ECLÉTICA química
www.scielo.br
Volume 36, número 4, 2011
Artigo/Article

SELEÇÃO DE VARIÁVEIS PARA CATEGORIZAÇÃO DE AMOSTRAS QUÍMICAS

M. J. Anzanello (anzanello@producao.ufrgs.br)
Departamento de Engenharia de Produção e Transportes – Universidade Federal do Rio Grande do Sul
Av. Osvaldo Aranha, 99, 5 andar, Porto Alegre, RS, 90035-190

This paper presents a method to select the best variables to categorize chemical samples into two classes, say conforming or non-conforming. For that matter, PLS regression is combined with a data mining tool, the k-Nearest Neighbor classification technique, through an iterative variable selection process. The recommended subset of variables is chosen based on several criteria: sensitivity, specificity and percent of retained variables. When applied to two datasets related to wine analysis and one associated to QSAR, the proposed method significantly reduced the number of variables required for classification, while yielding superior categorization performance when compared to using all original variables.
Keywords: variable selection; chemical samples; PLS regression

Introdução

Sistemáticas voltadas à seleção de variáveis com propósito de predição têm encontrado vasta aplicação em processos químicos. O objetivo geral consiste na identificação de um subconjunto de variáveis independentes (geralmente associadas a níveis ou propriedades químicas dos reagentes) que conduzam à melhor predição da variável dependente (usualmente característica ou especificação de um produto ou substância). Ferramentas como regressão PLS (Partial Least Squares) e algoritmo genético são algumas das técnicas utilizadas com esse propósito [1,2,3].

Os parâmetros da regressão PLS têm oferecido suporte na identificação das variáveis mais relevantes para predição em processos químicos distintos: reciclagem de papel [1], fabricação de aço [4] e produção de antibióticos, látex e óxido de titânio [2,5]. A regressão PLS também tem sido aplicada na seleção de variáveis em dados espectrais de análises químicas, os quais são usualmente caracterizados por milhares de variáveis independentes. Nestas abordagens, os parâmetros PLS permitem minimizar o impacto de variáveis irrelevantes em modelos de predição [6,7]. Com objetivos similares, áreas relacionadas à QSAR (Quantitative Structure-Activity Relationship) têm utilizado métodos de seleção baseados em PLS para identificar as variáveis independentes com maior impacto na descrição das propriedades de um composto ou produto (descrição de propriedades biológicas, reativas, níveis de toxicidade e atividade biológica ou química) [1,8]. Abordagens baseadas em algoritmo genético também têm sido utilizadas na seleção de variáveis em QSAR, como em Ferreira et al. [3], Gao et al. [9], Sagrado e Cronin [10] e Tang e Li [11].

Diversas aplicações práticas, no entanto, priorizam a categorização de uma amostra em classes de acordo com determinada especificação (como qualidade final do produto, nível de toxidade de um composto químico e nível de reatividade de uma substância, entre outros), em detrimento à predição da variável dependente. Nesta natureza de aplicação, a disponibilidade de um conjunto reduzido de variáveis relevantes é fundamental para categorizações precisas. A correta classificação de amostras químicas em classes, especialmente durante os primeiros estágios de produção, permite ajustes sobre os parâmetros do processo com vistas à correção de inconsistências no mesmo.

Métodos de seleção de variáveis com propósitos de classificação, no entanto, são incipientes na lit-

eratura quando comparados aos métodos voltados à predição [12]. Este artigo apresenta um método para seleção de variáveis com vistas à categorização de amostras químicas em duas classes com base em diversas medidas de desempenho de classificação. Para tanto, aplica-se a regressão PLS em dados históricos de amostras, de maneira a capturar as relações entre as variáveis dependentes e independentes. O coeficiente da regressão PLS é então utilizado como indicador da importância de cada variável independente [2]. Na sequência, inicia-se um procedimento iterativo de eliminação das variáveis independentes com base nos coeficientes PLS e classificação das amostras após cada eliminação de variável. O procedimento de classificação é operacionalizado através da ferramenta k-Nearest Neighbor (KNN). A precisão de categorização, medida após cada eliminação de variável, é avaliada através de sensibilidade e especificidade. O subconjunto de variáveis com o melhor compromisso entre sensibilidade, especificidade e percentual de variáveis retidas é escolhido como solução final. Ao ser aplicado em três bancos de dados (análises químicas de vinhos e QSAR), o método proposto reduziu significativamente o número de variáveis a serem utilizadas e aumentou a precisão de classificação frente à utilização de todas as variáveis.

Este artigo inova através da integração da técnica multivariada de regressão PLS à ferramenta de classificação KNN com vistas à categorização de amostras químicas em classes. Outra contribuição vem da utilização de diversos critérios para seleção do melhor subconjunto de variáveis: sensibilidade, especificidade e percentual de variáveis retidas. Grande parte das abordagens para identificação de variáveis utiliza apenas um critério de seleção, seja acurácia em procedimentos de classificação [13] ou soma dos erros em procedimentos de predição [2].

O restante deste artigo é organizado como segue. A Seção 2 apresenta os fundamentos da regressão PLS e KNN, enquanto que a Seção 3 descreve a metodologia sugerida. Exemplos númericos são apresentados na Seção 4, e uma conclusão encerra o artigo na Seção 5.

Referencial teórico

Essa seção descreve os fundamentos das ferramentas utilizadas no método proposto: regressão PLS e ferramenta de classificação KNN.

A regressão PLS é um modelo de regressão multivariado que relaciona as matrizes de variáveis X (independentes) e Y (dependentes). Tal regressão apresenta vantagens quando comparada à tradicional regressão linear múltipla, visto que não é afetada por variáveis altamente correlacionadas, elevados níveis de ruído e observações faltantes [14]. A regressão PLS também é recomendada em situações em que o número de variáveis é superior ao número de observações [1,15].

A regressão PLS é operacionalizada como segue. Geram-se L componentes ($t_1,t_2,\ldots,t_L$), onde $t_1 = w_{11}\, x_1 + w_{12}\, x_2 + \ldots + w_{1z} x_z = w_1 \not\in x$ é o primeiro componente da matriz X (com N observações e Z variáveis), e $w_1 = (w_{11}, w_{12}, \ldots, w_{1z})$ é o vetor de pesos do componente t_1. Uma relação similar é construída para a matriz Y (N observações ´ M variáveis), onde o primeiro componente é representado por $u_1 = c_{11}\, y_1 + c_{12}\, y_2 + \ldots + c_{1z} y_z = c_1 \not\in y$, e $c_1 = (c_{11}, c_{12}, \ldots, c_{1z})$ representa o vetor de pesos do componente u_1. Os vetores w_1 e c_1 são estimados através da maximização da covariância entre as combinações lineares de X e Y, restrito à condição de ortogonalidade de w [1].

De acordo com Wold et al.[1], os componentes t e u podem ser manipulados de forma a gerar o coeficiente de regressão b_{mz} [Eq. (1)], similar ao coeficiente da regressão linear múltipla. A magnitude de b_{mz} sinaliza a relevância da variável independente z para a explicação da variância da variável dependente m. l representa o número de componentes retidos para análise. De maneira geral, dois ou três componentes explicam a maior parte da covariância, justificando a larga utilização de PLS como técnica de redução dimensional em bancos de dados caracterizados por elevado número de variáveis. Maiores detalhes sobre a estrutura matemática da regressão PLS podem ser obtidos em Wold et al. [1,16].

$$b_{mz} = \sum_{a=1}^{l} c_{ma}\, w_{za} \qquad \text{m=1,..,M e z=1,..,Z} \tag{1}$$

A ferramenta KNN, por sua vez, tem sido amplamente utilizada em procedimentos de classificação por conta de sua simplicidade e robustez. KNN insere uma nova amostra à classe (categoria) com maior número de incidências entre as k amostras mais próximas. Considere N amostras em uma porção de treino com dimensões definidas pelas Z variáveis independentes. Objetiva-se classificar uma nova observação como 0 ou 1 (conforme ou não-conforme, respectivamente), utilizando somente variáveis independentes. A ferramenta KNN calcula a distância Euclidiana entre a nova amostra e as k amostras mais próximas. A classe das k amostras mais próximas é conhecida, 0 ou 1. A

nova amostra é então classificada como 0 se a maioria das k amostras mais próximas pertencem à classe 0. O parâmetro k pode ser obtido através de validação cruzada na porção de treino, maximizando-se indicadores de performance de classificação como sensibilidade e especificidade, entre outros. Maiores detalhes sobre KNN podem ser obtidos em Duda et al. [17], enquanto que exemplos de aplicações são encontrados em Golub et al. [18], Weiss et al. [19] e Chaovalitwongse et al. [20].

Método

O método para seleção de variáveis com vistas à categorização de amostras químicas inicia com a aplicação da regressão PLS nos dados históricos. Cada observação do banco de dados representa uma amostra descrita por Z variáveis independentes (componentes ou substâncias químicas) e uma única variável dependente (usualmente uma especificação do produto medida em unidade apropriada). Cada amostra é categorizada em uma classe (não-conforme e conforme, por exemplo), valendo-se de um ponto de corte na variável dependente (esse ponto pode ser definido por especialistas de processo). Na sequência, as observações são randomicamente divididas em duas porções: a porção de treino é utilizada para identificar as variáveis mais importantes e a porção de teste representa novas amostras a serem categorizadas. Recomenda-se uma proporção de 60% para porção de treino e 40% para porção de teste [21]. A regressão PLS é então aplicada na porção de treino, visando capturar a relação entre as variáveis independentes e dependente. Os dados devem ser normalizados antes da aplicação de PLS para eliminar efeitos de escala.

Os parâmetros de interesse gerados pela regressão PLS são os coeficientes b_{mz} [ver equação (1)], os quais caracterizam a intensidade da relação entre as variáveis independentes e dependente. Tendo-se em vista que os dados históricos são normalizados antes da aplicação da regressão, o valor absoluto de b_{mz} pode ser utilizado para identificar as variáveis independentes que mais afetam a variância da variável dependente (elevados valores de b_{mz} sinalizam variáveis independentes mais importantes). De acordo com Duda et al. [17], variáveis com elevada variância conduzem a classificações mais precisas.

A próxima etapa do método consiste na eliminação das variáveis irrelevantes da porção de treino. Para tanto, as observações consistindo de Z variáveis independentes são classificadas como conformes ou não-conformes através de KNN, e medidas de sensibilidade e especificidade de classificação são calculadas. Essas medidas de precisão de classificação são definidas como segue [20]. Considere quatro possibilidades de classificação: 1) Positivos verdadeiros (PV), os quais denotam a correta classificação de amostras conformes, 2) Negativos verdadeiros (NV), indicando a correta categorização de amostras não-conformes; 3) Positivos Falsos (PF), indicando a equivocada classificação de amostras não-conformes na categoria conforme; e 4) Negativos Falsos (NF), indicando a equivocada categorização de amostras conformes na categoria não-conforme. Sensibilidade é definida como a fração de casos conformes corretamente categorizados, de acordo com a Eq. (2); similarmente, especificidade é dada pela fração de casos não-conformes corretamente categorizados, conforme a Eq. (3).

$$Sensibilidade = \frac{PV}{PV + NF} \quad (2)$$

$$Especificidade = \frac{NV}{NV + PF} \quad (3)$$

Na sequência, remove-se a variável com o menor valor absoluto de b_{mz} e classifica-se novamente a porção de treino consistindo de Z-1 variáveis. A sensibilidade e especificidade de classificação são novamente calculadas. Esse processo de eliminação e classificação é repetido até atingir-se um número mínimo de variáveis remanescentes; recomenda-se a retenção de duas variáveis para assegurar a consistência da ferramenta de classificação KNN.

Após concluir-se o processo de eliminação das variáveis, constrói-se um gráfico associando as medidas de sensibilidade e especificidade ao percentual de variáveis retidas. Cada ponto deste gráfico, referente ao desempenho de classificação decorrente da eliminação de uma variável, tem sua distância Euclidiana calculada em relação a um ponto do gráfico tido como ideal. As coordenadas do ponto ideal são definidas pelo usuário e devem ser coerentes com os critérios analisados: valores próximos a 1 para sensibilidade e especificidade e valores próximos a 0 para o percentual de variáveis retidas. O ponto com a menor distância ao ponto ideal identifica o melhor subconjunto de variáveis a ser retido.

Por fim, as amostras da porção de teste são categorizadas utilizando as variáveis selecionadas, calculando-se então sensibilidade e especificidade para aquela porção de dados.

Resultados e discussão

O método proposto foi aplicado em três bancos de dados de processos químicos, detalhados na Tabela 1. Os dois primeiros referem-se a análises de amostras de vinho tinto e vinho branco, sendo descritas por 11 variáveis independentes: (x_1) ácidos fixos, (x_2) acidez volátil, (x_3) ácido cítrico, (x_4) açúcar residual, (x_5) cloretos, (x_6) dióxido de enxofre livre, (x_7) dióxido de enxofre total, (x_8) densidade, (x_9) pH, (x_{10}) sulfatos, e (x_{11}) álcool. A variável de resposta, y, é uma escala de qualidade compreendida entre 0 (menor qualidade) e 10 (maior qualidade). O terceiro banco de dados refere-se à análise QSAR de aminoácidos, sendo descrito por variáveis independentes relacionadas à área de superfície e volume molecular do aminoácido, entre outras. A variável dependente refere-se à energia livre para desdobramento de uma proteína específica. Os dois primeiros bancos de dados estão disponíveis em Cortez et al. [22], enquanto que o terceiro está disponível em Wold et al. [1]. Por conta do reduzido número de amostras, gera-se somente a porção de treino para o banco QSAR.

Tabela 1: Bancos de dados analisados

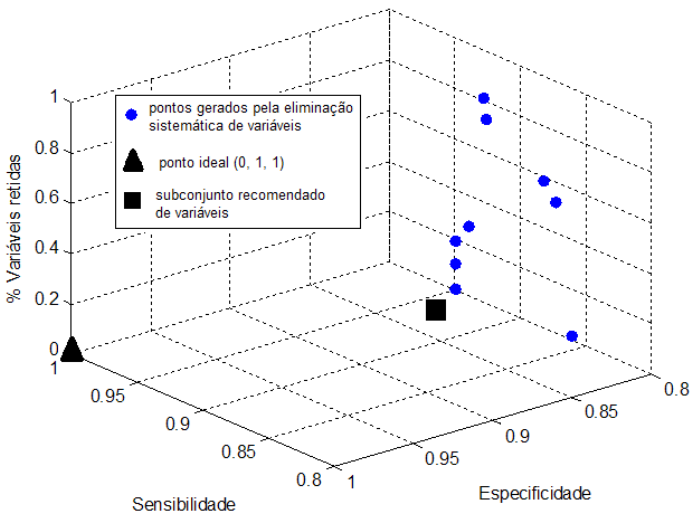
Banco de dados	Número de variáveis independentes	Número de observações	
		Porção de treino	Porção de teste
Vinho tinto	11	1000	599
Vinho branco	11	3500	1398
QSAR	7	19	-

As observações de cada banco de dados (representando amostras) foram classificadas em dois níveis de qualidade, de acordo com especificações das variáveis y (amostra conforme =1, amostra não-conforme=0). As especificações para os bancos relativos aos processos de produção de vinho foram levantadas junto a especialistas do setor, os quais definem que um vinho com escala de qualidade igual ou superior a seis é tido como de elevada qualidade ou conforme (ou seja, amostras de vinho com escala inferior a 6 são classificadas como não-conformes). Procedimento similar foi desdobrado para o banco QSAR, cuja especificação procede de Wold et al. [1].

A regressão PLS foi então aplicada à porção de treino dos processos da Tabela 1. Três componentes PLS foram retidos em cada banco de dados, com base no percentual de variância explicada em Y (Vinho tinto=81%, Vinho branco=77%, QSAR=70%). O parâmetro k da ferramenta KNN foi determinado por intermédio de 15 repetições de validação cruzada, sendo que cada banco de dados foi dividido em 5 porções em cada procedimento [20]. Definiu-se k=3 para os três bancos de dados ao maximizar-se sensibilidade e especificidade simultaneamente. Todos os procedimentos computacionais foram realizados em Matlab 7.4®.

A Figura 1 traz o perfil de sensibilidade e especificidade gerado com a eliminação de variáveis para o banco de dados de vinho tinto. O ponto superior direito do gráfico refere-se à sensibilidade e especificidade quando todas as variáveis são utilizadas no procedimento de categorização; os pontos subsequentes denotam os mesmos indicadores à medida que as variáveis são removidas de acordo com a magnitude do coeficiente de regressão b_{mz} e novas classificações são efetuadas. O ponto realçado com um quadrado representa o melhor subconjunto de variáveis a ser retido, visto que detém a menor distância ao ponto ideal (0, 1, 1), representado por um triângulo no extremo esquerdo do gráfico.

Figura 1: Perfil de sensibilidade e especificidade com a eliminação de variáveis



Os valores de sensibilidade e especificidade para as porções de treino dos três bancos, bem como as variáveis selecionadas, são apresentadas na Tabela 2. Foram retidas apenas 3 variáveis para todos os casos. As variáveis retidas conduziram a categorizações 3,3%

mais precisas em termos de sensibilidade e 8% em termos de especificidade, considerando a média sobre os três bancos analisados. Esse incremento deve-se à eliminação de variáveis ruidosas e irrelevantes do procedimento de categorização. Além disso, percebe-se que as variáveis mais adequadas para categorização das amostras de vinho diferem entre si: as melhores variáveis para o vinho tinto são acidez volátil, sulfatos e álcool, ao passo que as amostras de vinho branco são mais precisamente classificadas pela acidez volátil, açúcar residual e álcool.

e especificidade na classificação da porção de teste (Tabela 3), frente à porção de treino (Tabela 2). Tal ocorrência é natural em procedimentos de classificação e predição em que os dados históricos são divididos em duas porções; as precisões são mais elevadas na porção em que o modelo foi gerado (porção de treino), visto que peculiaridades daqueles dados são capturadas pelo modelo. Ao aplicar-se o mesmo modelo sobre dados de teste, percebe-se que algumas peculiaridades dos dados daquela porção não são corretamente capturadas, diminuindo a acurácia de classificação. Tal comportamento é corroborado em [12, 17, 20, 23], entre outros.

Tabela 2: Desempenho de classificação na porção de treino

Banco de dados (number de variáveis originais)	Sensibilidade na porção de treino (%)		Especificidade na porção de treino (%)		Variáveis retidas para categorização
	Método proposto	Sem seleção de variáveis	Método proposto	Sem seleção de variáveis	
VINHO TINTO (11)	86,3	87,9	88,3	84,2	x_2, x_{10} e x_{11}
VINHO BRANCO (11)	92,2	91,0	76,5	74,8	x_2, x_4 e x_{11}
QSAR (7)	90,9	81,8	75,0	63,0	x_2, x_5 e x_7
Média	89,8	86,9	79,9	74,0	

Tabela 3: Precisão de classificação na porção de teste

Banco de dados (number de variáveis originais)	Sensibilidade na porção de teste (%)		Especificidade na porção de teste (%)	
	Método proposto	Sem seleção de variáveis	Método proposto	Sem seleção de variáveis
VINHO TINTO (11)	76,9	60,8	62,3	55,7
VINHO BRANCO (11)	83,0	77,6	48,2	37,7
Média	80,0	69,2	55,3	46,7

Por fim, as variáveis selecionadas são utilizadas na classificação de novas amostras, representadas pela porção de teste (Tabela 3). Verifica-se um aumento de 15,6% na sensibilidade e de 18,4% na especificidade ao utilizarem-se apenas as variáveis selecionadas no procedimento de categorização. É importante salientar que se verifica uma redução dos níveis de sensibilidade

Conclusão

Bancos de dados de processos compostos por elevado número de variáveis são amplamente encontrados em aplicações industriais e químicas, demandando

o desenvolvimento de métodos robustos para seleção das variáveis mais relevantes. A ampla diversidade de abordagens para seleção de variáveis com propósitos de predição contrasta, no entanto, com a reduzida disponibilidade de sistemáticas de seleção voltadas a procedimentos de classificação.

Este artigo reporta um método para seleção de variáveis com vistas à categorização de amostras químicas em duas classes com base em diversas medidas de desempenho de classificação. A regressão PLS é inicialmente aplicada em dados históricos de amostras, gerando o coeficiente da regressão PLS. Na sequência, as variáveis são eliminadas com base na magnitude daquele coeficiente e as amostras então classificadas através de KNN. O subconjunto de variáveis com o melhor compromisso entre sensibilidade, especificidade e percentual de variáveis retidas aponta a solução final. Ao ser aplicado em dois processos de elaboração de vinho e um processo do tipo QSAR, o método reduziu significativamente o número de variáveis a serem utilizadas e aumentou a precisão de classificação frente à utilização de todas as variáveis. Além disso, percebe-se que variáveis distintas devem ser utilizadas para classificar amostras de produtos semelhantes, como vinho branco e tinto.

Desdobramentos futuros do método proposto incluem a utilização de outras ferramentas de classificação, como Máquina de Suporte Vetorial e Árvores de Decisão. O desenvolvimento de índices robustos de importância para as variáveis também se constitui em tópico de interesse.

Referências

[1] S. Wold, M. Sjostrom, L. Eriksson, Chemometr. Intell. Lab. 58 (2001) 109.

[2] J. Gauchi, P. Chagnon, Chemometr. Intell. Lab. 58 (2001) 171.

[3] M. Ferreira, C. Montanari, A. Gaudio, Quím. Nova 25 (2002) 439.

[4] I. Chong, C. Jun, Chemometr. Intell. Lab. 78 (2005) 103.

[5] A. Lazraq, R. Cleroux, J. Gauchi, Chemometr. Intell. Lab. 66 (2003) 117.

[6] A. Hoskuldsson, Chemometr. Intell. Lab. 55 (2001) 23.

[7] L. Xu, J. Jiang, H. Wu, G. Shen, R. Yu, Chemometr. Intell. Lab. 85 (2007) 140.

[8] H. Zhai, X. Chen, Z. Hu, Chemometr. Intell. Lab. 80 (2006) 130.

[9] H. Gao, M. Lajiness, J. Drie, J. Mol. Graph. Model. 20 (2002) 259.

[10] S. Sagrado, M. Cronin, Anal. Chim. Acta, 609 (2008) 169.

[11] K. Tang, T. Li, Chemometr. Intell. Lab. 64 (2002) 55.

[12] M. Anzanello, S. Albin, W. Chaovalitwongse, Chemometr. Intell. Lab. 97 (2009) 111.

[13] M. Anzanello, Gestão & Produção 16 (2009) 1.

[14] P. Nelson, J. MacGregor, P. Taylor, Chemometr. Intell. Lab. 80 (2006) 1.

[15] H. Abdi, H. *Partial Least Squares (PLS) Regression*, in Encyclopedia of Social Sciences Research Methods. Thousand Oaks: Sage, 2003.

[16] S. Wold, J. Trygg, A. Berglund, H. Antii, Chemometr. Intell. Lab. 58 (2001) 131.

[17] R. Duda, P. Hart, D. Stork, Pattern Classification, New York: Wiley-Interscience, 2nd ed., 2001.

[18] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomeld, E. Lander, Science, 286 (1999) 531.

[19] S. Weiss, C. Apte, D. Dameray, D. Johnson, F. Ples, T. Goetz, T. Hampp, IEEE Intell Syst. 14 (1999) 63.

[20] W. Chaovalitwongse, Y. Fan, C. Sachdeo, IEEE Trans Syst Man Cy A 37 (2007) 1005.

[21] I. Chong, S. Albin, C. Jun, IIE Trans. 39 (2007) 795.

[22] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis. Decis. Support. Syst. 47 (2009) 547. Disponível em http://archive.ics.uci.edu/ml/datasets/Wine+Quality ou http://www3.dsi.uminho.pt/pcortez (Acesso em 05/2011).

[23] X. Wu, V. Kumar, J, Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand, D. Steinberg. Knowl. Inf. Syst. 14