multivariada, álgebra matricial e análise numérica (BEEBE et al, 1998). Essas técnicas constituem no momento na melhor alternativa para a interpretação de dados e para a aquisição do máximo de informação sobre o sistema (BROWN, 1995; NETO et al ,2006; PAKYARI, 2008).

A calibração multivariada utiliza muitas variáveis $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ simultaneamente para quantificar outra variável de interesse $\mathbf{y}$. O método mais simples de calibração multivariada é

a regressão linear múltipla (*Multiple Linear Regression*, MLR). A MLR faz a predição da variável $\mathbf{y}$ pela estimativa da combinação linear das variáveis independentes da matriz $\mathbf{X}$,

$$\mathbf{Y} = \mathbf{Xb} \quad (1)$$

$$\mathbf{y} = \mathbf{b}_1\mathbf{x}_1 + \mathbf{b}_2\mathbf{x}_2 + \ldots + \mathbf{b}_n\mathbf{x}_n \quad (2)$$

onde $\mathbf{b}$ é o vetor com os coeficientes de regressão. A solução em mínimos quadrados para $\mathbf{b}$ é dada por $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

A MLR é um bom método para sistemas com respostas lineares e sem colinearidade. A principal limitação da MLR vem do fato que ela usa toda a informação contida na matriz $\mathbf{X}$ mesmo não que não seja importante para a construção do modelo. Assim, grande quantidade de informação sem interesse é incluída no modelo (SENA, 2000).

Em algumas situações é possível combinar a MLR com outros modelos para melhorar sua capacidade preditiva mantendo os benefícios da simplicidade que a MLR proporciona.

Os métodos de reamostragem usam múltiplas versões de um conjunto de dados de treinamento. Cada versão é criada selecionando um subconjunto de treinamento que é utilizado para treinar diferentes modelos de regressão componentes e a regressão final é construída por meio da contribuição individual de cada modelo. A técnica tem sido aplicada com sucesso a problemas de regressão e classificação em diferentes áreas, mostrando que é possível obter resultados mais precisos a partir da combinação de diferentes modelos de regressão do que a partir dos dados originais. Atualmente, vários métodos para realizar combinações vêm sendo pesquisados (SIMAR, 2003).

O objetivo deste trabalho e fazer a calibração multivariada empregando a MLR e as técnicas de reamostragem *bagging* e *subbaging* (BREIMAN, 1996; EFRON; TIBSHIRANI, 1993, FILHO, 2007; FOX, 2002) para obter a concentração dos analitos em um conjunto de dados proveniente de amostras de aço-ligas contendo Manganês (Mn), Cromo (Cr), Molibdênio (Mo), Níguel (Ni), Ferro (Fe).

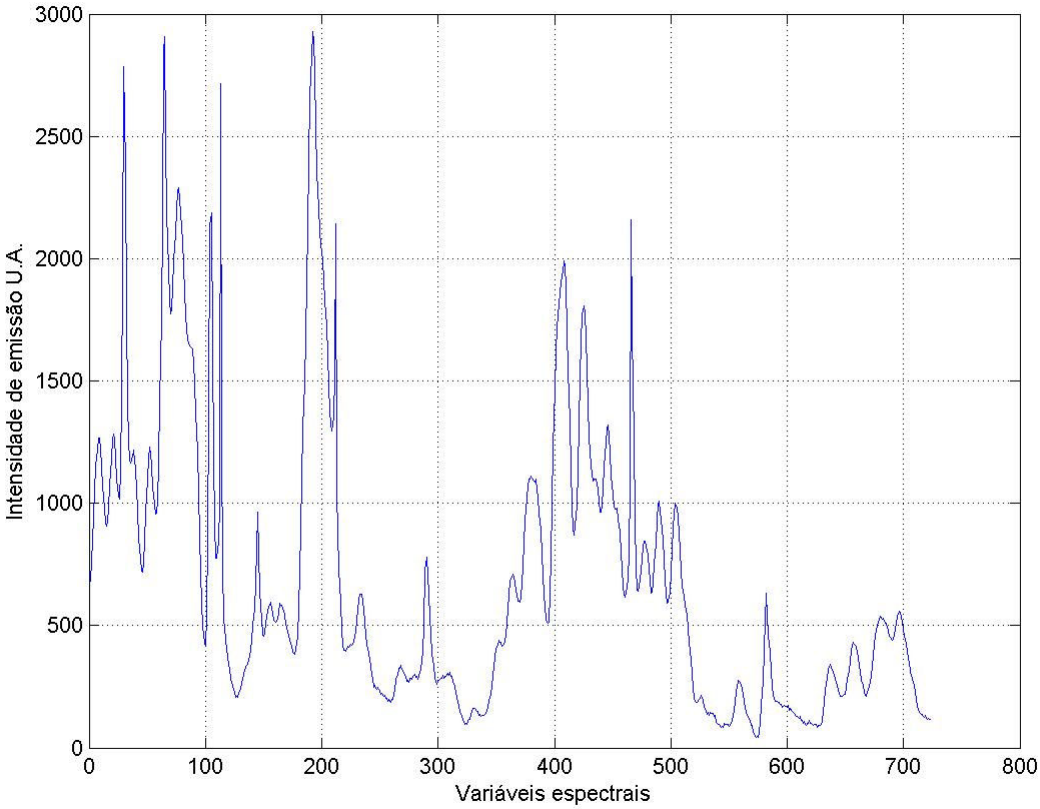

Figura 1. Espectro de uma amostra contendo Mn, Mo, Cr, Ni, Fe.

Vale ressaltar que o software desenvolvido para a calibração usando dados de aço-ligas pode ser usado para a obtenção da concentração de analitos no contexto da análise ambiental.

2. Técnicas de Reamostragem

O *bootstrap* é um método simples, porém preciso para análises estatísticas de simulações aleatórias. Seu principal objetivo é reamostrar o conjunto de dados para gerar amostras que possam ser utilizadas na estimação de um parâmetro de interesse (TIMM,2002). A rigor obteém-se pseudo-amostras, uma vez que são obtidas da amostra original seguindo um procedimento específico de reamostragem.

O *bootstrap* é um procedimento computacional que fornece a medida do nível de precisão da inferência estatística, como por exemplo, média, desvio padrão e erro padrão (LEVINE et al ,2005). A idéia básica do *bootstrap* é inferir estatisticamente as propriedades de interesse a

partir da análise de amostras aleatórias extras e amostras da população original sem a necessidade de conhecer a distribuição dos dados a priori o que constitui em significativa vantagem sobre outras técnicas (HÄRDLE; TIMM, 2002).

Se o parâmetro populacional de interesse é desconhecido, procura-se estimá-lo por reamostragem aleatória. A amostra é um conjunto de *n* elementos independentes retirados da população $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$.

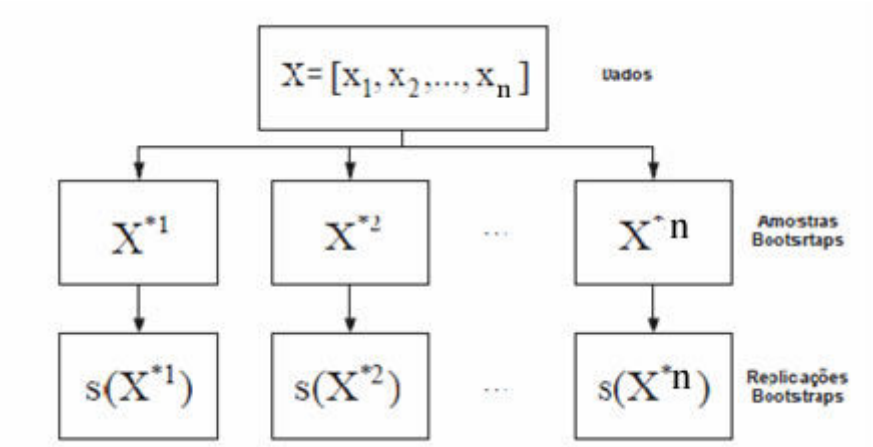

Figura 1. Reamostragem de amostras usando a técnica *bootstrap*.

A Figura 1 ilustra o processo de reamostragem de amostras usando a técnica *bootstrap.* O primeiro retângulo na Figura 1 representa a população original **X**. Os retângulos imediatamente abaixo representam a extração das amostras aleatórias *bootstrap* da população **X**. Os retângulos imediatamente abaixo representam as replicações bootstrap, onde é aplicada uma função estatística $s(\gamma)$ que é recalculada *n* vezes.

A técnica *bagging* (*bootstrap aggregating*) proposta por Breiman (BREIMAN, 1996) tem origem na técnica *bootstrap*. O *bagging* tem como objetivo obter diferentes conjuntos de calibração a partir da modelagem empregando o *bootstrap* e posteriormente faz a combinação de diferentes modelos gerados pela reamostragem aleatória dos dados disponíveis (BRAGA, 2008).

A técnica *subagging* (*subsample aggregating*) basea-se na reamostragem sem reposição (ou subamostragem). Cada modelo é construído de maneira similar ao *bagging*, através de amostras *bootstrap*, porém sem repetir amostras.

3. Resultados

Após a reamostragem dos dados é feita a calibração multivariada com a MLR-*bagging* e MLR-*subbaging*. A capacidade de predição do modelo é verificada com o erro médio de predição (*Root Mean Square Error of Prediction*, RMSEP). Esse índice é definido para um número n_t de pontos de teste, não usados na calibração do modelo, dado por

$$RMSEP = \sqrt{\frac{1}{n_t}\sum(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2} \quad (3)$$

onde $\mathbf{y}_i$ é o valor real de $\mathbf{y}_i$ e $\hat{\mathbf{y}}_i$ é o valor predito para $\mathbf{y}_i$.

Tabela 1. RMSEP das amostras aço-ligas usando as técnicas de reamostragem.

RMSEP	Mn	Cr	Mo	Ni	Fe	Média
RMSEP MLR-tradicional	0,2282	2,8239	0,3706	2,3717	6,0749	2,3739
RMSEP MLR-*bagging*	0,1588	2,5305	0,2676	1,4062	2,6004	1,3927
RMSEP MLR-*subagging*	0,0885	2,1070	0,2018	1,3953	2,0119	1,1609

A Tabela 1 mostra os RMSEP calculados para todos os analitos do conjunto de dados de aço-ligas com a MLR-tradicional e empregando as técnicas de reamostragem MLR-*bagging* e MRL-*subagging*.

É possível concluir que o RMSEP calculado com a técnica MLR-*bagging* apresenta um erro de predição 41,33% menor em média do que aqueles calculados com a MLR-tradicional. O RMSEP calculado com a técnica MLR-*subbaging* apresenta um erro de predição 51,09% menor em média do que os calculados com a MLR-tradicional.

4. Conclusão

Os conceitos relacionados às técnicas de reamostragem *bootstrap*, *bagging* e *subagging* são apresentados. Aplica-se duas técnicas de reamostragem conhecidas como *bagging* e *subbaging* aos dados de aço-ligas com objetivo aprimorar as predições das concentrações dos analitos. É feita a regressão linear múltipla para a comparação dos resultados produzidos. O melhor resultado é obtido com a técnica MLR-*subagging* que apresenta melhora em média de 51,09% em relação à MLR-tradicional e 16% em relação a MLR-*bagging*. No entanto, observa-se que o MLR-*subagging* tem custo computacional maior em torno de 72,25% em relação ao MLR-*bagging*.

5. Referências Bibliográficas

1. BEEBE, K.R; PELL, R.J; SEASHOLTZ, M.B. ***Chemometrics: A Practical Guide***, Editora John Wiley & Sons, New York, 1998.

2. BRAGA, J.W.B. **Aplicação e validação de modelos de calibração de segunda ordem em química analítica**. 2008.300f. Tese (Doutorado em Química) Universidade Estadual de Campinas, Instituto de Química, Campinas.

3. BREIMAN, L. **Bagging predictors**. *Machine Learning*, 1996.

4. BROWN, S. D. *Has the chemometrics revolution ended? Some views on the past, present and future of chemometrics.* ***Chemometrics and Intelligent Laboratory Systems*** **30**, 49-58, 1995.

5. CHATTERJEE, S; ALI S. H; PRICE, B. ***Regression Analysis by Example***. Editora John Wiley and Sons, Inc., 3.ed., 2000.

6. DRAPER, N.R; SMITH, H. ***Applied Regression Analysis***. Editora John Wiley and Sons, Inc.,3.ed., 1998.

7. EFRON, B; TIBSHIRANI, R. ***An introduction to the bootstrap***. Editora Chapman e Hall, 1993.

8. FERREIRA, M. M. C; ANTUNES, A. M; MARISA, S. M et al. Quimiometria I: Calibração Multivariada, um tutorial. **Química Nova**, v. 22, n. 5, p. 724–731, Nov.1999.

9. FILHO, A. R. G. **Técnicas de Reamostragem em Calibração Multivariada**. 2007. 42f. Monografia Universidade Católica de Goiás, Goiânia, 2007.

10. FOX, *John. Bootstrapping Regression Models:* ***Appendix to An R and S-PLUS Companion to Applied Regression*** .Jan.2002.

11. GAHÃO, S.R; TURKMAN, M.A.A. Predição bootstrap via amostragem Gibbs do montante anual de indemnizações. **Actas do XIV Congresso Anual da SPE**.

12. GALVÃO, R. K. H., Araújo, M. C. U., and et al., M. N. M. *An application of subagging for the improvement of prediction accuracy of multivariate calibration models*. **Chemometrics and Intelligent laboratory systems**, vol.81, n.1, p. 60-67, 2006.

13. GARCIA, L. A. C et al. Método bootstrap na estimação de parâmetros genéticos populacionais. ***Scientia*** **Agricola**, v.58, n.4, p.785-793, out./dez. 2001.

14. HÄRDLE, W; SIMAR, L. ***Applied Multivariate Statistical Analysis***. Editora Tech, 2003.

15. JOHNSON, R.A; WICHERN, D. W. ***Applied Multivariate Statistical Analysis***. Editora Prentice Hall,5.ed.,2001.

16. LAWSON, C. L; HANSON, R. J.***Solving Least Squares Problems***. Englewood Cliffs, Editora NJ: Prentice-Hall, 1974.

17. LEVINE, D. M; STEPHAN D.F; KREHBIEL T.C; BERENSON, M.L. Estatística - Teoria e Aplicações. Editora LTC, 2005.

18. MARTENS, H., NAES, T. ***Multivariate Calibration***, John Wiley & Sons, London, 1989. NETO, B. B; SCRAMINIO, I. S; BRUNS, R. E, 25 anos de quimiometria no Brasil, **Química Nova**, Vol. 29, No. 6, 1401-1406, 2006.

19. PAKYARI, R. *On Bagging and Estimation in Multivariate Mixtures*. Metodoloski zvezki,, Vol. 5, No. 1, p.9-18, 2008.

20. SKOOG, D. A; LEARY, J. J. ***Principles of Instrumental Analysis,*** Saunders College Publishing, Philadelphia, 4.ed., 1991.

21. SANTOS, R.G. S; COELHO, C.J. Estudo da Relação Linear entre Conjuntos de Dados Químicos Empregando Análise de Correlação Canônica. **Revista de Iniciação Científica da Sociedade Brasileira de Computação**. Dezembro, 2010.

22. SENA, M. M; POPPI, R. P. Avaliação do uso de métodos quimiométricos em análise de solos, **Química Nova**, v. 23, n. 4. P. 547-556, 2000.

23. TIMM, N. H. ***Applied Multivariate Analysis.*** Editora Spring-Verlag, 2002.

24. WOLD, S. SJÖSTRÖM, M. Chemometrics, present and future success*,* ***Chemometrics and Intelligent Laboratory Systems***, v. 44, p. 3-14, 1998.