Produto & Produção, vol. 10, n. 3, p. 19 - 27, out. 2009

Identificação das variáveis mais relevantes para categorização de bateladas de produção: reduzindo a variância do percentual de variáveis retidas

Michel J. Anzanello, PhD
Programa de Pós-Graduação em Engenharia de Produção – UFRGS
anzanello@producao.ufrgs.br

Susan L. Albin, PhD
Rutgers – The State University of New Jersey
salbin@rci.rutgers.edu

Wanpracha A. Chaovalitwongse, PhD
Rutgers – The State University of New Jersey
wchaoval@rci.rutgers.edu

O desenvolvimento de métodos de seleção de variáveis em processos produtivos tem encontrado suporte no elevado volume de informações coletadas para fins de monitoramento e controle do processo. Embora métodos para a seleção de variáveis com propósitos de predição venham sendo amplamente sugeridos na literatura, a seleção de variáveis com vistas à classificação de observações em processos industriais permanece pouco explorada. Este artigo sugere extensões no método em Anzanello *et al*. (2009) com vistas à redução da variância do percentual de variáveis retidas para a classificação de bateladas de produção em duas classes. As variáveis de processo são analisadas pela regressão por mínimos quadrados parciais (*Partial Least Squares* — PLS) e ordenadas em termos de importância. As observações (representando bateladas de produção) são classificadas através da ferramenta *k*-vizinhos mais próximos (KVP) à medida que as variáveis são eliminadas. O melhor subconjunto de variáveis é escolhido via análise de Pareto. O método sugerido reduziu o percentual e a variância das variáveis retidas, e conduziu a incrementos sensíveis de acurácia, tanto em dados simulados como em dados de processos industriais.

Palavras chave: seleção de variáveis, PLS, classificação de bateladas de produção

Methods for variable selection have been massively developed due to the increasing volume of process data collected by sensors. Although selecting variables for the prediction purpose has been widely discussed, few studies have focused on variable selection for classification in industrial applications. In this paper, we extend the method proposed in Anzanello *et al*. (2009) in order to reduce the percent of retained variables for the classification of production batches into two classes. The method applies Partial Least Squares (PLS) regression to characterize the process variables, which are then ranked according to importance. Observations representing production batches are classified by means of the *k*-Nearest Neighbor technique as variables are removed. The best subset of variables is identified via Pareto Optimal analysis. When applied to simulated and real datasets, the proposed method reduced the percent and variance of retained variables, and yielded slight increments on classification accuracy.

Keywords: variable selection, PLS, classification of production batches

1 Introdução

A crescente utilização de sensores em ambientes industriais, aliada ao aumento de recursos computacionais para o armazenamento de dados, tem conduzido a cenários complexos em termos da manipulação e análise das informações coletadas (KETTANEH *et al.*, 2005). A seleção das variáveis de processo mais relevantes passou a cons-

tituir-se em tópico de fundamental importância para o eficiente monitoramento e otimização dos parâmetros do processo produtivo.

A vasta maioria dos métodos de seleção de variáveis sugeridos na literatura visa a identificar variáveis de processo (variáveis independentes) que conduzam a predições acuradas das especificações do produto (variáveis dependentes). Muitas aplicações práticas de produção, no entanto, priorizam a categorização de uma batelada de produção em classes de acordo com determinada especificação (como qualidade ou nível de lucratividade, entre outros). Nesta natureza de aplicação, a disponibilidade de um conjunto reduzido de variáveis relevantes é fundamental para categorizações precisas.

Diversas abordagens têm sido propostas para identificar as variáveis mais importantes para classificação de observações (ORTEGA, 2000; NG; LIU, 2000; MALLET *et al.*, 1998; PIRAMUTHU, 2004; LIU; YU, 2005). Em termos de aplicações industriais, algumas dessas abordagens utilizam Análise de Componentes Principais (ACP) para identificar as variáveis de processo com maior variância, e então utilizam as variáveis selecionadas para classificação (GUO *et al.*, 2002). A ACP, contudo, não captura as relações entre as variáveis de processo (independentes) e de produto (dependentes), e importantes informações entre os dois blocos de variáveis são perdidas. Neste contexto, a regressão por quadrados parciais mínimos (*Partial Least Squares* – PLS) aparece como ferramenta alternativa de análise, visto que, diferentemente da ACP, captura as relações entre as variáveis independentes e dependentes.

A integração de PLS a ferramentas de classificação para seleção de variáveis com vistas à categorização de bateladas em duas classes foi proposta em Anzanello *et al.* (2009). Naquele método, as variáveis são ordenadas de acordo com sua importância, com base nos parâmetros da regressão PLS. Uma medida de *performance* de classificação, como acurácia, é calculada após cada eliminação de variável. Por fim, o subconjunto de variáveis responsável pela máxima acurácia é apontado como a melhor solução.

A seleção de variáveis baseada na máxima acurácia apresenta duas desvantagens: 1. um elevado número de variáveis pode ser retido (causando *overfitting*), visto que tal sistemática não penaliza soluções com essa característica; e 2. a variância do percentual de variáveis retidas pode ser elevada, uma vez que a máxima acurácia pode estar atrelada a um elevado número de variáveis em determinadas situações e a um número reduzido em outras. Simulações baseadas no método em Anzanello *et al.* (2009) sugerem que diversos picos de acurácia podem ocorrer durante a eliminação de variáveis, como ilustrado na Figura 1. Esses picos alternativos, capazes de conciliar reduzido percentual de variáveis retidas e níveis satisfatórios de acurácia, devem ser avaliados como soluções potenciais.

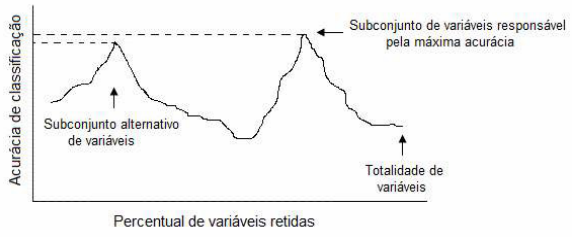

Figura 1 — Exemplo de perfil de acurácia com solução alternativa

Este artigo apresenta um método para reduzir a ocorrência de *overfitting* no processo de seleção de variáveis, com base em Anzanello *et al.* (2009). Dados de processo são inicialmente modelados por intermédio da regressão PLS. As variáveis independentes são ordenadas de acordo com sua relevância através do índice de importância sugerido por Wold *et al.* (2001). As observações, representando bateladas de produção, são então classificadas via *k*-vizinhos mais próximos (KVP) e a acurácia é calculada. A variável menos importante é removida e uma nova classificação é efetuada; esse processo é repetido até atingir-se um número pré-definido de variáveis remanescentes. A Análise de Pareto é então aplicada ao perfil de acurácia gerado, e as distâncias dos pontos da fronteira eficiente com relação a um ponto hipoteticamente definido como ideal são estimadas. O ponto de fronteira com a menor distância indica o melhor subconjunto de variáveis para classificação de bateladas de produção.

O método proposto apresentou significativa redução no percentual de variáveis retidas e na sua variância, ao mesmo tempo em que manteve a acurácia de classificação em patamares satisfatórios. Ao ser testado em dados simulados, o método reteve apenas 6% das variáveis frente a 16% retidas pelo método em Anzanello *et al.* (2009). A variância do percentual de variáveis retidas pelo método proposto foi de 3%, enquanto o método baseado na máxima acurácia apresentou 14% de variância. Ao ser aplicado em dados reais de processo, o método utilizou, em média, apenas 6% das variáveis originais e elevou a acurácia em 8%, de 78% para 84%.

O restante deste artigo é assim organizado. A Seção 2 apresenta os fundamentos da regressão PLS, KVP e Pareto Ótimo. A Seção 3 descreve a metodologia sugerida. A Seção 4 apresenta os detalhes da simulação, ao passo que os resultados numéricos são apresentados na Seção 5. A Seção 6 traz uma conclusão.

2 Referencial teórico

Uma breve revisão das ferramentas utilizadas no método proposto é apresentada nas seções seguintes.

2.1 Regressão PLS

A regressão PLS tem sido amplamente utilizada para selecionar as variáveis mais importantes com o propósito de predição (LINDGREN *et al.*, 1994; FORINA *et al.*, 1999; SARABIA *et al.*, 2001) e classificação (ANZA-NELLO *et al.*, 2009). Tal regressão apresenta satisfatória *performance* quando aplicada a variáveis altamente correlacionadas, além de poder ser utilizada em situações nas quais o número de variáveis é superior ao de observações (WOLD *et al.*, 2001; KETTANEH *et al.*, 2005; NELSON *et al.*, 2006; HOSKULDSSON, 2001). Os parâ-metros da regressão PLS podem ser estimados pelo algoritmo NIPALS (GOUTIS, 1997; DAYAL e MAC-GREGOR, 1997; ABDI, 2003).

Considere as matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$, compostas por *J* variáveis independentes e *M* variáveis dependentes, respectivamente, e *N* observações. A observação *i* pode ser representada pelo vetor x_i (x_{i1}, x_{i2},..., x_{iJ}) associado às variáveis independentes, e pelo vetor y_i (y_{i1}, y_{i2},..., y_{iM}) associado às variáveis dependentes. Caso exista somente uma variável dependente, a observação *i* é representada por y_i.

A regressão PLS gera combinações lineares das variáveis independentes (também referidas como componentes t_a) conforme a equação (1), onde $a = 1,\ldots,A$ e $A \leq J$. O número de componentes *A* é tipicamente pequeno quando comparado ao número de variáveis de processo. O número apropriado de componentes pode ser definido através de validação cruzada (HOSKULDSSON, 1998) ou pelo algoritmo inferencial sugerido por Lazraq e Cleroux (2001).

$$t_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + \ldots + w_{Ja}x_{iJ} = \boldsymbol{w_a}'\boldsymbol{x_i} \quad (1)$$

O vetor $\boldsymbol{w_a} = (w_{1a}, w_{2a},\ldots, w_{Ja})'$ representa os pesos, enquanto que o elemento w_{ja} denota o peso da variável independente *j* no componente *a*. Os pesos fornecem informações sobre a forma com que as variáveis independentes e dependentes interagem para gerar as matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ (WOLD *et al.*, 2001). De maneira semelhante, componentes podem ser gerados para as variáveis dependentes, conforme a equação (2).

$$u_{ia} = c_{1a}y_{i1} + c_{2a}y_{i2} + \ldots + c_{Ma}y_{iM} = \boldsymbol{c_a}'\boldsymbol{y_i} \quad (2)$$

onde $\boldsymbol{c_a} = (c_{1a}, c_{2a},\ldots, c_{Ma})'$ são os pesos das variáveis dependentes.

Os pesos $\boldsymbol{w_a}$ e $\boldsymbol{c_a}$ são calculados de forma a maximizar a covariância entre os componentes $\boldsymbol{t_a}$ e $\boldsymbol{u_a}$. Os componentes resultantes são independentes, ou seja, ortogonais entre si (WOLD *et al.*, 2001; XU; ALBIN, 2002).

Por sua vez, o vetor de cargas, $\boldsymbol{p_a} = (p_{1a}, p_{2a},\ldots, p_{Ja})'$ é gerado através da regressão das colunas de $\boldsymbol{X}$ sobre os componentes $\boldsymbol{t_a}$. De acordo com Wold *et al.* (2001), os parâmetros de carga fornecem importantes informações sobre as variáveis independentes quando multiplicados por seus componentes, $\boldsymbol{t_a}$.

Alternativamente, os pesos e cargas das variáveis independentes podem ser manipulados para gerar os pesos w^*_{ja} de acordo com $w^*_{ja} = w_{ja}(p_{ja}w_{ja})^{-1}$. Wold *et al.* (2001) sugerem que os pesos w^*_{ja} enfatizam a importância das variáveis independentes, gerando modelos mais estáveis e auxiliando na identificação das variáveis mais relevantes. Detalhes adicionais sobre w^*_{ja} podem ser obtidos em Manne (1987).

2.2 Classificação por k-vizinhos mais próximos (KVP)

A ferramenta de classificação *k*-Vizinhos mais Próximos (KVP) tem sido amplamente utilizada em aplicações práticas por conta de sua fundamentação teórica simplificada e larga disponibilidade em aplicativos computacionais.

Considere *N* observações em uma porção de treino com dimensões definidas pelas *J* variáveis independentes. Objetiva-se classificar uma nova observação como 0 ou 1 (não conforme ou conforme, respectivamente), utilizando somente variáveis independentes. O algoritmo KVP calcula a distância Euclidiana entre a nova observação e as *k* observações mais próximas. A classe das *k* observações mais próximas é conhecida, 0 ou 1. A nova observação é então classificada como 0 se a maioria das *k* observações mais próximas pertencem à classe 0. Formas alternativas de classificação utilizando KVP podem ser encontradas em Chaovalitwongse *et al.* (2007). O valor de *k* pode ser obtido através de validação cruzada na porção de treino, maximizando-se indicadores de *performance* como acurácia, sensitividade ou especificidade, entre outros. Maiores detalhes sobre KVP podem ser obtidos em Ridgeway (2003), ao passo que exemplos de aplicações são encontrados em Golub *et al.* (1999), Weiss *et al.* (1999) e Chaovalitwongse *et al.* (2007).

2.3 Pareto ótimo

A análise conhecida como Pareto Ótimo (PO) identifica um subconjunto de soluções distintas em aplicações caracterizadas por múltiplas funções objetivo. Tais aplicações geralmente não apresentam uma única solução, mas uma série de possíveis soluções. Procedimentos de seleção de variáveis nos quais medidas de *performance* de classificação devem ser maximizadas e percentual de

variáveis retidas minimizado são exemplos de aplicações com diversas soluções possíveis. Por conta de sua vasta aplicabilidade, a análise de PO tem sido integrada a algoritmos focados na otimização de cenário complexos, conforme Deb *et al.* (2002a; b).

As soluções apontadas pelo Pareto são definidas como soluções "não-dominadas", ou seja, soluções que não podem ser superadas por soluções vizinhas em termos dos objetivos avaliados (AZAPAGIC, 1999). As soluções "não-dominadas" são frequentemente ilustradas em um contorno gráfico denominado Fronteira do Pareto (ou Fronteira Eficiente). A Fronteira do Pareto facilita a identificação da melhor solução (ou grupo de melhores soluções), visto que o conjunto de potenciais soluções é reduzido de forma significativa (HORN *et al.*, 1994; ZITZLER e THIELE, 1999; TABOADA; COIT, 2007, 2008).

3 Método

O método para seleção de variáveis com vistas à classificação de bateladas de produção é composto por seis passos, descritos na sequência.

Passo 1: Aplique a regressão PLS na porção de treino de dados históricos de processo.

Considere as matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ descritas na Seção 2.1. Cada observação representa uma batelada de produção descrita por J variáveis independentes e uma única variável dependente. Classifique cada observação como 0 ou 1 (não conforme e conforme, respectivamente), utilizando um ponto de corte na variável dependente. As observações são então randomicamente divididas em duas porções, sendo que a porção de treino é utilizada para identificar as variáveis mais importantes e a porção de teste representa novas observações. Recomenda-se uma proporção de 60% para porção de treino e 40% para porção de teste (CHONG *et al.*, 2007).

Na sequência, aplique a regressão PLS na porção de treino, visando a caracterizar a relação entre as variáveis independentes e a variável dependente. Recomenda-se normalizar os dados antes da aplicação de PLS para eliminar efeitos de escala. Os parâmetros de interesse gerados pela regressão PLS incluem os pesos w_{ja}, cargas p_{ja} e percentual de variação em $\boldsymbol{Y}$ explicado pelo componente a, R^2_{Ya}.

Passo 2: Gere um índice de importância para as variáveis independentes (s).

O índice de importância visa a remover variáveis independentes irrelevantes que apresentam demasiado ruído. O índice de importância da variável j é expresso por s_j, j=1,…,J, sendo que o conjunto de índices para as J variá-veis é representado pelo vetor $\boldsymbol{s} = (s_1, s_2, \ldots, s_J)'$. Valores elevados de s_j denotam variáveis importantes.

O índice utilizado neste artigo foi inicialmente proposto por Wold *et al.* (2001) para a identificação das variáveis mais relevantes em modelos de predição. Tal índice é baseado nos pesos modificados w^*_{ja} e na fração de variância em $\boldsymbol{Y}$ explicada pelo componente a = 1,…,A, conforme apresentado na equação (3).

$$s_j = \sum_{a=1}^{A} (w^*_{ja})^2 R^2_{Ya} \qquad j=1,\ldots,J. \qquad (3)$$

Passo 3: Classifique a porção de treino utilizando KVP e elimine variáveis irrelevantes.

As observações da porção de treino, consistindo de J variáveis independentes, são classificadas como 0 ou 1 e a acurácia de classificação é calculada. Acurácia é aqui definida como a razão entre o número de classificações corretas e o número total de classificações efetuadas. O parâmetro k é obtido através de validação cruzada na porção de treino.

Na sequência, remova a variável com o menor s_j e classifique novamente a porção de treino consistindo de J-1 variáveis. Calcule novamente a acurácia de classificação. Esse processo de eliminação e classificação é repetido até atingir-se um número mínimo de variáveis remanescentes, sendo que duas variáveis é o número recomendado.

Passo 4: Gere um gráfico relacionando acurácia e percentual de variáveis retidas e aplique a análise de Pareto no perfil obtido.

Construa um gráfico associando acurácia ao percentual de variáveis retidas, conforme ilustrado na Figura 1. A análise de pareto é aplicada ao perfil de acurácia visando a identificar soluções que maximizem a acurácia e minimizem o percentual de variáveis retidas.

Passo 5: Calcule a distância dos pontos da fronteira do pareto ao ponto representando a solução ideal.

Neste passo, cada ponto localizado na fronteira do Pareto tem sua distância Euclidiana calculada em relação ao ponto definido como ideal. As coordenadas do ponto ideal são definidas pelo usuário e devem apresentar as seguintes propriedades: (*i*) devem ser contidas no intervalo [0,1], e (*ii*) devem ser coerentes com os critérios analisados (ou seja, valores próximos a 1 para a acurácia e valores próximos a 0 para o percentual de variáveis retidas). Considere, por exemplo, as coordenadas do ponto ideal como (PVR,1), onde PVR denota o Percentual de Variáveis Retidas quando o processo de eliminação de variáveis é concluído, e 1 representa a máxima acurácia. O ponto da fronteira com a menor distância ao ponto ideal identifica o melhor subconjunto de variáveis a ser

retido, conforme ilustrado na Figura 2. Esse método é referido como Distância de Pareto (DP).

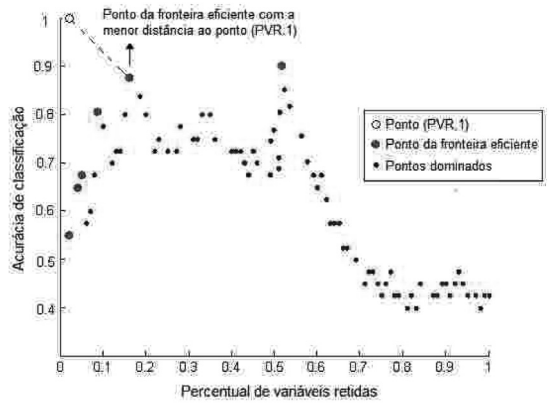

Figura 2 — Distância dos pontos de fronteira ao ponto ideal

Passo 6: Classifique as observações da porção de teste utilizando as variáveis selecionadas.

Aplique KVP na porção de teste contendo apenas o subconjunto de variáveis selecionadas e calcule a acurácia de classificação.

Para fins de análise de eficiência do método DP, o mesmo é confrontado com o método em Anzanello *et al.* (2009) por intermédio de simulação. Naquele estudo, as variáveis são selecionadas através dos passos 1, 2 e 3 anteriormente descritos, sendo escolhido o subconjunto de variáveis conduzindo ao máximo valor de acurácia. Esse método será referido como Máxima Acurácia (MA).

4 Detalhes da simulação

Os dados gerados na simulação são baseados em dados reais de um processo de produção de látex. Assume-se que um modelo PLS, conforme a equação (4), seja satisfatório para descrever tal processo. O banco de dados original é constituído por 100 variáveis independentes, x_{ij}, uma variável dependente, y_i, e 200 observações (sendo que cada observação representa uma batelada de produção).

$$y_i = \sum_{j=1}^{J} b_j x_{ij} + \varepsilon_i \qquad j=1,\ldots,J \qquad (4)$$

onde b_j é o coeficiente da regressão PLS e $\varepsilon_i \sim N(0, \sigma^2)$.

A matriz $\boldsymbol{X}$ (consistindo das variáveis independentes x_{ij}) é gerada de acordo com uma distribuição multinormal com média m_j para as J variáveis independentes e matriz de correlação Γ. Os parâmetros m_j e Γ são estimados com base no banco de dados de produção de látex.

A regressão PLS é então aplicada aos dados originais para estimar b_j, sendo que três componentes são retidos no modelo. A variância do termo de erro σ^2 é estimada através da soma do resíduo de predição SRP_A [conforme equação (5)] do modelo PLS gerado, onde A denota o número de componentes retidos e n é o número total de observações (DENHAM, 2000).

$$\hat{\sigma}^2 = \frac{SRP_A}{n - A - 1} \qquad (5)$$

A simulação utiliza 3 fatores entendidos como relevantes para identificação de variáveis em processos industriais: (*i*) variância do erro, (*ii*) correlação entre as variáveis, e (*iii*) razão entre o número de observações e o número de variáveis. Os níveis nominais para variância do erro (σ^2) e correlação (Γ) são extraídos do banco de dados do processo de látex, e então manipulados conforme apresentado na Tabela 1 para gerar os níveis alto e baixo de cada fator. O terceiro fator apresenta dois níveis: razão 0,5, onde o número de observações é menor que o número de variáveis (característica de processos em batelada), e razão 5. Foram gerados 18 (=3 × 3 × 2) casos distintos, com 200 repetições por caso.

Tabela 1 — Fatores e níveis da simulação

FATORES	NÍVEIS
Variância do erro	$0{,}5\sigma^2$; σ^2; $2\sigma^2$
Correlação entre as variáveis	Γ^3; Γ; Γ 1/3
Razão entre o número de observações e o número de variáveis	0.5; 5

As observações geradas são então classificadas como conforme ou não conforme de acordo com o ponto de corte da variável dependente.

5 Resultados

Os resultados da aplicação do método em dados simulados e reais são apresentados na sequência.

5.1 Resultados da simulação

Os métodos DP (sugerido neste estudo) e MA (ANZANELLO *et al.*, 2009) são comparados utilizando os dados simulados na Seção 4. Tendo-se em vista que a acurácia e o percentual de variáveis retidas apresentaram tendências similares frente aos diferentes fatores e níveis, optou-se por apresentar somente os resultados médios, calculados sobre todas as combinações de fatores e níveis.

A Tabela 2 traz os resultados da simulação. O método DP retém apenas 6% das variáveis independentes, ao passo que o método MA requer 16% de variáveis para classificação.

Tabela 2 — Desempenho dos métodos na porção de treino dos dados simulados (resultados médios sobre todos os fatores e níveis)

	Máxima acurácia (MA)	Distância do Pareto (DP)
Acurácia de Classificação (%)	90	89
Desvio padrão de Acurácia de Classificação (%)	3	3
Variáveis Retidas (%)	16	6
Desvio Padrão das Variáveis Retidas (%)	14	3

A acurácia gerada pelos métodos é semelhante. Em termos de estabilidade, o método DP apresenta menor desvio padrão no percentual de variáveis retidas quando comparado ao método MA: 3% contra 14%.

As Figuras 3 e 4 trazem as distribuições do percentual de variáveis retidas para os níveis nominais da simulação (variância do erro σ^2 correlação Γ e razão 2) para 200 repetições. O método sugerido apresenta significativa redução na variância das variáveis retidas. Por conta da sua maior consistência e reduzido percentual de variáveis retidas, o método DP será utilizado em dados reais de processo na Seção 5.2.

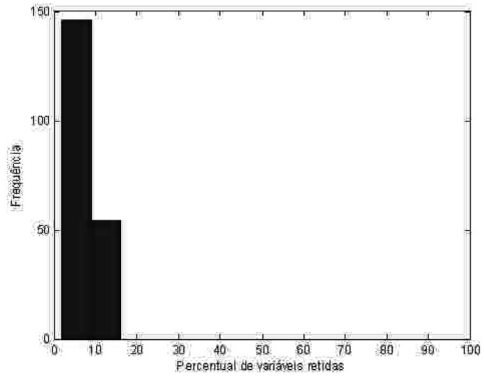

Figura 3 – Distribuição das variáveis retidas pelo método DP

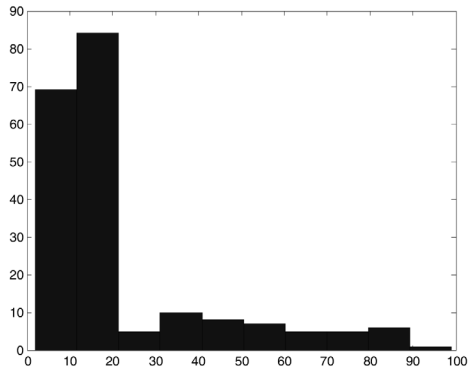

Figura 4 – Distribuição das variáveis retidas pelo método MA

5.2 Aplicação do método DP em dados reais de processo

O método DP é aplicado em 4 bancos de dados de processos químicos distintos. A Tabela 3 traz o número de variáveis independentes e observações para cada banco. Tais variáveis descrevem temperaturas, pressões e concentrações, entre outros. Existe apenas uma variável dependente por banco, a qual mede uma propriedade específica do produto. O banco PROC1 refere-se a um processo de produção de nylon; PROC2 refere-se à polimerização em processo de produção de látex; PROC3 descreve um processo de obtenção de antibióticos; e PROC4 refere-se à reciclagem de papel.

Tabela 3 – Banco de dados para 4 processos industriais

Banco de Dados	Número de variáveis independentes	Número de Observações	
		Porção de Treino	Porção de Teste
PROC1	100	57	14
PROC2	117	210	52
PROC3	54	192	192
PROC4	96	115	29

As observações de cada banco de dados foram classificadas em duas categorias, conforme e não-conforme, de acordo com pontos de corte definidos por especialistas de processo. Na sequência, a regressão PLS foi aplicada na porção de treino de cada banco de dados. Três componentes foram retidos em cada banco através de 10 repetições de validação cruzada. Os componentes retidos explicam 94% da variância em $\boldsymbol{Y}$ para o PROC1, 77% para o PROC2, 68% para o PROC3 e 71% para o PROC4. O parâmetro k para a ferramenta de classificação KVP foi

Tabela 4 – Desempenho do método DP na porção de teste dos dados reais

Banco de dados (número de variáveis originais)	Acurácia de classificação utilizando todas as variáveis (%)	Acurácia de classificação utilizando método DP (%)	Variáveis retidas utilizando método PD (%)
PROC1 (100)	79	86	5
PROC2 (117)	79	87	7
PROC3 (54)	75	79	6
PROC4 (96)	78	83	4
Média	78	84	6

definido com base em 10 repetições de validação cruzada na porção de treino. Obteve-se k=9 para o PROC1, k=3 para o PROC2, k=3 para o PROC3 e k=9 para o PROC4.

A Tabela 4 apresenta a acurácia de classificação e percentual de variáveis retidas na porção de teste nos dados de processo. Em média, o método DP utilizou 6% das variáveis originais e aumentou a acurácia em 8%, de 78% para 84%.

6 Conclusões

Este artigo apresentou um método para seleção de variáveis com vistas à redução da variância do percentual de variáveis retidas para classificação de bateladas. Para tanto, a regressão PLS foi integrada às ferramentas *k* vizinhos mais próximos e análise por Pareto Ótimo. No método sugerido, as variáveis de processo são sistematicamente eliminadas, de acordo com sua importância, definida com base nos parâmetros da regressão PLS. Após cada eliminação, efetua-se uma nova classificação utilizando as variáveis remanescentes e recalcula-se a acurácia. Um conjunto reduzido de possíveis soluções é então apontado pelo Pareto, ao passo que a melhor solução é definida com base na distância a um ponto considerado ideal.

Ao ser testado em dados simulados, o método reteve apenas 6% das variáveis, enquanto que o método em Anzanello *et al.* (2009) reteve 16%. A variância do percentual de variáveis retidas pelo método proposto foi de 3% frente à 14% gerada pelo método MA. Ao ser aplicado em quatro bancos de dados reais, o método utilizou, em média, apenas 6% das variáveis originais e elevou a acurácia em 8%.

Desdobramentos futuros incluem a extensão do método sugerido a cenários com três ou mais classes de qualidade. Avaliam-se ainda adequações do método a processos descritos por mais de uma variável de produto.

Referências

ABDI, H. Partial Least Squares (PLS) Regression. In *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 levels. *In: Chemometrics and Intelligent Laboratory Systems,* v. 97, p. 111-117, 2009.

AZAPGIC, A. Life cycle assessment and its application to process selection, design and optimization. *Chemical Engineering Journal*, v. 73, n. 1, p.1-21, 1999.

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, n. 6, p. 1005-1016, 2007.

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.

DAYAL, B.; MACGREGOR, J. Improved PLS algorithms. *Journal of Chemometrics***,** v.11, p. 73-85, 1997.

DEB, K.; PRATAP, A.; AGARWAL, S.; MEYARIVAN, T. A fast and elitist multiobjective genetic algoritm: NSGA-II. *IEEE Transactions on Evolutionary Computation,* v. 6, n. 2, 2002a.

DEB. K.; THIELE, L.; LAUMANNS, M.; ZITZLER, E. Scalable multi-objective optimization test problems. Proceedings of the 2002 *Congress on Evolutionary Computation,* v. 1, p. 825-830, 2002b.

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.

FORINA, M.; CASOLINO, C., MILLAN, C. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *Chemometrics and Intelligent Laboratory Systems,* v. 13, p. 165-184, 1999.

GOLUB, T.; SLONIM, D.; TAMAYO, P.; HUARD, C.; GAASENBEEK, M.; MESIROV, J.; COLLER, H.; LOH, M.; DOWNING, J.; CALIGIURI, M.; BLOOMELD, C.; LANDER, E. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science,* v. 286, n. 5439, p. 531-537, 1999.

GOUTIS, C. A fast method to compute orthogonal loadings partial least squares. *Journal of Chemometrics,* v. 11, p. 13-32, 1997.

GUO, Q.; WU, W.; MASSART, D.; BOUCON, C.; JONG, S. Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems,* v. 61, p. 123-132, 2002.

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

HOSKULDSSON, A. PLS regression methods. *Journal of Chemometrics.* v. 2, p. 211-228, 1988.

HOSKULDSSON, A. Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems,* v. 55, p. 23-38, 2001.

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS in very large datasets. *Computational Statistics & Data Analysis,* v. 48, p. 69-85, 2005.

LAZRAQ, A.; CLEROUX, R. The PLS multivariate regression model: tehsting the significance of successive PLS components. *Journal of Chemometrics,* v. 15, p. 523-536, 2001.

LINDGREN, F.; GELADI, P.; RANNAR, S.; WOLD, S. Interactive variable selection (IVS) for PLS: Part 1. Theory and algorithms. *Journal of Chemometrics,* v.8, p. 349-363, 1994.

LIU, H; YU, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering,* v. 17, n. 4, p. 491-502, 2005.

MALLET, Y.; de VEL, O.; COOMANS, D. Integrated feature extraction using adaptive wavelets. *In:* LIU, H; MOTODA, H. *Feature extraction, construction and selection:* A Data mining perspective, p. 175-189, 1998.

MANNE, R. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems,* v. 2, p. 187-197, 1987.

NELSON, P.; MACGREGOR, J; TAYLOR, P. *The impact of missing measurements on PCA and PLS prediction and monitoring applications,* v. 80, p. 1-12, 2006.

NG, K.; LIU, H. Customer retention via data mining. Artificial Intelligence Review – *An International Science and Engineering Journal,* v. 14, p. 569-590, 2000.

ORTEGA, J. Issues on the application of data mining. *Artificial Intelligence Review* – An International Science and Engineering Journal, v. 14, 2000.

PIRAMUTHU, S. Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research,* v.156, n.2, p. 483-494, 2004.

RIDGEWAY, G.(Editor), *The handbook of data mining.* Lawrence: New Jersey, 2003.

SARABIA, L.; ORTIZ, M.; SANCHEZ, A. Dimension wise selection in partial least squares regression a bootstrap estimated signal-noise relation to weight the loadings, *PLS and Related Methods*, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA, Editeur, Paris, p. 327-339, 2001.

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.

_____. Multi-objective scheduling problems: Determination of pruned Pareto sets. *IIE Transactions,* v. 40, p. 552-564, 2008.

WEISS, S.; APTE, C.; DAMERAY, D.; JOHNSON, D.; PLES, F.; GOETZ, T.; HAMPP, T. Maximizing text-mining performance. *IEEE Intell.* Syst. 14(4), p. 63-69, 1999.

WOLD, S.; SJOSTROM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* v. 58, p. 109-130, 2001.

ZITZLER, E.; THIELE, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach, *IEEE Transactions on Evolutionary Computation,* v. 3, n. 4, p. 257-271, 1999.