IDENTIFICAÇÃO DE VARIÁVEIS RELEVANTES PARA CATEGORIZAÇÃO DE BATELADAS DE PRODUÇÃO COM BASE EM CRITÉRIOS DE SENSIBILIDADE E ESPECIFICIDADE

Michel J Anzanello
Universidade Federal do Rio Grande do Sul
Av. Osvaldo Aranha, 99 – 5° andar, Porto Alegre – RS
anzanello@producao.ufrgs.br

Flavio S Fogliatto
Universidade Federal do Rio Grande do Sul
Av. Osvaldo Aranha, 99 – 5° andar, Porto Alegre – RS
ffogliatto@producao.ufrgs.br

RESUMO

Variáveis correlacionadas e ruidosas tendem a reduzir a eficiência de métodos para controle e monitoramento de processos industriais. Este artigo apresenta um método para seleção das variáveis mais relevantes para categorização de bateladas de produção em duas classes valendo-se de múltiplos critérios de desempenho de classificação (sensibilidade e especificidade). Um índice de importância das variáveis de processo é derivado dos parâmetros da regressão PLS (*Partial Least Squares*). As bateladas são então classificadas através da ferramenta *k*-Nearest Neighbor (KNN) e as medidas de desempenho avaliadas. Na sequência, a variável com menor índice de importância é removida e uma nova classificação é executada com base nas variáveis remanescentes. Esse processo é repetido até que reste apenas uma variável. A distância Euclidiana identifica o subconjunto de variáveis recomendado. A aplicação do método proposto em seis processos industriais distintos reduziu significativamente o número de variáveis retidas e aumentou o desempenho de classificação.

PALAVRAS CHAVE: Seleção de variáveis, sensitividade, especificidade, regressão PLS

Área principal: PO na Indústria

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 two classes based on multiple criteria (e.g., sensibility and specificity). 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 the Euclidian distance is generated to identify the recommended subset of variables. When applied to six datasets of industrial, the proposed method significantly reduced the number of retained variables and increased the classification performance.

KEYWORDS: Variable selection, sensitivity, specificity, PLS regression

Main area - PO na Indústria

1. Introdução

A maioria dos métodos para seleção das variáveis mais relevantes em processos industriais tem priorizado a predição de variáveis de produto (GAUCHI; CHAGNON, 2001; MEIRI; ZAHAVI, 2006; OZTURK et al., 2006; OLAFSSON et al., 2008). Este artigo, no entanto, objetiva selecionar as variáveis de processo mais relevantes com vistas à categorização de bateladas de produção com base em múltiplos critérios de desempenho de classificação, como sensibilidade e especificidade.

A acurácia de classificação, definida como fração de observações corretamente classificadas, tem figurado com principal critério para avaliação de desempenho em procedimentos de seleção de variáveis com propósito de categorização (ANZANELLO et al., 2009). No entanto, existem situações em que outros critérios são mais apropriados. Um exemplo vem da indústria farmacêutica, onde a incorreta classificação de uma batelada de medicamento 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. Em contrapartida, a classificação equivocada de uma batelada conforme pode acarretar impactos financeiros elevados em diversas aplicações industriais, apontando sensibilidade (fração de bateladas conformes corretamente classificadas) como o critério mais apropriado.

A utilização de critérios como sensibilidade e especificidade como balizadores para a seleção de variáveis não tem encontrado aplicação recente em cenários industriais. Abordagens deste cunho têm sido sugeridas em reconhecimento de texto, análise financeira e sistemas de segurança (ROSE-PEHRSSON et al., 2000; DOAN; HORIGUCHI, 2004, PIRAMUTHU, 2004; PENDARAKI et al., 2005; HUANG et al., 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).

O método proposto neste artigo aplica a regressão PLS (*Partial Least Squares*) na porção de treino de um banco de dados. Os parâmetros gerados dão origem a um índice de importância para cada variável de processo j, v_j, o qual baliza um processo de eliminação de variáveis seguindo a lógica *backward*. As observações (bateladas) da porção de treino descritas por todas as variáveis são categorizadas em duas classes através da ferramenta *k-Nearest Neighbor* (KNN), e as medidas de sensibilidade e especificidade são computadas. Na sequência, a variável com o menor v_j 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 reste apenas uma variável. O subconjunto ideal é identificado com base na distância Euclidiana de cada subconjunto candidato a um ponto hipotético tido como ideal. O subconjunto selecionado é validado na porção de teste.

A aplicação do método na porção de teste de seis bancos de dados industriais reteve, em média, 12% das variáveis originais. As variáveis selecionadas elevaram a sensibilidade de classificação em 9%, de 0,78 para 0,85, enquanto que a especificidade aumentou 20%, de 0,64 para 0,77.

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

2. Referencial Teórico

Esta seção traz os fundamentos da regressão PLS e ferramenta de classificação *k*-Nearest Neighbor (KNN).

A regressão PLS vem sendo amplamente utilizada em aplicações industriais onde as variáveis de processo apresentam elevados níveis de correlação, ruído, observações faltantes e desequilíbrio na proporção de variáveis e observações; ver Wold et al. (2001a) Kettaneh, et al. (2005), Nelson et al. (2006), e Hoskuldsson (2001). Tal regressão gera um reduzido número de

combinações lineares independentes (também chamadas de componentes PLS) das variáveis de processo. Essas combinações respondem por parte significativa da variância das variáveis originais do processo. Normalmente, apenas três ou quatro componentes PLS são retidos para representar dezenas ou mesmo centenas de variáveis de processo.

A regressão PLS é operacionalizada como segue. 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,

$$t_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + ... + w_{Ja}x_{iJ} = \mathbf{w}_a'\mathbf{x}_i, \quad (1)$$

com $A \leq J$. O número de componentes, *A*, é geralmente pequeno, e pode ser definido através de validação cruzada. 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 et al., 2001a). Similarmente, componentes são construídos para as variáveis de produto **Y**, ou seja,

$$u_{ia} = c_{1a}y_{i1} + c_{2a}y_{i2} + ... + c_{Ma}y_{iM} = \mathbf{c}_a'\mathbf{y}\boldsymbol{i}, \quad (2)$$

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 (XU; ALBIN, 2002). Outro parâmetro de interesse é o vetor de carga, $\mathbf{p}_a = (p_{1a}, p_{2a}, ..., p_{Ja})'$, gerado pela regressão das colunas de **X** em relação a $\mathbf{t}_a$. Tal parâmetro cumpre importante função na geração do índice de importância proposto na Seção 3.

Os parâmetros da regressão PLS podem ser estimados através do algoritmo NIPALS; ver Goutis (1997), Abdi (2003) e Geladi e Kowalski (1986). Detalhes matemáticos da regressão PLS podem ser obtidos em Westerhuis et al. (1998) e Wold et al. (2001a, b). A regressão PLS pode ser operacionalizada através do toolbox PLS, encontrado em pacotes estatísticos como Matlab® e R®.

A ferramenta de classificação KNN, por sua vez, encontra ampla utilização por conta de sua simplicidade conceitual e disponibilidade em pacotes estatísticos. Considere *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.

Algumas aplicações da ferramenta KNN incluem a classificação de genes em Golub et al. (1999), reconhecimento de texto em Weiss et al. (1999), detecção de atividade cerebral anormal em Chaovalitwongse et al. (2007) e seleção de atributos sensoriais em Anzanello et al. (2011). 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 três passos: (1) geração de um índice de importância das variáveis de processo com base nos parâmetros da regressão PLS; (2) categorização das bateladas em duas classes e eliminação das variáveis irrelevantes; e (3) aplicação da análise de Pareto Ótimo no perfil de desempenho gerado pela eliminação de variáveis e identificação da melhor solução da fronteira do Pareto via distância Euclidiana. Tais passos são detalhados na sequência.

No primeiro passo do método proposto, gera-se um índice de importância da variável de processo com base nos parâmetros da regressão PLS. Para tanto, dados de N bateladas são randomicamente divididos 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 et al., 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 et al., 2009).

O índice de importância v_j é gerado na equação (4), sendo que w^*_{ja} é obtido através da equação (3).

$$w^*_{ja} = w_{ja}(p_{ja}w_{ja})^{-1} \tag{3}$$

Wold et al. (2001a) 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).

$$v_j = \sum_{a=1}^{A} (w^*_{ja})^2 R^2_{Ya} \qquad j=1,\ldots,J. \tag{4}$$

Wold et al. (2001a) 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 et al., 2009).

No segundo passo do método proposto, inicia-se um processo iterativo para 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 (sensibilidade e especificidade) são calculados. Tais critérios de desempenho são definidos como segue.

Considere quatro possibilidades de classificação (CHAOVALITWONGSE et al., 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 (5); similarmente, especificidade é dada pela fração de bateladas não-conformes corretamente categorizadas, conforme a equação (6).

$$Sensibilidade = \frac{PV}{PV + FN} \tag{5}$$

$$Especificidade = \frac{NV}{NV + FP} \quad (6)$$

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.

O terceiro passo inicia após concluir-se o processo de eliminação das variáveis. Constrói-se inicialmente um gráfico associando sensibilidade, especificidade e percentual de variáveis retidas. Cada ponto deste gráfico refere-se ao desempenho de classificação 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 critérios de desempenho e percentual de variáveis retidas.

Na sequência, aplica-se a análise de Pareto Ótimo (PO) para identificar soluções diferenciadas no perfil gerado pela eliminação das variáveis. 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 (AZAPAGIC, 1999). 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 et al., 1994; ZITZLER; THIELE, 1999; TABOADA; COIT, 2007, 2008).

Por fim, os pontos da fronteira do perfil gerado pela eliminação das variáveis 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. Tais coordenadas são definidas pelo usuário. O ponto de fronteira com a menor distância ao ponto ideal consiste no subconjunto recomendado pelo método, sendo estão validado na porção de teste do banco de dados.

4. Estudo de caso

O método proposto é aplicado em dados de seis processos industriais obtidos em Gauchi e Chagnon (2001) e Wold et al. (2001a). 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 que a variável de resposta denota uma característica do produto, como viscosidade ou teor de pureza.

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	Emulsão 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

Fonte: autores

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 et al. (2001a).

Na sequência, aplicou-se a regressão PLS na porção de treino de cada processo, retendo-se três componentes para cada processo através de validação cruzada (ver WOLD et al., 2001a). A variância em **Y** explicada pelos componentes retidos em cada banco de dados, bem como o melhor parâmetro *k* para a ferramenta de classificação KNN (estimado através de validação cruzada), são apresentados na Tabela 2.

Tabela 2. Variância explicada em **Y** pelos componentes retidos e parâmetro *k*

Banco de dados	Variância em **Y** explicada pelos componentes retidos (%)	Parâmetro *k*
ADPN	94	3
GRANU	75	3
LATEX	77	3
OXY	94	3
PAPER	68	3
SPIRA	71	9

A Figura 1 ilustra o perfil de sensibilidade/especificidade gerado para a porção de treino do processo GRANU à medida que variáveis de processo são eliminadas. O método sugere que somente 24% das variáveis originais sejam utilizadas, elevando a sensibilidade em 7%, de 0,88 para 0,94, e a especificidade em 2%, de 0,88 para 0,90.

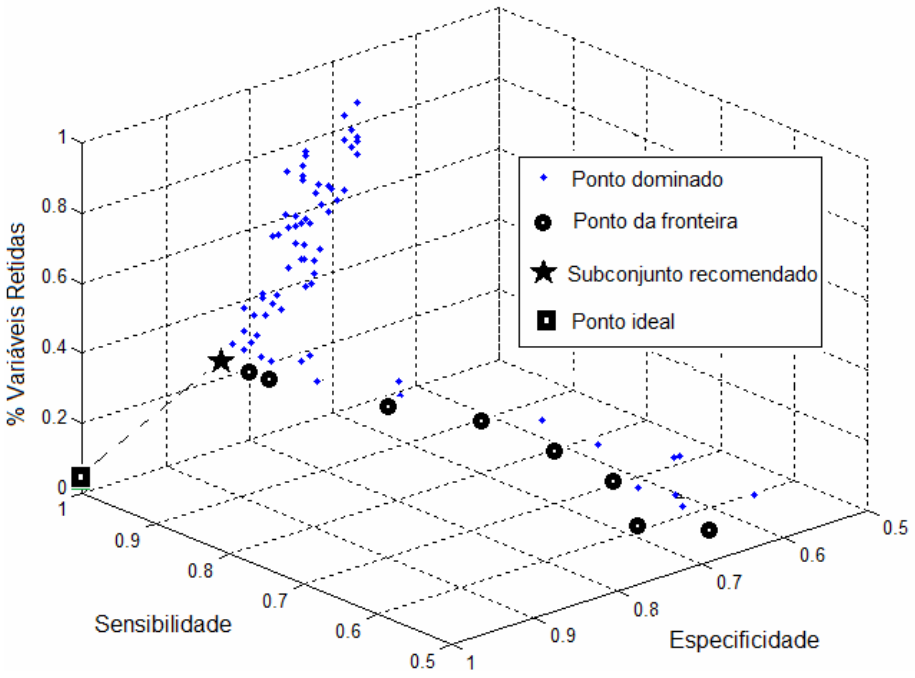

Figura 1. Perfil de sensibilidade/especificidade para o processo GRANU

Em relação aos seis processos analisados, o método proposto reteve, em média, 12% das variáveis originais (Tabela 3). A sensibilidade da porção de treino foi elevada em 8%, de 0,86 para

0,93, enquanto que a especificidade da porção de treino aumentou 14%, de 0,78 para 0,89. Variações no desempenho de classificação são justificadas pela diversidade de processos analisados.

Tabela 3. 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	Sem seleção de variáveis	Método proposto	Sem seleção de variáveis	
ADPN (100)	99	93	71	63	7
GRANU (78)	94	88	90	88	24
LATEX (117)	97	94	92	80	7
OXY (95)	98	98	95	67	8
PAPER (54)	75	48	94	93	9
SPIRA (96)	96	96	92	79	19
Média	93	86	89	78	12

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 4).

Tabela 4. 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	Sem seleção de variáveis	Método proposto	Sem seleção de variáveis
ADPN (100)	100	100	62	25
GRANU (78)	95	87	78	80
LATEX (117)	93	81	71	73
OXY (95)	96	97	88	57
PAPER (54)	35	20	90	86
SPIRA (96)	92	83	75	65
Média	85	78	77	64

5. Conclusão

Este artigo apresentou um método para seleção de variáveis com base em múltiplos critérios de desempenho. As etapas do método são: (1) geração de um índice de importância da variável de processo com base nos parâmetros da regressão PLS; (2) categorização das bateladas em duas classes e eliminação das variáveis irrelevantes; e (3) identificação do melhor subconjunto de variáveis através da análise de Pareto Ótimo.

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.

Desdobramentos futuros incluem a extensão do método proposto para cenários em que 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 múltiplas classes (três ou mais), o que demanda aprimoramento na ferramenta de classificação KNN.

Referências

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

Anzanello, M.; Albin, S. e Chaovalitwongse, W. (2009), Selecting the best variables for classifying production batches into two quality classes. *Chemometrics and Intelligent Laboratory Systems*, 97, 111-117.

Aragonés-Beltrán, P.; Aznar, J.; Ferrís-Oñate, J. e García-Melón, M., (2008) Valuation of urban industrial land: An analytic network process approach. *European Journal of Operational Research*, 185, 322-339.

Chaovalitwongse, W.; Fan, Y. e Sachdeo, C., (2007), On the time series k-nearest neighbor classification of abnormal brain activity. *IEEE Transactions on System and Man Cybernetics A*, 37, 1005-1016.

Chong, I.; Albin, S. e Jun, C. (2007), A data mining approach to process optimization without an explicit quality function. *IIE Transactions*, 39, 795-804.

Doan, S. e Horiguchi, S. An efficient feature selection using multi-criteria in text categorization. In: Fourth International Conference on Hybrid Intelligent Systems, 2004, HIS '04, 86-91, 2004.

Gauchi, J. e Chagnon, P., (2001), Comparison of selection methods of exploratory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, 58, 171-193.

Geladi, P. e Kowalski, B. (1986), Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185, 1-17.

Golub, T.; Slonim, D.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.; Coller, H.; Loh, M.; Downing, J.; Caligiuri, M.; Bloomeld, C. e Lander, E., (1999), Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.

Goutis, C. (1997), A fast method to compute orthogonal loadings partial least squares. *Journal of Chemometrics*, 11, 13-32.

Horn, J.; Nafpliotis, N. e Goldberg, D. (1994), A niched pareto genetic algorithm for multiobjective optimization. In: Proceedings of the First IEEE Conference on Evolutionary Computation, *IEEE World Congress on Computational Intelligence*, 1, 82-87.

Hoskuldsson, A. (2001), Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 55, 23-38.

Huang, J.; Tzeng, G. e Ong, C. (2006), Optimal fuzzy multi-criteria expansion of competence sets using multi-objectives evolutionary algorithms. *Expert Systems with Applications*, 30, 739-745.

Kettaneh, N.; Berglund, A. e Wold, S. (2005), PCA and PLS in very large datasets. *Computational Statistics & Data Analysis*, 48, 69-85.

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

Meiri, R. e Zahavi, J. (2006), Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 171, 842-858.

Nelson, P.; Macgregor, J. e Taylor, P. (2006), The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemometrics and Intelligent Laboratory Systems*, 80, 1-12.

Olafsson, S.; Li, X. e Wu, S. (2008), Operations research and data mining. *European Journal of Operational Research*, 187, 1429-1448.

Ozturk, A.; Kayaligil, S. e Ozdemirel, N. (2006), Manufacturing lead time estimation using data mining. *European Journal of Operational Research*, 173, 683-700.

Pendaraki, K.; Zopounidis, C. e Doumpos, M. (2005), 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*, 163, 462-481.

Piramuthu, S. (2004), Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research*, 156, 483-494.
Ridgeway, G. *The handbook of data mining*, in N. Ye (Editor). Lawrence: New Jersey, 2003.
Rose-Pehrsson, S.; Shaffer, R.; Hart, S.; Williams, D.; Gottuk, D.; Strehlen, B. e HILL, S. (2000), Multi-criteria fire detection systems using a probabilistic neural network. Sensors and Actuators B: Chemical, 69, 325-335.
Sueyoshi, T. (2006), DEA-Discriminant Analysis: Methodological comparison among eight discriminant analysis approaches. *European Journal of Operational Research*, 169, 247-272.
Taboada, H. e Coit, D. (2007), Data clustering of solutions for multiple objective system reliability optimization problems. *Quality Technology & Quantitative Management Journal*, 4, 35-54.
Taboada, H. e Coit, D. (2008), Multi-objective scheduling problems: Determination of pruned Pareto sets. *IIE Transactions*, 40, 552-564.
Weiss, S.; Apte, C.; Dameray, D.; Johnson, D.; Ples, F.; Goetz, T. e Hampp, T. (1999), Maximizing text-mining performance. *IEEE Intelligent Systems*, 14, 63-69.
Westerhuis, J.; Kourti, T. e Macgregor, J. (1998), Analysis of multiblock and hierarquical PCA and PLS models. *Journal of Chemometrics*, 12, 301-321.
Wold, S.; Sjostrom, M. e Eriksson, L. (2001a), PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109-130.
Wold, W.; Trygg, J.; Berglund, A. e Antti, H. (2001b), Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 58, 131-150.
Zitzler, E. e Thiele, L. (1999), Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach, *IEEE Transactions on Evolutionary Computation*, 3, 257-271.
Zopounidis, C. e Doumpos, M. (2002), Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, 138, 229-246.