PRODUÇÃO
PRODUCTION

Produção, v. 23, n. 4, p. 858-865, out./dez. 2013
http://dx.doi.org/10.1590/S0103-65132013005000001

Seleção de variáveis para classificação de bateladas produtivas com base em múltiplos critérios

Michel José Anzanello[a*]

[a*]michel.anzanello@gmail.com, UFRGS, Brasil

Resumo

Processos industriais são frequentemente descritos por um elevado número de variáveis correlacionadas e ruidosas. Este artigo apresenta um método para seleção das variáveis mais relevantes para classificação de bateladas de produção valendo-se de múltiplos critérios de desempenho (sensibilidade e especificidade). As bateladas são categorizadas em duas classes (conforme ou não conforme, por exemplo). O método utiliza a regressão PLS (*Partial Least Squares*) para derivar um índice de importância das variáveis de processo. Um procedimento iterativo de classificação das bateladas e eliminação das variáveis é então conduzido. Por fim, uma medida de distância euclidiana ponderada é aplicada para selecionar o melhor subconjunto de variáveis. Ao ser aplicado em dados de processos industriais, o método proposto reteve, em média, 12% das variáveis originais, elevando a sensibilidade em 9%, de 0,78 para 0,85, e a especificidade em 20%, de 0,64 para 0,77. Estudos de simulação permitiram avaliar o desempenho do método frente a cenários distintos.

Palavras-chave

Seleção de variáveis. Múltiplos critérios. Regressão PLS.

1. Introdução

O elevado volume de dados coletados em processos industriais tem incentivado o desenvolvimento de métodos para seleção das variáveis mais relevantes. A maioria dos estudos tem selecionado variáveis de processo (temperatura, pressão e concentração de componentes, entre outras) com vistas à predição de uma ou mais variáveis de produto (GAUCHI; CHAGNON, 2001; MEIRI; ZAHAVI, 2006; OZTURK; KAYALIGIL; OZDEMIREL, 2006; OLAFSSON; LI; WU, 2008). O objetivo deste estudo, no entanto, é selecionar as variáveis de processo mais relevantes com vistas à categorização de bateladas de produção valendo-se de múltiplos critérios de desempenho de classificação, como sensibilidade e especificidade.

O critério de desempenho tradicionalmente utilizado em procedimentos de seleção de variáveis com propósito de classificação é a acurácia, definida como a fração das bateladas corretamente classificadas (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009). No entanto, existem situações em que outros critérios são mais apropriados. Na indústria farmacêutica, por exemplo, a incorreta classificação de uma batelada não conforme como conforme pode acarretar sérias consequências. Neste caso, o critério especificidade (fração de bateladas não conformes corretamente classificadas) deve ser avaliado em detrimento à acurácia. Por outro lado, cenários onde a classificação equivocada de uma batelada conforme pode acarretar impactos financeiros elevados devem ser analisados através do critério sensibilidade (fração de bateladas conformes corretamente classificadas). No método proposto, o custo para coleta e análise de variáveis também pode ser avaliado como critério para seleção das variáveis, juntamente com sensibilidade e especificidade.

Abordagens para a seleção de variáveis utilizando múltiplos critérios têm sido sugeridas em aplicações de reconhecimento de texto, análise financeira e sistemas de segurança (ROSE-PEHRSSON et al., 2000; DOAN; HORIGUCHI, 2004; PIRAMUTHU, 2004; PENDARAKI;

*UFRGS, Porto Alegre, RS, Brasil
Recebido 16/02/2011; Aceito 06/08/2012

ZAPOUNIDIS; DOUMPOS, 2005; HUANG; TZENG; ONG, 2006; PASIOURAS et al., 2007; ARAGONÉS-BELTRÁN et al., 2008). Uma revisão abrangente sobre critérios múltiplos com foco em tomada de decisão é apresentada em Zopounidis e Doumpos (2002) e Sueyoshi (2006). A utilização de tais critérios em cenários industriais, no entanto, não tem encontrado aplicação recente.

O método proposto para seleção de variáveis com base em múltiplos critérios de desempenho é operacionalizado como segue. A regressão PLS (*Partial Least Squares*) é inicialmente aplicada na porção de treino de um banco de dados para geração de um índice de importância para cada variável de processo. Em seguida, as observações (bateladas) da porção de treino descritas por todas as variáveis são categorizadas em duas classes através da ferramenta *k-Nearest Neighbor* (KNN). Critérios para medição do desempenho de classificação (sensibilidade, especificidade, custo das variáveis retidas etc.) são computados. Na sequência, a variável com o menor índice de importância é eliminada, uma nova classificação é realizada utilizando as variáveis remanescentes, e o desempenho de classificação é reavaliado. Tal procedimento iterativo é mantido até que apenas uma variável reste. Uma análise de Pareto Ótimo aponta os subconjuntos de variáveis candidatas à melhor solução. O subconjunto ideal é identificado com base em uma medida de distância ponderada de cada subconjunto candidato a um ponto hipotético tido como ideal. Por fim, as variáveis selecionadas são validadas na porção de teste. Ao ser aplicado em seis bancos de dados industriais, o método proposto aumentou significativamente o desempenho de classificação (sensibilidade e especificidade) utilizando um percentual reduzido de variáveis. Experimentos de simulação permitem avaliar o desempenho do método frente a distintos níveis de ruído, colinearidade e proporção "número de observações/número de variáveis" nos bancos de dados.

O artigo está organizado como segue. A seção 2 traz os fundamentos da regressão PLS e da ferramenta de classificação KNN, enquanto a seção 3 descreve o método para seleção de variáveis com base em múltiplos critérios. A seção 4 apresenta os resultados do método aplicado em dados industriais reais. Uma conclusão é apresentada na seção 5.

2. Referencial teórico

A regressão PLS (*Partial Least Squares*), a exemplo da análise de componentes principais, gera um reduzido número de combinações lineares independentes das variáveis de processo. Essas novas variáveis, chamadas de componentes PLS, respondem pela maior parte da variância presente nas variáveis originais do processo. Normalmente, apenas três ou quatro componentes PLS são retidos para representar dezenas ou mesmo centenas de variáveis de processo.

Os principais parâmetros que resultam de regressão PLS são pesos e cargas. Esses parâmetros podem ser calculados através do algoritmo NIPALS; ver Goutis (1997), Abdi (2003) e Geladi e Kowalski (1986). Detalhes matemáticos da regressão PLS podem ser obtidos em Westerhuis, Kourti e MacGregor (1998), Wold, Sjostrom e Eriksson (2001) e Wold et al. (2001). A regressão PLS pode ser operacionalizada através do *toolbox* PLS, encontrado em pacotes estatísticos como Matlab® e R®.

Os fundamentos da regressão PLS são agora apresentados. Considere uma matriz X com n observações para cada uma das J variáveis de processo e uma matriz Y com n observações para cada uma das M variáveis de produto. As variáveis de processo e produto referentes a uma batelada i são representadas pelos vetores $\mathbf{x}_i$ $(x_{i1}, x_{i2}, ..., x_{iJ})$ e $\mathbf{y}_i$ $(y_{i1}, y_{i2}, ..., y_{iM})$, respectivamente.

A regressão PLS gera A combinações lineares (componentes) das variáveis de processo, $t_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + ... + w_{ja}x_{ij} = \mathbf{w}_a'\mathbf{x}_i$, com $A \leq J$. O número de componentes, A, é geralmente pequeno, e pode ser definido através do método de validação cruzada proposto em Hoskuldsson (1988). O vetor $\mathbf{w}_a = w_{1a}, w_{2a}, ..., w_{ja}$ quantifica a influência de cada variável na composição do componente e, por consequência, na composição das variáveis de processo e produto (WOLD; SJOSTROM; ERIKSSON, 2001). Componentes são similarmente construídos para as variáveis de produto Y, ou seja, $u_{ia} = c_{1a}y_{i1} + c_{2a}y_{i2} + ... + c_{Ma}y_{iM} = \mathbf{c}_a'\mathbf{y}_i$, onde $\mathbf{c}_a = (c_{1a}, c_{2a}, ..., c_{Ma})$ é o peso das variáveis de produto.

Os vetores de peso $\mathbf{w}_a$ e $\mathbf{c}_a$ são estimados com vistas à maximização da covariância entre os componentes $\mathbf{t}_a$ and $\mathbf{u}_a$. Tais pesos são ortogonais entre si, garantindo a independência dos componentes gerados. Por fim, o vetor de carga, $\mathbf{p}_a = (p_{1a}, p_{2a}, ..., p_{Ja})$ é gerado pela regressão das colunas de X em relação a $\mathbf{t}_a$, fornecendo informações relevantes sobre as variáveis do processo.

A regressão PLS pode ser utilizada em situações onde as variáveis de processo apresentam elevados níveis de correlação, ruído, observações faltantes e desbalanço na proporção de variáveis e observações. Tais condições são frequentemente encontradas em aplicações industriais; ver Wold, Sjostrom e Eriksson (2001), Kettaneh, Berglund e Wold (2005), Nelson, MacGregor e Taylor (2006) e Hoskuldsson (2001).

Por sua vez, a ferramenta de classificação *k-Nearest Neighbor* (KNN) encontra ampla utilização por conta de sua simplicidade conceitual e disponibilidade em pacotes estatísticos. Algumas aplicações incluem a classificação de genes em Golub et al. (1999),

reconhecimento de texto em Weiss et al. (1999) e detecção de atividade cerebral anormal em Chaovalitwongse, Fan e Sachdeo (2007). Na KNN existem N observações em um conjunto de dados de treino composto por J variáveis de processo. O objetivo é classificar uma nova observação em conforme ou não conforme (1 ou 0, respectivamente), utilizando-se apenas as variáveis do processo. O algoritmo KNN mede a distância euclidiana entre a nova observação e os k vizinhos mais próximos (ou seja, observações já existentes). A classe de cada um dos k vizinhos é previamente conhecida, 0 ou 1. Uma nova observação é classificada como 0 se a maioria dos seus vizinhos mais próximos pertencer a 0. O número de vizinhos k, é definido através da maximização de uma medida de desempenho de classificação na porção de treino, onde a classe de cada observação é conhecida. Mais detalhes sobre KNN podem ser encontrados em Ridgeway (2003).

3. Método

O método proposto para seleção de variáveis utilizando múltiplos critérios de desempenho é operacionalizado em quatro passos: (1) Aplicação da regressão PLS no banco de dados e geração de um índice de importância da variável de processo; (2) Categorização das bateladas em duas classes e eliminação das variáveis irrelevantes; (3) Geração do perfil de desempenho e aplicação da análise de Pareto Ótimo; e (4) Identificação da melhor solução da fronteira do Pareto. Tais passos são detalhados na sequência.

O primeiro passo aplica a regressão PLS no banco de dados e gera um índice de importância de cada variável de processo. Para tanto, considere dados de entrada para N bateladas. Uma batelada i=1, ..., N é descrita pelo vetor de variáveis de processo $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iJ})$, enquanto a variável de produto é descrita por $\mathbf{y}_i$. O banco de dados é randomicamente dividido em duas porções: treino (N_{tr}) e teste (N_{ts}), com $N = N_{tr} + N_{ts}$. Recomenda-se manter 60% das observações na porção de treino (CHONG; ALBIN; JUN, 2007).

A regressão PLS é aplicada na porção de treino (N_{tr}). Os parâmetros de interesse gerados pela regressão incluem os pesos w_{ja}, as cargas p_{ja} e o percentual de variância em Y explicado pelo a-ésimo (a = 1, ..., A) componente retido, R^2_{Ya}. Tais parâmetros são utilizados na geração de um índice de importância das variáveis de processo com vistas à eliminação das variáveis ruidosas e menos relevantes. O índice de importância da variável j é definido como v_j, j = 1,..., J. Valores elevados de v_j indicam as variáveis mais importantes para propósitos de classificação (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

O índice de importância v_j é gerado na Equação 1, sendo que w^*_{ja} é obtido através da relação $w^*_{ja} = w_{ja}(p_{ja}w_{ja})^{-1}$. Wold, Sjostrom e Eriksson (2001) afirmam que o peso ajustado w^*_{ja} conduz a processos de seleção mais estáveis do que o peso original (w_{ja}). Detalhes sobre w^*_{ja} podem ser obtidos em Manne (1987).

$$v_j = \sum_{a=1}^{A} (w^*_{ja})^2 R^2_{Ya} \qquad j = 1,..., J. \tag{1}$$

Wold, Sjostrom e Eriksson (2001) inicialmente sugeriram o índice v_j para seleção de variáveis com propósitos de predição. No entanto, o mesmo apresentou resultados satisfatórios em procedimentos de seleção com vistas à classificação de bateladas produtivas (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2009).

A etapa seguinte consiste na eliminação das variáveis irrelevantes da porção de treino. Para tanto, as bateladas descritas por J variáveis independentes são classificadas como conformes ou não conformes através da ferramenta de classificação KNN, e múltiplos critérios de avaliação de desempenho de classificação (ex. sensibilidade e especificidade) são calculados. Tais critérios são definidos como segue.

Considere quatro possibilidades de classificação (CHAOVALITWONGSE; FAN; SACHDEO, 2007): 1) Positivos verdadeiros (PV), os quais denotam a correta classificação de bateladas conformes; 2) Negativos verdadeiros (NV), indicando a correta categorização de bateladas não conformes; 3) Positivos falsos (PF), indicando a equivocada classificação de bateladas não conformes na categoria conforme; e 4) Negativos falsos (NF), indicando a equivocada categorização de bateladas conformes na categoria não conforme. Sensibilidade é definida como a fração de bateladas conformes corretamente categorizadas, de acordo com a Equação 2; similarmente, especificidade é dada pela fração de bateladas não conformes corretamente categorizadas, conforme a Equação 3.

$$Sensibilidade = \frac{PV}{PV + FN} \tag{2}$$

$$Especificidade = \frac{NV}{NV + FP} \tag{3}$$

Na sequência, remove-se a variável com o menor valor absoluto de v_j e classifica-se novamente a porção de treino consistindo das J-1 variáveis remanescentes. A sensibilidade e especificidade de classificação são novamente calculadas. Esse processo de eliminação e classificação é repetido até que exista apenas uma variável remanescente.

Após concluir-se o processo de eliminação das variáveis, inicia-se o terceiro passo com a construção

de um gráfico associando os critérios de desempenho (sensibilidade, especificidade, entre outros), ou percentual de custo das variáveis retidas, ao percentual de variáveis retidas. Cada ponto deste gráfico refere-se ao desempenho de classificação ou custo das variáveis remanescentes decorrente da eliminação de uma variável. No caso de mais de três critérios serem considerados na análise, o gráfico é substituído por uma tabela descrevendo os valores de desempenho e percentual de variáveis retidas.

Na sequência, aplica-se a análise de Pareto Ótimo (PO) para identificar soluções diferenciadas. As soluções apontadas pela análise de PO são definidas como soluções "não dominadas" em aplicações caracterizadas por múltiplas funções objetivo, ou seja, soluções que não podem ser superadas por soluções vizinhas em termos dos objetivos avaliados. As soluções "não dominadas" são ilustradas em um contorno gráfico denominado fronteira do Pareto. Tal fronteira facilita a identificação da melhor solução (ou grupo de melhores soluções), visto que o conjunto de potenciais soluções é reduzido de forma significativa (HORN; NAFPLIOTIS; GOLDBERG, 1994; ZITZLER; THIELE, 1999; TABOADA; COIT, 2007, 2008).

No quarto passo, os pontos da fronteira do perfil de sensibilidade/especificidade têm suas distâncias euclidianas calculadas em relação a um ponto do gráfico tido como ideal. As coordenadas do ponto ideal devem ser coerentes com os critérios analisados: valores próximos a 1 para os critérios de desempenho de classificação e valores próximos a 0 para o percentual de variáveis retidas e custo dessas variáveis. Tais coordenadas são definidas pelo usuário. O ponto de fronteira com a menor distância ao ponto ideal é definido como ponto base (PB), e sua distância definida como d_{PB}. Um exemplo hipotético considerando três critérios (dois a serem maximizados, como sensibilidade e especificidade, e um a ser minimizado, como percentual de variáveis retidas) é apresentado na Figura 1. O ponto hipotético tido como ideal, neste caso, é representado por (1,1,0).

Na sequência, sugere-se a varredura dos pontos de fronteira vizinhos ao ponto PB. Com isso, pretende-se avaliar soluções que aumentem significativamente um ou mais critérios de desempenho por conta da inclusão de variáveis adicionais na solução final. Considere os F (f = 1, ..., F) pontos de fronteira contendo um ou mais critérios de desempenho superiores aos obtidos no ponto PB. Calcula-se a variação de desempenho do critério c, var_{cf}, com base no percentual de variáveis retidas no ponto f, conforme a Equação 4:

$$var_{cf} = \frac{crit_{cf} - crit_{cPB}}{ret_f - ret_{PB}} \qquad c = 1,..., C;\ f = 1,..., F \quad (4)$$

onde $crit_{cf}$ denota o critério de desempenho c avaliado no ponto f, $crit_{cPB}$ denota o valor daquele critério no ponto PB, ret_f é o percentual de variáveis retidas no ponto f e ret_{PB} indica o percentual de variáveis retidas no ponto PB. Tal relação é calculada para os C critérios de desempenho considerados.

Na sequência, calcula-se a distância euclidiana ponderada de cada ponto f, d_{wf}. Tal distância é calculada através da Equação 5 que, para fins de ilustração, considera os critérios sensibilidade (*sens*) e especificidade (*espec*). A Equação 5 pode ser facilmente desdobrada para mais de três critérios.

$$dw_f = \sqrt{\frac{(1-sens_f)^2 + (1-espec_f)^2 + (ret_f)^2}{var_{sensf} \times var_{especf}}} \qquad f = 1, ..., F \quad (5)$$

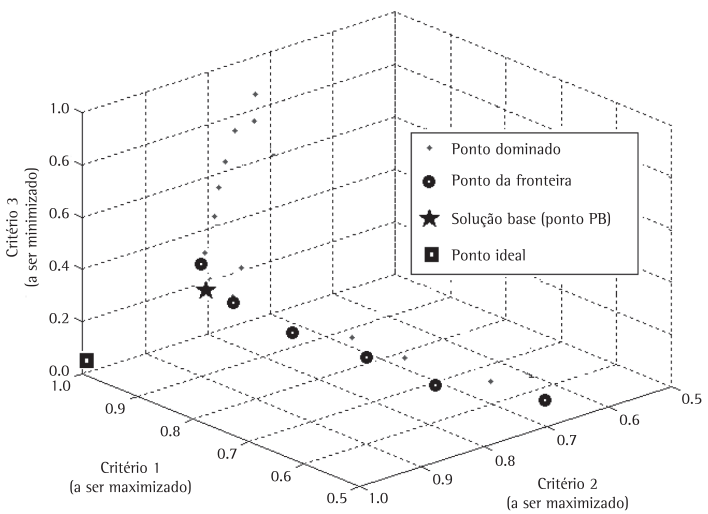

Figura 1. Perfil hipotético gerado com a eliminação sistemática de variáveis de processo.

A obtenção de valores negativos em var_f indica que o desempenho de classificação foi reduzido com a inclusão de variáveis adicionais. Tal ponto não deve ser avaliado na Equação 5. Além disso, a ordem da subtração no numerador da Equação 4 deve ser alterada quando o critério "custo das variáveis retidas" for analisado (quanto mais baixo o custo, melhor).

Por fim, o menor valor de dw_f ou d_{PB} identifica o ponto da fronteira associado ao melhor subconjunto de variáveis para classificação. As variáveis selecionadas são então utilizadas na categorização das bateladas da porção de teste (simbolizando novas observações), e o desempenho da classificação avaliada segundo os critérios utilizados no processo de seleção.

4. Aplicação em dados industriais

O método proposto é aplicado em dados de seis processos industriais obtidos em Gauchi e Chagnon (2001) e Wold, Sjostrom e Eriksson (2001). A Tabela 1 traz o nome de cada processo, natureza de aplicação, número de variáveis de processo e número de observações (bateladas) nas porções de treino e teste. Tais variáveis de processo referem-se a temperaturas, pressões e concentrações de reagentes químicos, enquanto a variável de resposta denota uma característica do produto, como viscosidade ou teor de pureza. As observações de cada banco de dados foram classificadas em dois níveis de qualidade (conforme ou não conforme), seguindo especificações na variável de resposta fornecidas por Gauchi e Chagnon (2001) e Wold, Sjostrom e Eriksson (2001).

Na sequência, aplicou-se a regressão PLS na porção de treino de cada processo, sendo retidos três componentes para cada processo através de validação cruzada (ver WOLD; SJOSTROM; ERIKSSON, 2001). A variância em **Y** explicada pelos três componentes foi 94% para ADPN, 75% para GRANU, 77% para LATEX, 94% para OXY, 68% para PAPER e 71% para SPIRA. Um procedimento de validação cruzada também foi utilizado pra definir o melhor parâmetro k para a ferramenta de classificação KNN. Os resultados são apresentados entre parênteses: ADPN (3), GRANU (3), LATEX (3), OXY (3), PAPER (3), e SPIRA (9).

A Tabela 2 apresenta os critérios de sensibilidade, especificidade e percentual de variáveis retidas na porção de treino em cada banco. Na mesma tabela, o método proposto é comparado ao Q^2_{cum}, originalmente proposto por Gauchi e Chagnon (2001) para seleção de variáveis com propósitos de predição e aqui adaptado para fins de classificação. O método proposto conduz a melhores resultados acerca de desempenho de classificação frente ao Q^2_{cum}, retendo menor percentual de variáveis na maioria dos bancos analisados. A justificativa para o melhor resultado frente ao método Q^2_{cum} decorre da integração da ferramenta KNN na classificação das bateladas e pela utilização de um índice de importância de variáveis mais consistente que o utilizado por Gauchi e Chagnon (2001), que ordenam suas variáveis valendo-se apenas do coeficiente de regressão PLS.

Para a porção de teste (composta por observações não utilizadas na geração do modelo), a sensibilidade foi elevada em 9%, de 0,78 para 0,85, enquanto que a especificidade aumentou 20%, de 0,64 para 0,77 (Tabela 3).

Tabela 1. Processos industriais analisados.

Banco de dados	Natureza de aplicação	Número de variáveis de processo	Número de observações	
			Porção de treino	Porção de teste
ADPN	Produção de nylon	100	57	14
GRANU	Emulção na indústria de papel	78	300	200
LATEX	Polimerização em um processo de látex	117	210	52
OXY	Produção de óxido de titânio	95	300	200
PAPER	Reciclagem de papel	54	192	192
SPIRA	Produção de antibióticos	96	115	29

Tabela 2. Desempenho do método proposto na porção de treino dos bancos de dados.

Banco de dados (número de variáveis originais)	Sensibilidade na porção de treino (%)			Especificidade na porção de treino (%)			Variáveis retidas (%)	
	Método proposto	Q^2 cum	Sem seleção de variáveis	Método proposto	Q^2 cum	Sem seleção de variáveis	Método proposto	Q^2 cum
ADPN (100)	99	98	93	71	56	63	7	13
GRANU (78)	94	90	88	90	93	88	24	9
LATEX (117)	97	93	94	92	83	80	7	21
OXY (95)	98	98	98	95	81	67	8	13
PAPER (54)	75	63	48	94	98	93	9	11
SPIRA (96)	96	77	96	92	84	79	19	16
Média	93	87	86	89	83	78	12	14

5. Experimentos de simulação

Por fim, a robustez do método proposto é analisada através de experimentos de simulação. Para tanto, geram-se bancos com base em dados reais de um processo de produção de látex. Assume-se que um modelo PLS é satisfatório para descrever tal processo. O banco de dados original é constituído por 100 variáveis independentes, uma variável dependente e 100 observações, sendo que cada observação representa uma batelada de produção.

As variáveis independentes são geradas de acordo com uma distribuição multinormal, com média e matriz de correlação extraídas do banco de dados de produção de látex. A regressão PLS é então aplicada aos dados originais para estimar os coeficientes de regressão PLS, sendo que três componentes são retidos no modelo. A variância do erro é estimada como sugerido em Denham (2000).

A simulação utiliza três fatores (com seus respectivos níveis entre parênteses) entendidos como relevantes para identificação de variáveis em processos industriais (JUN; CHONG, 2005): (*i*) variância do erro – (0,5 *ve*, *ve* e 1,5 *ve*); (*ii*) correlação entre as variáveis – (1/3 *cv*, 1 *cv* e 3 *cv*); e (*iii*) razão entre o número de observações e o número de variáveis – (0,2 *nv* e 10 *nv*). Os níveis nominais para variância do erro e correlação são extraídos do banco de dados do processo de látex. O terceiro fator apresenta dois níveis: razão 0,2, onde o número de observações é menor que o número de variáveis (característica de processos em batelada), e razão 10. Foram geradas 500 repetições por caso.

Os resultados da simulação são apresentados na Tabela 4, onde são calculadas a média e a variância para as 500 repetições de cada cruzamento de níveis dos fatores. Percebe-se uma diminuição dos critérios de desempenho de classificação (sensibilidade e especificidade) à medida que maior ruído é inserido no banco de dados. A adição de ruído reduz a precisão dos parâmetros gerados pela regressão PLS, afetando

Tabela 3. Desempenho do método proposto na porção de teste dos bancos de dados.

Banco de dados (número de variáveis originais)	Sensibilidade na porção de teste (%)			Especificidade na porção de teste (%)		
	Método proposto	Q^2 cum	Sem seleção de variáveis	Método proposto	Q^2 cum	Sem seleção de variáveis
ADPN (100)	100	90	100	62	50	25
GRANU (78)	95	86	87	78	80	80
LATEX (117)	93	67	81	71	71	73
OXY (95)	96	94	97	88	72	57
PAPER (54)	35	16	20	90	86	86
SPIRA (96)	92	75	83	75	83	65
Média	85	71	78	77	74	64

Tabela 4. Desempenho do método proposto nos experimentos simulados.

		0,5 nv								
		1/3 cv			cv			3cv		
		0,5 ve	ve	1,5 ve	0,5 ve	ve	1,5 ve	0,5 ve	ve	1,5 ve
Média	Sensibilidade	0,8508	0,7808	0,7496	0,7867	0,7717	0,7564	0,7686	0,7839	0,7548
	Especificidade	0,8299	0,7868	0,7484	0,7764	0,7663	0,7542	0,7777	0,7622	0,7498
	Variáveis retidas	0,0942	0,0883	0,0972	0,1194	0,1256	0,118	0,117	0,1286	0,1198
Variância	Sensibilidade	0,0125	0,0183	0,0219	0,0186	0,0197	0,017	0,0167	0,0167	0,0162
	Especificidade	0,0142	0,015	0,0178	0,0144	0,0189	0,0162	0,0167	0,0178	0,0174
	Variáveis retidas	0,0046	0,004	0,0065	0,0056	0,0087	0,0063	0,0063	0,0087	0,0065
		10 nv								
		1/3 cv			cv			3 cv		
		0,5 ve	ve	1,5 ve	0,5 ve	ve	1,5 ve	0,5 ve	ve	1,5 ve
Média	Sensibilidade	0,8809	0,7114	0,6639	0,7022	0,6512	0,6309	0,681	0,6343	0,6304
	Especificidade	0,8068	0,7162	0,6705	0,7024	0,6551	0,6404	0,6668	0,6519	0,6268
	Variáveis retidas	0,0822	0,0932	0,0964	0,122	0,109	0,107	0,1208	0,1129	0,1004
Variância	Sensibilidade	0,0038	0,0054	0,0059	0,0053	0,0072	0,0069	0,0059	0,0052	0,0061
	Especificidade	0,0039	0,0053	0,0056	0,0048	0,0069	0,0068	0,0055	0,0058	0,0069
	Variáveis retidas	0,0017	0,004	0,0045	0,0044	0,0048	0,0049	0,0053	0,0055	0,0046

o índice de importância das variáveis da Equação 1 e alterando a ordem de eliminação das variáveis no passo 2 do método proposto. O percentual de variáveis permanece praticamente inalterado com a adição de maiores níveis de ruído, porém variáveis menos relevantes deslocam outras mais relevantes no ordenamento de importância. Um maior ruído, como esperando, tende a elevar a variabilidade dos índices de desempenho (medida pela variância na Tabela 4), porém não estabelece um padrão nítido de influência sobre o percentual de variáveis retidas.

De maneira similar, o aumento da correlação entre as variáveis também reduz o desempenho de classificação do método proposto. Apesar de mais robusta do que a tradicional regressão linear múltipla, elevados níveis de colinearidade entre as variáveis prejudicam a estimação dos parâmetros da regressão PLS e, por consequência, reduzem a precisão dos índices de importância gerados.

Um aumento no número de observações eleva o desempenho das classificações. Tal situação é justificada pela maior disponibilização de informação à regressão PLS, a qual gera parâmetros mais precisos e conduz a índices de importância mais confiáveis, beneficiando o processo de eliminação de variáveis. Uma maior proporção de observações frente às variáveis também reduz a variabilidade nos critérios de desempenho e percentual de variáveis retidas.

6. Conclusão

Este artigo apresentou um método para seleção de variáveis com base em múltiplos critérios de desempenho. As etapas do método são: (1) Aplicação da regressão PLS no banco de dados e geração de um índice de importância da variável de processo; (2) Categorização das bateladas em duas classes e eliminação sistemática das variáveis irrelevantes até que exista apenas uma variável remanescente; (3) Geração do perfil de desempenho e aplicação da análise de Pareto Ótimo; e (4) Identificação da melhor solução da fronteira do Pareto através de uma distância euclidiana ponderada.

Quando aplicado na porção de teste de seis bancos de dados industriais, o método reteve, em média, 12% das variáveis originais. As variáveis selecionadas elevaram a sensibilidade de classificação da porção de teste em 9%, de 0,78 para 0,85, enquanto que a especificidade da mesma porção aumentou 20%, de 0,64 para 0,77. Verificou-se ainda redução significativa no percentual de custo e número de variáveis retidas ao incluir-se um quarto critério (custo de medição/coleta das variáveis retidas). Experimentos de simulação permitiram avaliar o impacto gerado por diferentes níveis de ruído, colinearidade e proporção de observações/variáveis sobre o método proposto.

Desdobramentos futuros incluem a extensão do método proposto para cenários onde diversas variáveis de resposta são encontradas. O desafio está na elevada correlação entre tais variáveis, responsável pela redução da eficiência dos métodos de classificação. Outro potencial desenvolvimento está ligado à categorização de bateladas em diversas classes (três ou mais), o que demanda aprimoramento na ferramenta de classificação KNN.

Referências

ABDI, H. Partial Least Squares (PLS) Regression. In: LEWIS-BECK, M.; BRYMAN, A.; FUTING, T. (Eds.). *Encyclopedia of Social Sciences Research Methods*. Thousand Oaks: Sage, 2003.

ANZANELLO, M.; ALBIN, S.; CHAOVALITWONGSE, W. Selecting the best variables for classifying production batches into two quality classes. *Chemometrics and Intelligent Laboratory Systems*, v. 97, n. 2, p. 111-117, 2009. http://dx.doi.org/10.1016/j.chemolab.2009.03.004

ARAGONÉS-BELTRÁN, P. et al. Valuation of urban industrial land: An analytic network process approach. *European Journal of Operational Research*, v. 185, p. 322-339, 2008. http://dx.doi.org/10.1016/j.ejor.2006.09.076

CHAOVALITWONGSE, W.; FAN, Y.; SACHDEO, C. On the time series k-nearest neighbor classification of abnormal brain activity. *IEEE Transactions on System and Man Cybernetics A*, v. 37, p. 1005-1016, 2007. http://dx.doi.org/10.1109/TSMCA.2007.897589

CHONG, I.; ALBIN, S.; JUN, C. A data mining approach to process optimization without an explicit quality function. *IIE Transactions*, v. 39, p. 795-804, 2007. http://dx.doi.org/10.1080/07408170601142668

DENHAM, M. Choosing the number of factors in partial least square regression: estimating and minimizing the mean squared error of precision. *Journal of Chemometrics*, v. 14, p. 351-361, 2000. http://dx.doi.org/10.1002/1099-128X(200007/08)14:4%3C351::AID-CEM598%3E3.0.CO;2-Q

DOAN, S.; HORIGUCHI, S. An efficient feature selection using multi-criteria in text categorization. In: *Fourth International Conference on Hybrid Intelligent Systems*, p. 86-91, 2004. http://dx.doi.org/10.1109/ICHIS.2004.20

GAUCHI, J.; CHAGNON, P. Comparison of selection methods of exploratory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, v. 58, p. 171-193, 2001. http://dx.doi.org/10.1016/S0169-7439(01)00158-7

GELADI, P.; KOWALSKI, B. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, v. 185, p. 1-17, 1986. http://dx.doi.org/10.1016/0003-2670(86)80028-9

GOLUB, T. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, v. 286, p. 531-537, 1999. http://dx.doi.org/10.1126/science.286.5439.531

GOUTIS, C. A fast method to compute orthogonal loadings partial least squares. *Journal of Chemometrics*, v. 11, p. 13-32, 1997. http://dx.doi.org/10.1002/(SICI)1099-128X(199701)11:1%3C33::AID-CEM432%3E3.0.CO;2-C

HORN, J.; NAFPLIOTIS, N.; GOLDBERG, D. A niched pareto genetic algorithm for multiobjective optimization. *IEEE World Congress on Computational Intelligence*, v. 1, p. 82-87, 1994. Proceedings of the First IEEE

Conference on Evolutionary Computation. http://dx.doi.org/10.1109/ICEC.1994.350037

HOSKULDSSON, A. PLS regression methods. *Journal of Chemometrics*, v. 2, p. 211-228, 1988. http://dx.doi.org/10.1002/cem.1180020306

HOSKULDSSON, A. Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, v. 55, p. 23-38, 2001. http://dx.doi.org/10.1016/S0169-7439(00)00113-1

HUANG, J.; TZENG, G.; ONG, C. Optimal fuzzy multi-criteria expansion of competence sets using multi-objectives evolutionary algorithms. *Expert Systems with Applications*, v. 30, p. 739-745, 2006. http://dx.doi.org/10.1016/j.eswa.2005.07.033

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS in very large datasets. *Computational Statistics & Data Analysis*, v. 48, p. 69-85, 2005. http://dx.doi.org/10.1016/j.csda.2003.11.027

MANNE, R. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, v. 2, p. 187-197, 1987. http://dx.doi.org/10.1016/0169-7439(87)80096-5

MEIRI, R.; ZAHAVI, J. Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, v. 171, p. 842-858, 2006. http://dx.doi.org/10.1016/j.ejor.2004.09.010

NELSON, P.; MacGREGOR, J.; TAYLOR, P. The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemometrics and Intelligent Laboratory Systems*, v. 80, p. 1-12, 2006. http://dx.doi.org/10.1016/j.chemolab.2005.04.006

OLAFSSON, S.; LI, X.; WU, S. Operations research and data mining. *European Journal of Operational Research*, v. 187, p. 1429-1448, 2008. http://dx.doi.org/10.1016/j.ejor.2006.09.023

OZTURK, A.; KAYALIGIL, S.; OZDEMIREL, N. Manufacturing lead time estimation using data mining. *European Journal of Operational Research*, v. 173, p. 683-700, 2006. http://dx.doi.org/10.1016/j.ejor.2005.03.015

PENDARAKI, K.; ZOPOUNIDIS, C.; DOUMPOS, M. On the construction of mutual fund portfolios: A multicriteria methodology and an application to the Greek market of equity mutual funds. *European Journal of Operational Research*, v. 163, p. 462-481, 2005. http://dx.doi.org/10.1016/j.ejor.2003.10.022

PIRAMUTHU, S. Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research*, v. 156, p. 483-494, 2004. http://dx.doi.org/10.1016/S0377-2217(02)00911-6

RIDGEWAY, G. Strategies and Methods for Prediction. In: YE, N. (Ed.). *The handbook of data mining.* Lawrence: New Jersey, 2003.

ROSE-PEHRSSON, S. et al. Multi-criteria fire detection systems using a probabilistic neural network. *Sensors and Actuators B*: Chemical, v. 69, p. 325-335, 2000. http://dx.doi.org/10.1016/S0925-4005(00)00481-0

SUEYOSHI, T. DEA-Discriminant Analysis: Methodological comparison among eight discriminant analysis approaches. *European Journal of Operational Research*, v. 169, p. 247-272, 2006. http://dx.doi.org/10.1016/j.ejor.2004.05.025

TABOADA, H.; COIT, D. Data clustering of solutions for multiple objective system reliability optimization problems. *Quality Technology & Quantitative Management Journal*, v. 4, p. 35-54, 2007.

TABOADA, H.; COIT, D. Multi-objective scheduling problems: Determination of pruned Pareto sets. *IIE Transactions*, v. 40, p. 552-564, 2008. http://dx.doi.org/10.1080/07408170701781951

WEISS, S. et al. Maximizing text-mining performance. *IEEE Intelligent Systems*, v. 14, p. 63-69, 1999. http://dx.doi.org/10.1109/5254.784086

WESTERHUIS, J.; KOURTI, T.; MacGREGOR, J. Analysis of multiblock and hierarquical PCA and PLS models. *Journal of Chemometrics*, v. 12, p. 301-321, 1998. http://dx.doi.org/10.1002/(SICI)1099-128X(199809/10)12:5%3C301::AID-CEM515%3E3.0.CO;2-S

WOLD, S.; SJOSTROM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, v. 58, p. 109-130, 2001. http://dx.doi.org/10.1016/S0169-7439(01)00155-1

WOLD, W. et al. Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, v. 58, p. 131-150, 2001. http://dx.doi.org/10.1016/S0169-7439(01)00156-3

ZITZLER, E.; THIELE, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary* Computation, v. 3, p. 257-271, 1999. http://dx.doi.org/10.1109/4235.797969

ZOPOUNIDIS, C.; DOUMPOS, M. Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, v. 138, p. 229-246, 2002. http://dx.doi.org/10.1016/S0377-2217(01)00243-0

A multiple criteria-based method for variable selection in industrial applications

Abstract

Several correlated and noisy variable are collected from industrial processes. This paper proposes a method for selecting the most relevant process variables aimed at classifying production batches into classes based on multiple criteria (e.g., sensibility and specificity). Production batches are inserted into two classes. The method first applies the PLS regression (Partial Least Squares) on process data and derives a variable importance index. A classification/elimination procedure is then carried out, and a weighted Euclidian distance is generated to identify the recommended variable subset. When applied to the testing set of real industrial data, the proposed method retained average 12% of original variables. The recommended subsets yielded 9% higher sensibility, from 0.78 to 0.85, and 20% higher specificity, from 0.64 to 0.77. Simulation experiments are also performed.

Keywords

Variable selection. Multiple criteria. PLS regression.