Seleção de variáveis com vistas à classificação de bateladas de produção em duas classes

Michel José Anzanello

Resumo

Bancos de dados caracterizados por elevado número de variáveis correlacionadas são usualmente encontrados em ambientes industriais, dificultando a identificação das variáveis de processo mais relevantes. A regressão por quadrados parciais mínimos (Partial Least Square – PLS) tem sido amplamente utilizada para a seleção de variáveis com propósitos de predição. No entanto, muitas aplicações práticas priorizam a correta categorização de lotes produtivos em classes, de acordo com determinada especificação do produto final. Neste artigo, a regressão PLS é integrada às ferramentas de classificação z vizinhos mais próximos (z-Nearest Neighbor) e máquina de suporte vetorial (Support Vector Machine) com visando a seleção de variáveis para fins de categorização de bateladas de produção em duas classes. Índices de Importância das Variáveis (IIV) baseados nos parâmetros da regressão PLS são desenvolvidos para o ordenamento das variáveis de processo, de acordo com sua relevância para a caracterização da variável de produto, e então integrados às ferramentas de classificação. O subconjunto de variáveis retidas é identificado através do monitoramento do perfil de acurácia gerado com a remoção sistemática das variáveis menos relevantes. Aplicada em três bancos de dados, a metodologia proposta reduziu o número de variáveis de processo necessárias para classificação de bateladas em 90,6% e elevou a acurácia média de classificação em 29,2%, quando comparada à aplicação de ferramentas de classificação na totalidade das variáveis.

Palavras-chave: *Seleção de variáveis. PLS. z vizinhos mais próximos. Máquina de suporte vetorial.*

1 *Introdução*

O problema de seleção de variáveis tem sido estudado em diversas áreas por diferentes motivos. Áreas da matemática e da estatística têm se focado na identificação de um subconjunto de variáveis *x* que conduzam à melhor predição da variável de resposta *y*, dentro do contexto de regressão linear. Regressões *Stepwise*, *Backward* e *Forward* têm sido aplicadas para esse propósito, além de técnicas bayesianas (PHILIPHS; GUTTMAN, 1998; GEORGE, 2000; MEROLA; ABRAHAM, 2001).

Em ambientes industriais, o controle do processo produtivo envolve um número elevado de variáveis, como temperaturas, pressões e concentrações de reagentes, entre outras. A crescente utilização de sensores, aliada ao aumento de recursos computacionais para o armazenamento de dados, tem conduzido a cenários complexos em termos da manipulação e análise desses dados (KETTANEH et al., 2005). De tal forma, a identificação das variáveis de processo mais relevantes constitui tópico de fundamental importância para o monitoramento dos parâmetros do processo produtivo, além de oferecer condições para a correta caracterização de produtos de acordo com as especificações desejadas.

A seleção de variáveis em processos industriais pode ser justificada pelos seguintes aspectos: i) um modelo composto por elevado número de variáveis pode apresentar aderência satisfatória aos dados modelados, porém não oferece garantias em termos de predição (por conta do demasiado número de variáveis independentes no modelo – *overfitting*) e classificação (devido ao ruído inserido por variáveis menos relevantes) (GUYSON; ELISSEEFF, 2003); ii) a identificação de variáveis com base no conhecimento empírico de especialistas é frequentemente sujeita a equívocos; e iii) a preferência por modelos reduzidos, por demandarem menor tempo de análise e serem menos complexos.

Diversas combinações de ferramentas estatísticas têm sido utilizadas para a identificação de variáveis de predição em processos industriais. Muitas aplicações práticas de produção,

no entanto, priorizam a categorização de um lote produtivo em classes de acordo com determinada especificação (como qualidade final do produto, confiabilidade e nível de lucratividade, entre outros), em detrimento da predição da variável de produto *y*. Nesta natureza de aplicação, a disponibilidade de um conjunto reduzido de variáveis relevantes é fundamental para categorizações precisas. O estudo aqui apresentado tem como objetivo combinar ferramentas de análise multivariada e de classificação para a seleção de variáveis que caracterizem bateladas de produção em duas classes. A correta classificação de bateladas em classes, especialmente durante os primeiros estágios de produção, permite o ajuste dos parâmetros do processo com vistas à correção de inconsistências ou, em situações extremas, sinaliza a necessidade de remanejo da batelada para destino alternativo.

Este artigo traz como contribuições originais: i) a integração de uma técnica multivariada de regressão (regressão por mínimos quadrados parciais – PLS) a ferramentas de *data mining* voltadas à classificação de observações (*z* vizinhos mais próximos – ZVP e máquinas de suporte vetorial – MSV); e ii) desenvolvimento de índices de importância para a identificação das variáveis de processo mais relevantes e a posterior comparação do desempenho desses índices quando combinados com as ferramentas de classificação.

O restante deste artigo é organizado como segue: a seção 2 apresenta os fundamentos de PLS, ZVP e MSV, enquanto a seção 3 descreve a metodologia sugerida. Exemplos númericos são apresentados na seção 4 e uma conclusão encerra o estudo na seção 5.

2 *Referencial teórico*

2.1 Regressão por quadrados parciais mínimos

A regressão por quadrados parciais mínimos (PLS) constitui-se em um método de análise multivariada que relaciona duas matrizes de variáveis, **X** (variáveis de processo) e **Y** (variáveis de produto), através de um modelo de regressão. O principal objetivo é maximizar a covariância das combinações lineares das variáveis (combinações essas definidas como componentes) (GARTHWAITE, 1994; WOLD et al., 2001a, b; MONTGOMERY, 2001). 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.

A regressão PLS gera L componentes (t_1, t_2, …, t_L), onde $t_1 = w_1 x_1 + w_{12} x_2 + ... + w_{1k} x_k = \mathbf{w}_1' \mathbf{x}$ é o primeiro componente da matrix **X** (dimensões $N \times K$), e $\mathbf{w}_1 = (w_{11}, w_{12}, ..., w_{1k})'$ representa o vetor dos pesos em relação ao componente $\mathbf{t}_1$. Os pesos fornecem informações importantes sobre a forma como as variáveis *x* se combinam para gerar as relações quantitativas entre as matrizes **X** e **Y**, sendo que de **w** possibilitam a identificação das variáveis responsáveis por informações relevantes (HOSKULDSSON, 1988).

Uma relação similar é construída para a matriz **Y** (dimensões $N \times M$), onde o primeiro componente é representado por $u_1 = c_1 y_1 + c_{12} y_2 + ... + c_{1k} y_k = \mathbf{c}_1' \mathbf{y}$, e $\mathbf{c}_1 = (c_{11}, c_{12}, ..., c_{1k})'$ representa o vetor de pesos do componente $\mathbf{u}_1$. Os vetores $\mathbf{w}_1$ e $\mathbf{c}_1$ são estimados através da maximização da covariância entre as combinações lineares de **X** e **Y**, restrito à condição de ortogonalidade de **w** (XU; ALBIN, 2002).

De acordo com Wold et al. (2001a), os componentes **t** e **u** podem ser manipulados de forma a gerar o coeficiente de regressão b_{mk} Equação (1), similar ao coeficiente da regressão linear múltipla. A magnitude de b_{mk} sinaliza a relevância da variável independente *k* para a predição da variável dependente *m; l* representa o número de componentes retidos para análise.

$$b_{mk} = \sum_{a=1}^{l} c_{ma} w_{ka} \qquad m = 1, .., M \text{ e } k = 1, .., K \qquad (1)$$

A regressão PLS apresenta vantagens quando comparada à tradicional regressão linear múltipla, visto que não é afetada por variáveis altamente correlacionadas, níveis de ruído elevados e observações faltantes. PLS também é recomendada em situações em que o número de variáveis é superior ao número de observações (WOLD et al., 2001a; ABDI, 2003).

A regressão PLS tem sido amplamente utilizada na seleção de variáveis em processos industriais com vistas à predição das características do produto final. Buscando eliminar variáveis em bancos de dados caracterizados por elevados níveis de ruído, Forina et al. (1999) propuseram a utilização iterativa de PLS para a identificação das variáveis independentes mais relevantes, ao passo que Sarabia et al. (2001) sugeriram um método gráfico baseado na ponderação do vetor **w** com o mesmo propósito. A utilização de PLS para seleção de variáveis em processos industriais também foi discutida por Wold et al. (2001a) em processos de reciclagem de papel e por Gauchi e Chagnon (2001) e Lazraq et al. (2003) em diversos processos químicos. Focados na eficiência de métodos de seleção de variáveis, Chong e Jun (2005) compararam diversos métodos em dados simulados com distintos níveis de multicolinearidade, proporções de variáveis relevantes e ruído das variáveis.

A regressão PLS também tem sido utilizada na seleção de variáveis em dados espectrais de análises químicas, os quais são caracterizados por milhares de variáveis independentes. Hoskuldsson (2001) propôs uma sistemática para a pré-seleção de variáveis e posterior aplicação de PLS nos dados resultantes, enquanto Xu et al. (2007) propuseram a inserção de vetores de ponderação para ajustar a contribuição de cada variável antes da aplicação de

PLS, visando aumentar a capacidade preditiva do modelo. Por fim, áreas de análise química têm aplicado PLS para avaliação da influência de estruturas moleculares sobre as propriedades de uma substância (WOLD et al., 2001a; ZHAI et al., 2006).

A aplicação de PLS com vistas à seleção de variáveis com propósitos de classificação é bastante restrita, limitando-se ao estudo de Anzanello et al. (2009). Neste estudo, PLS e ferramentas de classificação são combinadas, gerando-se significativa redução no percentual de variáveis retidas.

2.2 Ferramentas de classificação: z Vizinhos mais Próximos e Máquina de Suporte Vetorial

A primeira técnica de classificação, ZVP, insere uma nova observação à classe (categoria) com maior número de incidência entre as *z* observações (vizinhas) mais próximas. Considere *N* observações em uma porção de treino com dimensões definidas pelas *K* variáveis independentes. Objetiva-se classificar uma nova observação como 0 ou 1 (baixa qualidade ou elevada qualidade, respectivamente), utilizando somente variáveis de processo. O algoritmo ZVP calcula a distância euclidiana entre a nova observação e as *z* observações mais próximas. A classe das *z* observações mais próximas é conhecida, 0 ou 1. A nova observação é então classificada como 0 se a maioria das *z* observações mais próximas pertencer à classe 0. O parâmetro *z* pode ser obtido através de validação cruzada na porção de treino maximizando-se indicadores de performance como acurácia, entre outros. Maiores detalhes sobre ZVP podem ser obtidos em Duda et al. (2001), enquanto exemplos de aplicações são encontrados em Golub et al. (1999), Weiss et al. (1999) e Chaovalitwongse et al. (2007).

A segunda ferramenta de classificação, MSV, define um plano de separação entre dois grupos de observações, de forma que as observações da classe 0 sejam separados das observações da classe 1. O plano é construído utilizando dois subplanos paralelos auxiliares, cada um posicionado em um lado do plano principal. A idéia é maximizar a distância entre os dois subplanos, visto que seu afastamento denota maior poder de categorização da sistemática. O cálculo da distância entre os subplanos penaliza observações da classe 0 posicionadas no lado oposto do plano, ou seja, localizadas no "território" da classe 1. Com vistas ao aumento do poder de separação da MSV, transformações sobre os dados originais são efetivadas antes da construção do plano de separação. Essas transformações buscam "mover" as observações originais no espaço, de maneira a facilitar a construção do plano de separação. Essas funções são chamadas de kernel, sendo funções polinomiais e gaussiana exemplos de transformações usualmente utilizadas. Maiores detalhes da função kernel e de suas aplicações podem ser obtidos em Cristianini e Shawe-Taylor (2000), Duda et al. (2001) e Huang et al. (2006).

3 *Metodologia proposta*

A metodologia proposta visa identificar as variáveis de processo mais importantes para a classificação de bateladas de produção em duas classes de qualidade. A metodologia é composta por três passos: (1) aplicação de regressão PLS sobre dados de processo; (2) geração de Índices de Importância das Variáveis (IIV) com base nos parâmetros oriundos da regressão PLS; e (3) eliminação das variáveis de processo menos relevantes e categorização das observações através de uma ferramenta de classificação. Os passos da sistemática proposta são descritos a seguir.

3.1 Passo 1 – Aplicação de regressão PLS

Os dados de processo são divididos em dois blocos (porção de treino e porção de teste), mantendo uma proporção razoável de acordo com a natureza do processo e a disponibilidade de dados. A regressão PLS é então aplicada sobre a porção de treino. O número de componentes PLS a ser retido é definido através de validação cruzada com base no percentual de variância em $\mathbf{Y}$, R^2_{Ya}, explicado pelos componentes retidos (WOLD et al., 2001). Pode-se ainda utilizar o procedimento inferencial de Lazraq e Cleroux (2001) para a definição do número de componentes a ser retido.

A regressão PLS pode ser efetivada através do algoritmo NIPALS (GELADI; KOWALSKI, 1986), sendo as variáveis normalizadas no início do procedimento para eliminar efeitos de escala.

3.2 Passo 2 – Geração dos Índices de Importância das Variáveis

Os parâmetros da regressão PLS estimados no passo anterior são matematicamente manipulados para a geração de três Índices de Importância das Variáveis (IIV). Esses índices quantificam a influência das variáveis de processo de acordo com sua importância ao explicar a variância das variáveis de processo (x) e produto (y). Variáveis com IIV elevados são tidas como mais relevantes na explicação da variância das variáveis y e, por consequência, na discriminação de tais variáveis em classes (DUDA et al., 2001). Os IIVs são detalhados na sequência.

- IIV 1 – os coeficientes da regressão PLS, b_{mk}, descrevem a relação entre a variável de processo k ($k = 1, ..., K$) e a variável de resposta m ($m = 1, ..., M$), sendo sua magnitude um indicativo da importância de cada variável de processo para a predição de y. O índice IIV 1 aparece na Equação (2).

$$IIV1_k = \sum_{m=1}^{M} b_{mk}^2 \quad k = 1, ..., K \text{ variáveis de processo} \qquad (2)$$

- IIV 2 – o vetor **w** dita a influência da variável independente k na composição do componente a. De tal forma, o valor absoluto de **w** permite ordenar as variáveis de processo de acordo com sua relevância.

O índice proposto é gerado através da soma de $|w_k|$ sobre o conjunto de l componentes PLS retidos para análise. Valores elevados na Equação (3) apontam as variáveis de maior importância.

$$IIV2_k = \sum_{a=1}^{l} |w_{ak}| \; k = 1, \ldots, K \text{ variáveis de processo} \quad (3)$$

- IIV 3 é baseado na "Importância de Projeção de Variável", proposta por Wold et al. (2001a), e associa o vetor de pesos **w** do componente a à variância em **Y** explicada pelo mesmo componente (R^2_{Ya}), conforme apresentado na Equação (4). Tendo-se em vista que a magnitude dos elementos do vetor **w** quantifica a influência da variável de processo k sobre o componente a e R^2_{Ya} indica a importância do componente para predição de **Y**, o índice proposto é considerado consistente para seleção das variáveis.

$$IIV3_k = \sum_{a=1}^{l} |w_{ak}| R^2_{Ya} \; k = 1, \ldots, K \text{ variáveis de processo} \quad (4)$$

A utilização dos IIVs acima apresentados para seleção de variáveis relevantes é justificada i) por sua capacidade em identificar as variáveis x detentoras da maior variância em si; e ii) por selecionar as variáveis x mais aptas a explicarem a variância das variáveis de produto y. Enfatiza-se que cada vetor IIV pode dar origem a uma ordem distinta de importância para as variáveis x, visto que cada índice foi gerado utilizando-se parâmetros distintos.

A integração de IIV às ferramentas de classificação ZVP e MSV é descrita na sequência. Note que IIV cumpre a função de ordenar as variáveis mais relevantes para descrição do processo, ao passo que as ferramentas de classificação visam determinar o número recomendado de variáveis (de acordo com a ordem estabelecida pelos IIVs) a ser utilizado em procedimentos de classificação.

3.3 Passo 3 – Eliminação das variáveis de processo menos relevantes e categorização das observações em classes

As porções de treino e teste geradas no passo 1 são consideradas. A variável de produto y é utilizada para rotular cada observação (batelada de produção) de ambas as porções em uma de duas classes de qualidade através da comparação de y com a especificação definida para o produto. Em um processo do tipo maior-é-melhor, por exemplo, bateladas com y inferior ao limite mínimo são rotuladas como 0 (baixa qualidade), enquanto que bateladas com y superior ao limite assumem valor 1 (alta qualidade). O instrumento de medição de y depende da natureza do processo em análise e o limite de especificação é normalmente definido pelos engenheiros de desenvolvimento de produto.

Na sequência, uma classificação inicial utilizando as K variáveis de processo é realizada e a acurácia de classificação é computada. Neste estudo, acurácia é definida como a razão entre o número de classificações corretas e o número total de casos classificados. Uma classificação é dita correta quando a classe prevista pela ferramenta de classificação é idêntica à classe estabelecida através da comparação de y com a especificação definida para o produto.

IIV é então utilizado para guiar o processo de remoção das variáveis. A variável detentora do menor valor em IIV é eliminada das porções de treino e teste e uma nova classificação utilizando as K-1 variáveis remanescentes é realizada. A acurácia de classificação é novamente computada. O procedimento de classificação/eliminação de variáveis é interrompido quando restam apenas duas variáveis de processo. Esse limite inferior é adotado para preservar a habilidade de categorização das ferramentas de classificação, visto que a utilização de uma única variável de processo no procedimento de classificação não assegura categorizações precisas (DUDA et al., 2001).

O procedimento de classificação/eliminação é conduzido utilizando-se as duas ferramentas de classificação, ZVP e MSV, separadamente, com vistas à posterior comparação de desempenho. Os parâmetros para ZVP e MSV são estimados através de validação cruzada na porção de treino, maximizando-se a medida de acurácia. De forma semelhante, o procedimento de classificação/eliminação é repetido para todos os IIVs gerados no passo 2 para fins de comparação de desempenho. Objetiva-se, assim, eleger a combinação de IIV e ferramenta de classificação que conduza aos melhores índices de redução de variáveis e incremento na acurácia de classificação.

Um gráfico associando número de variáveis removidas e acurácia de classificação é gerado para monitorar o processo de eliminação. O ponto de máxima acurácia nesse gráfico indica o subconjunto de variáveis recomendadas para propósitos de classificação. No caso de distintos subconjuntos de variáveis conduzirem a picos idênticos de acurácia, opta-se pelo subconjunto com menor número de variáveis remanescentes (assumindo-se que um menor número de variáveis é desejado).

4 *Exemplo numérico*

Três bancos de dados foram analisados, conforme apresentado na Tabela 1. Dados do processo 1 descrevem um estágio de polimerização na produção de látex. As variáveis x indicam temperaturas, pressões e tempos de reação, enquanto que a variável y mede a quantidade de subprodutos gerados pelo processo, a qual deve ser minimizada. O processo 2 refere-se à geração de subcomponentes para fabricação de papel. As variáveis de processo medem temperaturas e concentrações de reagentes, ao passo que a variável y descreve a viscosidade do produto final.

Tabela 1. Bancos de dados considerados.

Banco de dados	Número de variáveis		Número de observações	
	y	*x*	Porção de treino	Porção de teste
Processo 1	1	121	195	52
Processo 2	1	73	34	13
Processo 3	4	20	1.000	500

O banco de dados para o processo 3 foi gerado através de simulação visando testar a consistência da metodologia quando aplicada em cenários caracterizados por elevado número de observações e múltiplas variáveis de produto. O processo simulado refere-se à produção de vinho, sendo que as variáveis de processo descrevem características qualitativas da uva (tida como principal matéria-prima), temperaturas e tempos de fermentação dos estágios iniciais de processo, além da quantidade de reagentes adicionados. Especialistas estimaram as correlações entre as variáveis. Uma mil e quinhentas observações normalmente distribuídas (simulando bateladas de produção) foram geradas em MATLAB 7.4 através do comando *mvnrnd*. Regressões lineares múltiplas geraram variáveis de produto descrevendo nível de álcool (y_1), acidez (y_2), resíduos sólidos (y_3) e nota final incluindo aspectos sensitivos (y_4). Os coeficientes das regresssões foram definidos por especialistas. Um termo de erro $N\sim(0, \sigma^2)$ foi adicionado às regressões, sendo σ^2 baseado na variância média das variáveis de processo.

As observações de cada banco de dados (representando bateladas de produção) foram classificadas em dois níveis de qualidade, de acordo com especificações das variáveis *y* (produto de alta qualidade = 1, produto de baixa qualidade = 0). As especificações de classificação foram determinadas por especialistas no processo.

A regressão PLS foi então aplicada à porção de treino dos processos da Tabela 1. Três componentes PLS foram retidos em cada processo, com base no percentual de variância explicada em **Y** (processo 1 = 91%, processo 2 = 83%, processo 3 = 87%).

Os parâmetros gerados pela regressão PLS foram manipulados matematicamente para obtenção de IIV, e então integrados às ferramentas de classificação ZVP e MSV, conforme descrito nos passos 2 e 3, respectivamente. Os parâmetros das ferramentas de classificação foram determinados por intermédio de 25 repetições de validação cruzada, sendo que cada banco de dados foi dividido em 5 partes em cada procedimento. A ferramenta ZVP utilizou $z = 3$ vizinhos mais próximos para os processos 1 e 2, e $z = 7$ para o processo 3. A ferramenta MSV utilizou as seguintes funções Kernel: processo 1 – polinômio de segundo grau, processo 2 – função radial gaussiana com amplitude 9, e processo 3 – polinômio de quarto grau.

A Figura 1 apresenta os perfis de acurácia gerados pelas duas ferramentas de classificação com a remoção de variáveis do processo 1 e utilização do IIV 1. A ferramenta ZVP apresenta patamares mais elevados de acurácia durante o processo de eliminação das variáveis. No entanto, ambos os métodos convergem a níveis semelhantes de performance no pico de acurácia, quando 117 variáveis das 121 iniciais são removidas pela ferramenta ZVP e 119 pela MSV.

A Tabela 2 traz um resumo da performance dos métodos e IIVs testados. O percentual de aumento de acurácia é calculado através da comparação da acurácia máxima gerada pelo método (pico de cada perfil) com a acurácia obtida ao aplicar-se ZVP e MSV ao banco de dados contendo todas as variáveis originais. Percebe-se que todos os cruzamentos de ferramentas de classificação e IIV conduzem a categorizações mais acuradas ao reduzir o número de variáveis retidas.

A identificação da melhor combinação de IIV e ferramenta de classificação é sistematizada através do Indicador Geral de Performance (IGP), o qual unifica o aumento de acurácia e a redução de variáveis geradas por cada combinação. O IGP foi concebido por conta da heterogeneidade dos resultados obtidos pelos diferentes cruzamentos entre IIV e ferramentas de classificação, o que dificulta a identificação da melhor combinação através da análise simples dos resultados. Valores elevados de IGP são desejáveis, visto que denotam significativa redução no número de variáveis e significativo aumento na acurácia de classificação. O IGP é apresentado na Equação (5):

$$IGP_{ic} = \left[\frac{AA_{ic}}{max\ AA} + \left(1 - \frac{VR_{ic}}{VI} \right) \right] \Big/ 2 \qquad 0 \le \text{IGP} \le 1 \tag{5}$$

em que AA_{ic} refere-se ao aumento de acurácia ao utilizar-se o índice de importância *i* (IIV 1, IIV 2 e IIV 3) e o método de classificação *c* (ZVP e MSV). *Max AA* refere-se ao máximo aumento de acurácia obtido em cada banco de dados (considerando-se todas as combinações possíveis de métodos e IIVs), enquanto que *VR* e *VI* indicam o número de variáveis retidas e o número inicial de variáveis, respectivamente. Elevados valores de IGP na Tabela 3 indicam combinações de IIV e ferramentas de classificação capazes de reduzir significativamente o número de variáveis retidas e, ao mesmo tempo, de conduzir a incrementos na acurácia de classificação.

A Tabela 3 aponta IIV 1 como o melhor índice para a identificação das variáveis de processo mais relevantes, visto que conduz ao maior valor médio de IGP. Em relação às ferramentas de classificação, o IGP médio sinaliza MSV como a melhor opção, conforme Tabela 4.

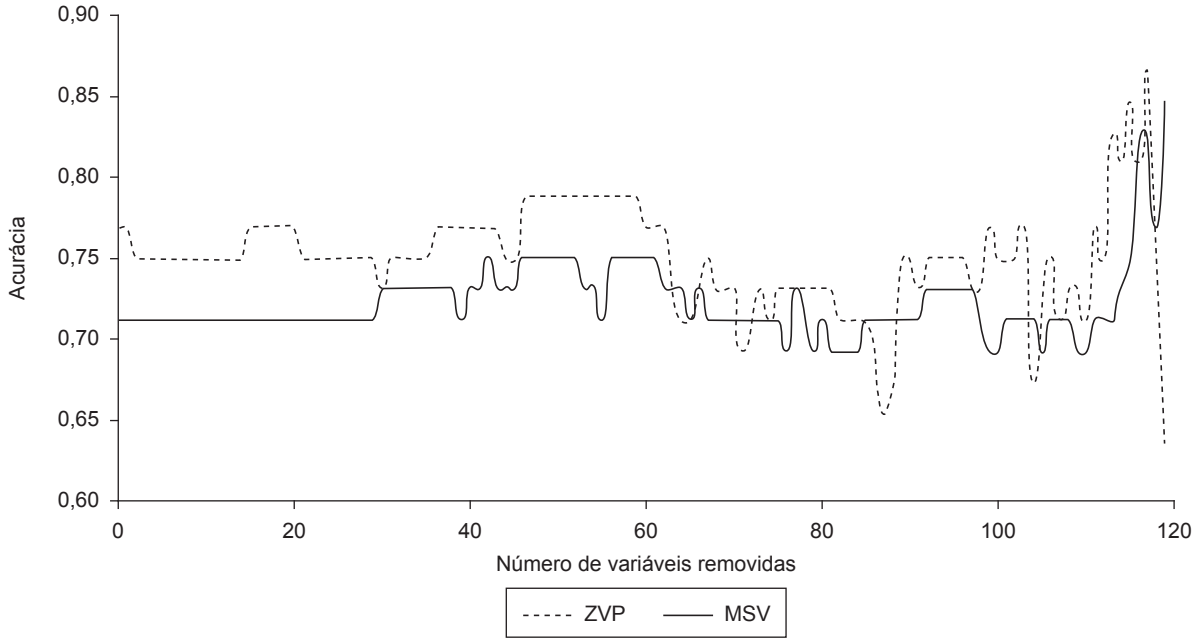

Figura 1. Evolução do perfil de acurácia com a remoção de variáveis do banco processo 1 utilizando IIV 1.

Tabela 2. Aumento de acurácia e nível de redução de variáveis geradas pelas distintas ferramentas de classificação e IIVs.

		Índice de importância da variável (IIV)					
		IIV 1		IIV 2		IIV 3	
Banco de dados (número original de variáveis)	Ferramenta de classificação	Aumento de acurácia (%)	Número de variáveis retidas	Aumento de acurácia (%)	Número de variáveis retidas	Aumento de acurácia (%)	Número de variáveis retidas
Processo 1 (121)	ZVP	12,5	4	12,5	8	12,5	9
	MSV	16,2	2	5,4	3	2,5	35
Processo 2 (73)	ZVP	0,3	9	8,3	8	8,3	2
	MSV	57,1	12	37,5	15	57,1	2
Processo 3 (20)	ZVP	10,6	5	10,6	5	11,3	4
	MSV	14,2	2	14,2	2	14,2	2

Tabela 3. Indicador Geral de Performance (IGP) para as distintas ferramentas de classificação e IIVs.

Banco de dados	Ferramenta de classificação	IIV 1	IIV 2	IIV 3
Processo 1	ZVP	0,87	0,85	0,85
	MSV	0,99	0,65	0,43
Processo 2	ZVP	0,42	0,50	0,55
	MSV	0,89	0,69	0,98
Processo 3	ZVP	0,75	0,75	0,80
	MSV	0,95	0,95	0,95
	Média desvio	0,81	0,73	0,76
		0,21	0,16	0,22

Tabela 4. Indicador Geral de Performance (IGP) médio para as ferramentas de classificação.

Ferramenta de classificação	Média	Desvio
ZVP	0,70	0,17
MSV	0,83	0,20

Combinando-se IIV 1 e a ferramenta de classificação MSV nos dados do processo 1, verifica-se que 2 das 121 variáveis iniciais devem ser utilizadas para classificação de bateladas de produção futuras, conduzindo a classificações 16,2% mais precisas quando comparada à performance da classificação obtida ao utilizar-se a totalidade de variáveis. Para o processo 2, IIV 1 e MSV utilizam somente 12 das 73 variáveis iniciais, representando um aumento de 57,1% na acurária de classificação. Ao aplicar-se a mesma combinação de método e IIV para o processo 3, verifica-se que 2 das 20 variáveis iniciais conduzem a categorizações 14,2% mais acuradas, quando comparadas à utilização variáveis. Em média, nos três bancos de dados, verificou-se redução de 90,6% no número de variáveis retidas e elevação da acurácia média de classificação em 29,2%, quando comparadas à aplicação da MSV sobre todas as variáveis.

5 *Conclusão*

Neste artigo, a regressão PLS foi integrada às ferramentas de classificação *z* vizinhos mais próximos (ZVP) e Máquina de Suporte Vetorial (MSV), com vistas à seleção de variáveis de processo para categorização de bateladas de produção em duas classes. Índices de Importância da Variável baseados nos parâmetros da regressão PLS foram desenvolvidos para o ordenamento das variáveis de processo de acordo com sua relevância, e então integrados ao ZVP e MSV. O monitoramento dos perfis de acurácia gerados com a remoção sistemática de variáveis permitiu definir o subconjunto recomendado de variáveis a ser utilizado na categorização de bateladas futuras de produção em duas classes de qualidade. Por fim, gerou-se o Índice Geral de Performance (IGP) para balizar a escolha da melhor combinação de IIV e ferramenta de classificação. O IGP avalia a performance da combinação em termos da capacidade de redução de variáveis e incremento na acurácia de classificação.

A metodologia foi aplicada em três bancos de dados (dois reais e um simulado) caracterizados por elevados níveis de correlação entre variáveis e ruído, além de distintas proporções entre número de observações e variáveis. O IGP médio apontou o índice IIV 1 e a ferramenta de classificação MSV como a combinação de melhor performance em termos de redução de variáveis e aumento da acurácia, revelando-a a combinação recomendada para seleção de variáveis com vistas à classificação de bateladas produtivas. A combinação IIV 1 e MSV reduziu o número médio de variáveis de processo necessárias para classificação de bateladas em 90,6% nos três processos analisados aumentou a acurácia média de classificação em 29,2%, quando comparada à aplicação da MSV sobre todas as variáveis.

Deve-se acrescentar que, por estar alicerçada na composição de técnicas distintas, a metodologia proposta demanda diversas suposições, como a consideração de que a variável de produto é tratada como binária (0 ou 1). Tais considerações, todavia, são compensadas pelo ganho de acurácia que a utilização das variáveis recomendadas para classificação garante e pelos ganhos advinhos da redução no número de variáveis a serem coletadas para controle de processo.

Desdobramentos futuros incluem a utilização de ferramentas alternativas de classificação, como Redes Neurais Probabilísticas (*Probabilistic Neural Network* – PNN), e posterior comparação da performance de classificação. A geração de índices de importância de variável focados em aspectos preditivos da variável *y* em cenários com múltiplas variáveis de resposta também constitui-se em tópico de interesse.

Identifying relevant variables for production batch categorization into quality levels

Abstract

A large number of correlated process variables are usually found in industrial environments, making it difficult for engineers to identify the key variables. Partial Least Squares (PLS) has been successfully applied to select the most relevant process variables for predicting response variables. However, many practical applications are more interested in correctly categorizing the final product into classes. This paper addresses this classification issue by integrating Partial Least Square (PLS) regression to the z-nearest neighbor rule and support vector machine for the categorization of production batches into two quality levels. Indices based on PLS parameters are developed for evaluating variable importance. The classification methods are then applied to reduce noisy and irrelevant variables based on the importance indices. The best subset of variables is identified by monitoring accuracy profile variations while variables are removed. In three datasets, the suggested approach reduced the number of variables necessary for classification of production batches by 90.6 per cent, while yielding 29.2 per cent more accurate classifications.

Keywords: *Variable selection. PLS. z-Nearest neighbors classification rule. Support vector machine.*

Referências bibliográficas

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. **Chemometrics and Intelligent Laboratory Systems,** v. 97, n. 2, p. 111-117, 2009.

CHAOVALITWONGSE, W.; FAN, Y.; SACHDEO, C. On the time series k-nearest neighbor classification of abnormal brain activity. **IEEE Transactions on Systems, Man and Cybernetics A,** v. 37, n. 6, p. 1005-1016, 2007.

CHONG, I.; JUN, C. Performance of some variable selection methods when multicollinearity is present. **Chemometrics and Intelligent Laboratory Systems,** v. 78, n. 1-2, p. 103-112, 2005.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines and other kernel-based learning methods.** Cambridge: Cambridge University Press, 2000.

DUDA, R.; HART, P.; STORK, D. **Pattern Classification.** 2 ed. New York: Wiley-Interscience, 2001.

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, n. 2, p. 165-184, 1999.

GARTHWAITE, P. An interpretation of Partial Least Squares. **Journal of the American Statistical Association,** v. 89, n. 425, p. 122-127, 1994.

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, n. 2, p. 171-193, 2001.

GELADI, P.; KOWALSKI, B. Partial least-squares regression: a tutorial. **Analytica Chimica Acta,** v. 185, p. 1-17, 1986.

GEORGE, E. The variable selection problem. **Journal of the American Statistical Association,** v. 95, n. 452, p. 1-12, 2000.

GOLUB, T. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. **Science,** v. 286, n. 5439, p. 531-537, 1999.

GUYSON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research,** v. 3, n. 7-8, p. 1157-1182, 2003.

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, n. 1-2, p. 23-38, 2001.

HUANG, T.; KEDMAN, V.; KOPRIVA, I. **Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi-supervised, and Unsupervised Learning.** Berlin, Heidelberg: Springer-Verlag, 2006.

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS for very large data sets. **Computational Statistics & Data Analysis,** v. 48, n. 1, p. 69-85, 2005.

LAZRAQ, A.; CLEROUX, R. The PLS multivariate regression model: testing the significance of successive PLS components. **Chemometrics and Intelligent Laboratory Systems,** v. 15, n. 6, p. 523-536, 2001

LAZRAQ, A.; CLEROUX, R.; GAUCHI, J. Selecting both latent and exploratory variables in the PLS1 regression model. **Chemometrics and Intelligent Laboratory Systems,** v. 66, n. 2, p. 117-126, 2003.

MEROLA, G.; ABRAHAM, B. Dimensionality reduction approach to multivariate prediction. **The Canadian Journal of Statistics,** v. 29, n. 2, p. 1-12, 2001.

MONTGOMERY, D. **Introduction to Statistical Quality Control.** 4 ed. New York: Wiley & Sons, 2001.

PHILIPHS, R.; GUTTMAN, I. A new criterion for variable selection. **Statistics & Probability Letters,** v. 38, n. 1, p. 11-19, 1998.

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

WEISS, S. et al. Maximizing text-mining performance. **IEEE Intelligent Systems,** v. 14, n. 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, n. 2, p. 109-130, 2001a.

WOLD, S. et al. Some recent developments in PLS modeling. **Chemometrics and Intelligent Laboratory Systems,** v. 58, n. 2, p. 131-150, 2001b.

XU, D.; ALBIN, S. Manufacturing start-up problem solved by mixed-integer quadratic programming and multivariate statistical modeling. **International Journal of Production Research,** v. 40, n. 3, p. 625-640, 2002.

XU, L. et al. Variable-weighted PLS. **Chemometrics and Intelligent Laboratory Systems,** v. 85, n. 1, p. 140-143, 2007.

ZHAI, H.; CHEN, X.; HU, Z. A new approach for the identification of important variables. **Chemometrics and Intelligent Laboratory Systems,** v. 80, n. 1, p. 130-135, 2006.

Sobre o autor

Michel José Anzanello

Universidade Federal do Rio Grande do Sul – UFRGS
Av. Paulo Gama, 110, CEP 90040-060, Porto Alegre – RS, Brasil
e-mail: anzanello@producao.ufrgs.br

Recebido em 7/5/2009
Aceito em 9/11/2009